Critical reading of scientific articles on health: A practical guide

(Spanish Edition)

South American Publishing

Critical reading of scientific articles on health: A practical guide (Spanish Edition)

Micaela Mariel Moncada Mercado
Freddy Ednildon Bautista-Vanegas
Luis Mariano Tecuatl Gómez
Mauricio Alejandro Paz del Rio
Jaykel Evelio Gómez Triana
Paul Cardozo-Gil
Jhossmar Cristians Auza-Santivañez
Carlos Alberto Paz-Roman
Blas Apaza-Huanca
Josué Elías Peca-Hoyos
Antonio Viruez-Soto
Nadia Sandra Orozco Vargas
Timothé Schenker
Ariel Sosa-Remón
Carmen Julia Salvatierra-Rocha
Jorge Marquez-Molina
Alberto Martin Diaz-Seminario
Rolando Antonio Quiroz Quispe
Alejandro Carías
María Lourdes del Rosario Escalera Rivero
Adrian Avila-Hilari
Edgar Juan José Chávez Navarro
Nayra Condori-Villca
Marco Antonio Gumucio-Villarroel
Isaura Oberson-Santander
María Valeria Canedo-Sánchez
Isis Scarleth Funes Galindo
Stanley Aguirre
Giovanni Callizaya-Macedo
Weymar Lehi Poma Luna
Guiselle Carol Cabrera Morales
Helen Fernández Burgoa
Pamela Gutierrez Villegas
Aaron Eduardo Carvajal-Tapia
José Bernardo Antezana-Muñoz
Adolfo Israel Vasquez Cuellar
Eloy Paycho Anagua
Maria Tereza Nieto Coronel
Boris Adolfo Llanos Torrico
Rita Marlene Mamani-Limachi
Henrry Temis Quisbert Vasquez
Fidel Aguilar Medrano
Leonel Rivero Castedo
Marizol Garcia-Leon
Ismael Vargas Gallego

South American Publishing

Authors

Micaela Mariel Moncada Mercado
https://orcid.org/0000-0002-2403-8930

Freddy Ednildon Bautista-Vanegas
https://orcid.org/0009-0005-5827-7066

Luis Mariano Tecuatl Gómez
https://orcid.org/0000-0003-3860-9332

Mauricio Alejandro Paz del Rio
https://orcid.org/0000-0003-1046-1567

Jaykel Evelio Gómez Triana
https://orcid.org/0000-0003-2571-925X

Paul Cardozo-Gil
https://orcid.org/0009-0005-8397-3901

Jhossmar Cristians Auza-Santivañez
https://orcid.org/0000-0002-7703-2241

Carlos Alberto Paz-Roman
https://orcid.org/0000-0001-9746-9523

Blas Apaza-Huanca
https://orcid.org/0009-0002-4738-6126

Josué Elías Peca-Hoyos
https://orcid.org/0009-0004-5321-8449

Antonio Viruez-Soto
https://orcid.org/0000-0002-0579-5186

Nadia Sandra Orozco Vargas
https://orcid.org/0000-0001-8693-6317

Timothé Schenker
https://orcid.org/0009-0007-7847-5662

Ariel Sosa-Remón
https://orcid.org/0000-0002-5128-4600

Carmen Julia Salvatierra-Rocha
https://orcid.org/0009-0004-5181-8977

Jorge Marquez-Molina
https://orcid.org/0000-0001-8251-490X

Alberto Martin Diaz-Seminario
https://orcid.org/0000-0002-3137-0587

Rolando Antonio Quiroz Quispe
https://orcid.org/0000-0002-2957-1232

Alejandro Carías
https://orcid.org/0000-0002-4578-1811

María Lourdes del Rosario Escalera Rivero
https://orcid.org/0009-0005-6986-2703

Adrian Avila-Hilari
https://orcid.org/0000-0001-9797-3033

Edgar Juan José Chávez Navarro
https://orcid.org/0009-0003-5994-3134

Nayra Condori-Villca
https://orcid.org/0000-0002-2631-5276

Marco Antonio Gumucio-Villarroel
https://orcid.org/0000-0001-6852-2663

Isaura Oberson-Santander
https://orcid.org/0009-0006-7563-5037

María Valeria Canedo-Sánchez
https://orcid.org/0009-0003-3435-2290

Isis Scarleth Funes Galindo
https://orcid.org/0009-0003-3153-4443

Giovanni Callizaya-Macedo
https://orcid.org/0009-0002-7509-5537

Weymar Lehi Poma Luna
https://orcid.org/0009-0003-4341-1512

Guiselle Carol Cabrera Morales
https://orcid.org/0009-0000-3798-227X

Helen Fernández Burgoa
https://orcid.org/0009-0009-4532-6737

Pamela Gutierrez Villegas
https://orcid.org/0009-0003-0648-8692

Aaron Eduardo Carvajal-Tapia
https://orcid.org/0000-0002-4900-2299

José Bernardo Antezana-Muñoz
https://orcid.org/0009-0001-7723-8274

Adolfo Israel Vasquez Cuellar
https://orcid.org/0000-0002-4363-2601

Eloy Paycho Anagua
https://orcid.org/0009-0003-4891-5867

Maria Tereza Nieto Coronel
https://orcid.org/0000-0002-5606-6434

Stanley Aguirre

Boris Adolfo Llanos Torrico
https://orcid.org/0009-0004-5943-7642

Rita Marlene Mamani-Limachi
https://orcid.org/0009-0005-3187-3677

Henrry Temis Quisbert Vasquez
https://orcid.org/0009-0002-8548-8529

Fidel Aguilar Medrano
https://orcid.org/0009-0001-8867-0580

Leonel Rivero Castedo
https://orcid.org/0000-0002-9115-1271

Marizol Garcia-Leon
https://orcid.org/0009-0005-2719-0400

Ismael Vargas Gallego
https://orcid.org/0000-0002-5031-9226

Copyright Page

© 2026. The authors. This is an open access book, distributed under the terms of a Creative Commons Attribution 4.0 license (https://creativecommons.org/licenses/by/4.0/) that allows use, distribution, and reproduction in any medium provided that the original work is properly cited.

This South American Publishing imprint is published by South American Publishing.

For more information, see South American Publishing's Open Access Policy: https://southam.pub/policies.html

ISBN (English edition, eBook): 978-9915-9928-0-8

ISBN (English edition, Softcover): 978-9915-9928-1-5

ISBN (Spanish edition, eBook): 978-9915-9928-3-9

ISBN (Spanish edition, Softcover): 978-9915-9928-2-2

This title is available in both print (softcover) and digital (open access PDF) formats.

DOI: 10.62486/978-9915-9928-3-9

Cataloging Data

Cataloging-in-Publication Data (CIP):

South American Publishing

Critical reading of scientific articles on health: A practical guide (Spanish Edition) / Micaela Mariel Moncada Mercado, Freddy Ednildon Bautista-Vanegas, Luis Mariano Tecuatl Gómez, Mauricio Alejandro Paz del Rio, Jaykel Evelio Gómez Triana, Paul Cardozo-Gil, Jhossmar Cristians Auza-Santivañez, Carlos Alberto Paz-Roman, Blas Apaza-Huanca, Josué Elías Peca-Hoyos, Antonio Viruez-Soto, Nadia Sandra Orozco Vargas, Timothé Schenker, Ariel Sosa-Remón, Carmen Julia Salvatierra-Rocha, Jorge Marquez-Molina, Alberto Martin Diaz-Seminario, Rolando Antonio Quiroz Quispe, Alejandro Carías, María Lourdes del Rosario Escalera Rivero, Adrian Avila-Hilari, Edgar Juan José Chávez Navarro, Nayra Condori-Villca, Marco Antonio Gumucio-Villarroel, Isaura Oberson-Santander, María Valeria Canedo-Sánchez, Isis Scarleth Funes Galindo, Stanley Aguirre, Giovanni Callizaya-Macedo, Weymar Lehi Poma Luna, Guiselle Carol Cabrera Morales, Helen Fernández Burgoa, Pamela Gutierrez Villegas, Aaron Eduardo Carvajal-Tapia, José Bernardo Antezana-Muñoz, Adolfo Israel Vasquez Cuellar, Eloy Paycho Anagua, Maria Tereza Nieto Coronel, Boris Adolfo Llanos Torrico, Rita Marlene Mamani-Limachi, Henrry Temis Quisbert Vasquez, Fidel Aguilar Medrano, Leonel Rivero Castedo, Marizol Garcia-Leon, and Ismael Vargas Gallego. Montevideo: South American Publishing, 2026.

ISBN (English edition, eBook): 978-9915-9928-0-8

ISBN (English edition, Softcover): 978-9915-9928-1-5

ISBN (Spanish edition, eBook): 978-9915-9928-3-9

ISBN (Spanish edition, Softcover): 978-9915-9928-2-2

BISAC classification:

MED112000 - MEDICAL / Evidence-Based Medicine

MED028000 - MEDICAL / Epidemiology

MED090000 - MEDICAL / Biostatistics

MED014000 - MEDICAL / Clinical Medicine

Editorial Notice and Acknowledgments

The publication of this book is part of South American Publishing's mission to promote open, ethical, and rigorous scientific communication across all areas of knowledge.

All books published by South American Publishing undergo a double-blind peer-review process and a technical evaluation in accordance with the publisher's editorial policies, aligned with COPE guidelines.

South American Publishing acknowledges the valuable collaboration of authors, reviewers, designers, and production teams who made this publication possible.

Preface

The ability to interpret scientific evidence has become one of the defining competencies of modern healthcare. Every day, clinicians encounter an expanding body of publications, rapidly evolving technologies, and increasingly complex decisions that demand more than familiarity with medical knowledge. They require the capacity to question, evaluate, compare, and apply evidence with intellectual rigor while remaining attentive to the individual circumstances of each patient.

This book was conceived as a practical and educational guide for that purpose. Rather than presenting Evidence-Based Medicine as a collection of isolated concepts, it approaches it as a continuous process of clinical reasoning. The chapters move from the foundations of research methodology and scientific communication to the interpretation of observational and experimental studies, systematic reviews, diagnostic research, and emerging approaches that incorporate artificial intelligence into evidence appraisal and clinical decision-making.

Throughout these pages, methodological principles are consistently connected with realistic clinical situations, allowing readers to appreciate not only how evidence is generated but also how it can be interpreted responsibly. Particular attention is given to recognizing methodological strengths, identifying potential sources of bias, understanding the limitations of different study designs, and distinguishing statistical findings from clinically meaningful conclusions.

The final chapters acknowledge that medicine is entering a new stage in which artificial intelligence is becoming an increasingly common companion in literature searching, evidence synthesis, and clinical support. These technologies offer remarkable opportunities, yet they also introduce ethical, methodological, and professional challenges that demand critical judgment rather than unquestioning acceptance.

This volume is intended for medical students, residents, clinicians, educators, and health researchers seeking to strengthen their analytical skills and cultivate a disciplined approach to scientific evidence. More than offering definitive answers, it invites readers to develop the habit of asking better questions, evaluating information carefully, and making decisions that remain grounded in scientific integrity and compassionate patient care.

Introduction

Evidence-Based Medicine has transformed the way healthcare professionals approach diagnosis, treatment, prevention, and clinical decision-making. Its central principle is straightforward: clinical decisions should emerge from the integration of the best available scientific evidence, professional expertise, and the values and preferences of patients. Translating this principle into everyday practice, though, requires methodological knowledge, critical thinking, and the ability to navigate an ever-growing scientific literature.

This book provides a structured pathway through the essential components of that process. It begins by introducing the principles of critical appraisal, research methodology, scientific writing, and literature searching. These foundations prepare readers to understand how evidence is generated before exploring the characteristics, strengths, and limitations of the principal epidemiological and clinical research designs, including descriptive studies, cohort studies, case-control studies, randomized controlled trials, diagnostic accuracy studies, quasi-experimental research, systematic reviews, and meta-analyses.

Recognizing the profound impact of digital technologies on scientific practice, the book also examines the expanding role of artificial intelligence in biomedical research. Rather than presenting AI as a replacement for scientific reasoning, the chapters describe it as a complementary resource whose value depends on transparent verification, methodological rigor, and continuous human oversight. Ethical considerations, algorithmic limitations, and responsible implementation therefore occupy a central place within the discussion.

Although each chapter can be consulted independently, together they build a coherent framework for understanding how scientific evidence is produced, evaluated, synthesized, and translated into clinical practice. Readers are encouraged to move beyond memorizing methodological concepts and instead cultivate a reflective mindset capable of interpreting evidence within its clinical, ethical, and societal context.

Ultimately, this book seeks to reinforce a simple but enduring idea: sound clinical decisions are not determined solely by the availability of information, but by the quality of the reasoning used to interpret it.

Abstract

The rapid expansion of biomedical literature has increased the need for healthcare professionals to critically evaluate scientific evidence before incorporating it into clinical practice. This book presents a comprehensive introduction to the principles and methods of Evidence-Based Medicine, integrating research methodology, critical appraisal, epidemiological study designs, literature searching, diagnostic evaluation, systematic evidence synthesis, and contemporary applications of artificial intelligence in healthcare. Beginning with the conceptual foundations of scientific reasoning, the chapters progressively examine the structure of scientific publications, research design, observational and experimental studies, diagnostic accuracy research, systematic reviews, meta-analyses, and evidence interpretation. Practical clinical scenarios accompany methodological explanations to illustrate how research findings can inform patient care while acknowledging the limitations, biases, and uncertainties inherent to scientific investigation. Particular attention is devoted to the responsible incorporation of artificial intelligence into literature retrieval, comparative evidence analysis, clinical decision support, and ethical deliberation. Throughout the volume, technological advances are presented as complementary resources that enhance, rather than replace, professional judgment. By combining methodological rigor with clinically relevant examples, the book offers readers a coherent framework for understanding how scientific evidence is generated, evaluated, synthesized, and responsibly applied. It is intended for students, clinicians, educators, and researchers who wish to strengthen their capacity for critical reasoning and evidence-informed healthcare practice.

Keywords: Evidence-Based Medicine; Critical Appraisal; Clinical Research Methodology; Scientific Evidence; Artificial Intelligence; Clinical Decision-Making

Index

Chapter 01. The importance of critical reading in medical practice | La importancia de la lectura crítica en la práctica médica 1

Chapter 02. Structure of a Scientific Article | Estructura de un artículo científico 9

Chapter 03. Introduction to Evidence-Based Medicine | Introducción a la Medicina Basada en la Evidencia 16

Chapter 04. The research process in the context of Evidence-Based Medicine | El proceso de investigación en el contexto de la Medicina Basada en la Evidencia 27

Chapter 05. How to find a scientific article on health? A step-by-step strategy | ¿Cómo buscar un artículo científico en salud? Estrategia paso a paso 41

Chapter 06. Búsqueda de artículos de investigación con inteligencia artificial | Searching for Research Articles with Artificial Intelligence 51

Chapter 07. Research Methodology in the Health Field | Metodología de la Investigación en el Área de la Salud 70

Chapter 08. Practical Guide for Searching Scientific Articles | Guía práctica para búsqueda de artículos científicos 83

Chapter 09. Case Report and Case Series | Reporte de Casos y Serie de Casos 98

Chapter 10. Descriptive studies (cross-sectional, ecological, agreement) | Estudios descriptivos (transversal, ecológico, concordancia) 116

Chapter 11. Diagnostic Test Accuracy Studies | Estudios de prueba diagnóstica 136

Chapter 12. Case-Control Studies | Estudios de casos y controles 156

Chapter 13. Cohort Studies | Estudios de Cohorte 178

Chapter 14. Randomized Controlled Trials | Ensayos Clínicos Aleatorizados 209

Chapter 15. Quasi-Experimental Studies | Estudios cuasiexperimentales 243

Chapter 16. Systematic Reviews and Meta-Analyses | Revisiones Sistemáticas y Metaanálisis 263

Chapter 17. Non-Systematized Evidence Sources in Medicine | Fuentes de evidencia no sistematizada en medicina 299

Chapter 18. Use of Artificial Intelligence in the Critical Appraisal of Scientific Medical Articles | Uso de la inteligencia artificial en la lectura crítica de artículos científicos médicos..... 311

Chapter 19. Use of NotebookLM in the Comparative Analysis of Scientific Evidence | Uso de NotebookLM en el análisis comparativo de evidencia científica 333

Chapter 20. Artificial Intelligence and Evidence-Based Clinical Decision-Making | Inteligencia Artificial y toma de decisiones clínicas basadas en evidencia 344

Chapter 21. Ethics and Limitations of Artificial Intelligence in Medicine | Ética y limitaciones de la inteligencia artificial en medicina 356

Chapter 1 / Capítulo 1

Critical reading of scientific articles on health: A practical guide (Spanish Edition)
ISBN: 978-9915-9928-3-9
DOI: 10.62486/978-9915-9928-3-9.Ch01
Pages: 1-8

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

The importance of critical reading in medical practice

La importancia de la lectura crítica en la práctica médica

Micaela Mariel Moncada Mercado
Freddy Ednildon Bautista-Vanegas

1 Universidad Técnica de Oruro. La Paz, Bolivia.
2 Neurokliniken Beelitz GmbH Neurologische Rehabilitationsklinik: Beelitz Heilstätten, Brandenburg, DE. Germany.
3 Neurologisches Fachkrankenhaus für Bewegungsstörungen/Parkinson: Beelitz Heilstätten, Brandenburg, DE. Germany.

ABSTRACT

Introduction: Critical appraisal is an essential skill for contemporary medical practice, given the exponential growth of scientific literature and the variability in the methodological quality of research. Its application allows for the systematic evaluation of the validity, clinical relevance, and applicability of evidence, promoting decisions based on the principles of Evidence-Based Medicine (EBM) and aimed at optimizing patient safety. Development: This chapter uses a clinical case to illustrate how the critical interpretation of a study prevents decisions based on insufficient or biased evidence. Emphasis is placed on evaluating the methodological design, the magnitude of the effect, conflicts of interest, and the applicability of the results. It also integrates the responsible use of artificial intelligence as a support tool, without replacing clinical judgment or scientific reasoning. Conclusions: Critical appraisal transcends the analysis of scientific articles, becoming a reflective process that strengthens clinical judgment, promotes safe therapeutic decisions, and fosters a culture of continuous learning. In an environment of increasing scientific output and technological development, mastery of this skill is an indispensable requirement for practicing ethical, rigorous, and patient-centered medicine, ensuring that clinical practice is based on solid and clinically relevant evidence.

Keywords: Evidence-Based Medicine; Critical Appraisal; Clinical Decision-Making; Biomedical Research; Artificial Intelligence

RESUMEN

Introducción: La lectura crítica constituye una competencia esencial para la práctica médica contemporánea, debido al crecimiento exponencial de la literatura científica y a la variabilidad en la calidad metodológica de las investigaciones. Su aplicación permite evaluar de manera sistemática la validez, relevancia clínica y aplicabilidad de la evidencia, favoreciendo decisiones fundamentadas en los principios de la Medicina Basada en la Evidencia (MBE) y orientadas a optimizar la seguridad del paciente. Desarrollo: El capítulo expone, mediante un caso clínico, cómo la interpretación crítica de un estudio evita decisiones sustentadas en evidencia insuficiente o sesgada. Se enfatiza la evaluación del diseño metodológico, la magnitud del efecto, los conflictos de interés y la aplicabilidad de los resultados, integrando además el uso responsable de la inteligencia artificial como herramienta de apoyo, sin sustituir el juicio clínico ni el razonamiento científico. Conclusiones: La lectura crítica trasciende el análisis de artículos científicos para convertirse en un proceso reflexivo que fortalece el juicio clínico, promueve decisiones terapéuticas seguras y fomenta una cultura de aprendizaje continuo. En un entorno de creciente producción científica y desarrollo tecnológico, su dominio representa un requisito indispensable para ejercer una medicina ética, rigurosa y centrada en el paciente, garantizando que la práctica clínica se sustente en evidencia sólida y clínicamente relevante.

Palabras clave: Evidence-Based Medicine; Critical Appraisal; Clinical Decision-Making; Biomedical Research; Artificial Intelligence

1. CASO CLÍNICO

Durante la consulta médica, un residente de Medicina Interna, el Dr. Martínez, evalúa a un paciente de 55 años de edad con antecedentes de hipertensión arterial y dislipidemia, ambos considerados factores de riesgo cardiovascular relevantes. Durante la entrevista clínica, el paciente consulta acerca de un nuevo medicamento que ha observado promocionado en redes sociales, cuya publicidad afirma que es capaz de reducir en un 50% el riesgo de infarto agudo de miocardio. Ante esta afirmación, el residente decide contrastar la información mediante una búsqueda sistemática en una base de datos biomédica especializada. Como resultado, identifica un estudio clínico reciente que evalúa la eficacia del fármaco en la prevención de eventos cardiovasculares. En una revisión preliminar, los hallazgos parecen alentadores, ya que los autores reportan una reducción del 50% en el riesgo de infarto en el grupo que recibió el tratamiento en comparación con el grupo control. No obstante, antes de incorporar esta intervención a su práctica clínica o formular una recomendación terapéutica, el residente, recurre a los principios de la Medicina Basada en la Evidencia (MBE) y procede a realizar una lectura crítica del artículo científico. Para ello, examina de manera rigurosa la validez metodológica del estudio, la magnitud real del efecto observado, la relevancia clínica de los resultados y la aplicabilidad de los hallazgos a las características específicas de su paciente.

Al profundizar en el análisis metodológico del estudio, el médico residente identifica diversos aspectos que limitan la validez interna y externa de los resultados reportados:

Diseño del estudio. El médico constata que la investigación corresponde a un estudio observacional y no a un ensayo clínico aleatorizado. Esta característica metodológica implica una menor capacidad para establecer relaciones causales entre la intervención y los resultados observados, debido a la posible influencia de factores de confusión y sesgos inherentes al diseño. En consecuencia, aunque el estudio sugiere una asociación entre el uso del medicamento y la reducción del riesgo de infarto, no permite afirmar con certeza que dicho efecto sea atribuible exclusivamente al fármaco.

• Tamaño muestral y selección de los participantes. Durante la evaluación de la población de estudio, observa que la muestra estuvo conformada por únicamente 500 pacientes reclutados en un solo hospital privado. Esta situación limita la representatividad de la muestra y reduce la capacidad de extrapolar los hallazgos a otras poblaciones con características sociodemográficas, clínicas y epidemiológicas diferentes. Asimismo, el reducido tamaño muestral podría afectar la precisión de las estimaciones obtenidas.

• Magnitud y relevancia clínica del efecto observado. Al examinar detenidamente los resultados, el residente advierte que la aparente reducción del 50% corresponde a una reducción del riesgo relativo (RRR), medida que puede magnificar la percepción del beneficio terapéutico cuando se presenta de forma aislada. Sin embargo, al calcular la reducción del riesgo absoluto (RRA), encuentra que esta alcanza únicamente el 1%, lo que indica que el beneficio clínico real es considerablemente menor de lo que inicialmente sugieren los resultados. Este hallazgo resalta la importancia de interpretar conjuntamente las medidas de efecto para valorar adecuadamente la relevancia clínica de una intervención.

• Posibles conflictos de interés y fuente de financiamiento. En la sección correspondiente al financiamiento y declaración de intereses, el médico identifica que la investigación fue patrocinada por la misma compañía farmacéutica responsable de la producción y comercialización del medicamento evaluado. Aunque esta situación no invalida automáticamente los resultados, constituye una potencial fuente de sesgo que debe ser considerada durante la interpretación de los hallazgos, especialmente en aspectos relacionados con el diseño del estudio, el análisis estadístico y la presentación de las conclusiones.

Tras completar la evaluación crítica de la evidencia disponible, el Dr. Martínez concluye que los resultados del estudio, si bien son preliminarmente prometedores, no proporcionan evidencia suficientemente robusta para respaldar una recomendación clínica inmediata. En consecuencia, decide mantener la conducta terapéutica basada en las evidencias científicas actualmente disponibles y esperar la publicación de ensayos clínicos aleatorizados, multicéntricos y metodológicamente rigurosos que permitan confirmar la eficacia y seguridad del medicamento. Esta decisión refleja la aplicación práctica de los principios de la Medicina Basada en la Evidencia, donde las decisiones clínicas deben sustentarse en información científica de alta calidad, integrada con la experiencia profesional y las características individuales de cada paciente.

https://doi.org/10.62486/978-9915-9928-3-9.Ch01

2. OBJETIVO

- Introducir al lector a la importancia de la lectura crítica en medicina.
- Explicar cómo la lectura crítica apoya la toma de decisiones clínicas basadas en evidencia y mejora la práctica médica.
- Preparar al lector para el contenido del libro y cómo utilizará la lectura crítica en su formación.

3. INTRODUCCIÓN

En las últimas décadas, el conocimiento científico en el ámbito de las ciencias de la salud ha experimentado un crecimiento acelerado y continuo. La producción de artículos científicos, revisiones sistemáticas, ensayos clínicos y estudios observacionales aumenta de manera exponencial, generando una gran cantidad de información disponible para los profesionales sanitarios. En este contexto, la capacidad de realizar una lectura crítica de la literatura científica constituye una competencia esencial para médicos, residentes, internos, licenciados en enfermería, bioquímicos, odontólogos, farmacéuticos y demás integrantes del equipo de salud. La lectura crítica es un proceso sistemático de análisis y evaluación de la evidencia científica que permite determinar la validez metodológica, la relevancia clínica y la aplicabilidad de los resultados de una investigación. Esta habilidad trasciende la simple comprensión del contenido de un artículo, ya que implica valorar rigurosamente la calidad del diseño del estudio, identificar posibles sesgos, interpretar adecuadamente los resultados y juzgar la pertinencia de las conclusiones emitidas por los autores.En el ejercicio cotidiano de la práctica clínica, las decisiones relacionadas con el diagnóstico, tratamiento, pronóstico y prevención de enfermedades deben sustentarse en información científica confiable y actualizada. Sin embargo, no toda la evidencia publicada posee el mismo nivel de calidad metodológica ni ofrece resultados clínicamente relevantes. La existencia de errores sistemáticos, limitaciones en el diseño de investigación, conflictos de interés o interpretaciones inadecuadas puede comprometer la validez de los hallazgos y conducir a decisiones clínicas incorrectas.

Por ello, la lectura crítica se constituye en una herramienta fundamental de la Medicina Basada en la Evidencia (MBE), permitiendo a los profesionales de la salud distinguir entre investigaciones de alta y baja calidad, identificar aquellas evidencias que aportan beneficios reales para los pacientes y evitar la adopción de intervenciones sustentadas en información insuficiente o metodológicamente deficiente. En consecuencia, el desarrollo de competencias en lectura crítica no solo fortalece el juicio clínico y la toma de decisiones informadas, sino que también contribuye a mejorar la calidad de la atención sanitaria, la seguridad del paciente y los resultados en salud.

4. IMPACTO DE LA LECTURA CRÍTICA EN LA TOMA DE DECISIONES CLÍNICAS

La toma de decisiones debe apoyarse en el conocimiento científico más reciente, en conjunto con la experiencia clínica y las preferencias del paciente. La lectura crítica facilita este proceso, ofreciendo una estructura para evaluar la validez de un artículo, su impacto potencial y su aplicabilidad en un contexto clínico específico.

Con el enfoque que se desarrollará en este libro, el lector aprenderá a interpretar los resultados de estudios científicos con un criterio riguroso, desde entender la introducción, analizar la metodología del estudio, detectar los posibles sesgos y limitaciones, y evaluar la relevancia de los hallazgos aplicado al contexto que se encuentra viviendo.

5. CÓMO ESTE LIBRO AYUDARÁ A DESARROLLAR LA HABILIDAD DE LECTURA CRÍTICA

- A lo largo de este libro, se analizará cada tipo de artículo científico con una metodología sistemática y clara. En cada capítulo, se encontrará:
- Casos clínicos introductorios: Situaciones clínicas que motivan la búsqueda y lectura de un tipo específico de artículo y que serán la guía para el resto del capítulo.
- Guías para entender la estructura y metodología: Se explicará la estructura de cada tipo de estudio y los elementos clave de su metodología.
- Interpretación estadística: Para aprender a interpretar las estadísticas específicas de cada tipo de estudio, permitiendo entender sus resultados en profundidad.

https://doi.org/10.62486/978-9915-9928-3-9.Ch01

Figura 1: importancia de la lectura crítica en la toma de decisiones clínicas.

Fuente:Elaboración propia

- Autoevaluación: del contenido para reforzar lo aprendido.
- Checklists: Para poder identificar rápidamente cada parte del artículo.

Este enfoque permitirá que, al final de cada capítulo, el lector sea capaz de analizar críticamente un tipo específico de artículo, comprendiendo en detalle su estructura, metodología, resultados y aplicabilidad clínica.

6. LA LECTURA CRÍTICA COMO BASE PARA LA MEDICINA BASADA EN LA EVIDENCIA (MBE)

La lectura crítica es uno de los pilares de la MBE. La MBE integra la mejor evidencia científica disponible con la experiencia clínica y las preferencias del paciente para tomar decisiones médicas. A medida que avances en este libro, el lector adquirirá habilidades que son esenciales para aplicar la MBE en la práctica médica, como formular preguntas clínicas, buscar y seleccionar evidencia, evaluar su validez y relevancia.

En futuras producciones de otros libros, se explicarán los pasos completos de la MBE, pero por ahora, el enfoque será dominar la lectura crítica de artículos científicos. Esto permitirá construir una base sólida para el análisis y la aplicación de la evidencia en la medicina.

7. LA INTELIGENCIA ARTIFICIAL EN LA LECTURA CRÍTICA

En la actualidad, la inteligencia artificial representa una herramienta complementaria y poderosa para el análisis de artículos científicos. Por ello, cada capítulo de esta primera parte del libro también incluirá secciones que muestran cómo utilizar la IA de forma ética y eficaz para facilitar la lectura crítica, sin reemplazar el juicio clínico ni la evaluación metodológica.Se explorará el uso de herramientas como ChatGPT, Gemini, Copilot, etc. motores de búsqueda científicos con IA, detectores de sesgos automatizados, y generadores de resúmenes inteligentes, siempre con un enfoque crítico y formativo. La IA no sustituye la formación médica, pero bien utilizada, potencia la capacidad del médico en formación para comprender, analizar y aplicar la evidencia.

8. CONCLUSIONES

La lectura crítica es una competencia esencial para todo personal de salud que aspire a

https://doi.org/10.62486/978-9915-9928-3-9.Ch01

brindar una atención basada en la mejor evidencia disponible. En un entorno donde la información científica crece de manera exponencial, distinguir entre estudios bien diseñados y aquellos con limitaciones metodológicas se vuelve una habilidad indispensable para la toma de decisiones clínicas seguras y efectivas.Dominar la lectura crítica no solo permitirá interpretar con mayor precisión la literatura médica, sino que también contribuirá a mejorar la seguridad y eficacia de tus decisiones clínicas, optimizando los resultados en la atención de sus pacientes.

9. PREGUNTAS DE AUTOEVALUACIÓN

PREGUNTAS DE OPCIÓN MÚLTIPLE

1. ¿Cuál es la principal utilidad de la lectura crítica en la práctica médica? a) Memorizar los resultados de los estudios más recientes.
b) Aceptar las conclusiones de los estudios sin cuestionarlas.
c) Evaluar la calidad y aplicabilidad de los estudios científicos antes de utilizarlos en la práctica clínica.
d) Utilizar únicamente estudios publicados en revistas de alto impacto
2. ¿Cuál de las siguientes afirmaciones es correcta respecto a la lectura crítica? a) Todos los estudios publicados tienen la misma calidad metodológica.
b) La lectura crítica solo es útil para investigadores, no para médicos clínicos ni personal de salud.
c) La lectura crítica ayuda a detectar posibles sesgos y errores metodológicos en un estudio.
d) Un estudio observacional siempre es superior a un ensayo clínico aleatorizado.
3. ¿Qué elemento noes un criterio fundamental al evaluar la validez de un estudio? a) Tamaño muestral adecuado.
b) Metodología bien definida.
c) Opinión de expertos sobre el artículo.
d) Identificación de posibles sesgos.
4. ¿Por qué es importante interpretar los resultados estadísticos en la lectura crítica? a) Porque permite entender si los hallazgos son clínicamente relevantes.
b) Porque siempre se deben aceptar los valores de p como definitivos.
c) Porque los estudios con significancia estadística siempre son aplicables a todos los pacientes.
d) Porque evita la necesidad de leer la metodología del estudio.
5. ¿Cuál de los siguientes conceptos está directamente relacionado con la MBE? a) Experiencia clínica del médico.
b) Preferencias del paciente.
c) Evidencia científica disponible.
d) Todas las anteriores.

PREGUNTAS ABIERTAS:

6. Explica en tus palabras por qué la lectura crítica es fundamental en la toma de decisiones clínicas.
7. Menciona y explica brevemente tres posibles consecuencias de no aplicar la lectura crítica al evaluar un estudio científico.
8. ¿Cómo se relaciona la lectura crítica con la Medicina Basada en la Evidencia? Proporciona un ejemplo.
9. Describe un caso en el que la lectura crítica haya evitado una mala decisión clínica. Puede ser un ejemplo real o hipotético.
10. ¿Cuáles son los principales desafíos que enfrenta un médico en formación al desarrollar la habilidad de lectura crítica y cómo pueden superarse?

RESPUESTAS:

1. c) Evaluar la calidad y aplicabilidad de los estudios científicos antes de utilizarlos en la práctica clínica. La lectura crítica permite discernir si un estudio es válido y útil para resolver una pregunta clínica específica, evitando decisiones basadas en evidencia débil.

https://doi.org/10.62486/978-9915-9928-3-9.Ch01

2. c) La lectura crítica ayuda a detectar posibles sesgos y errores metodológicos en un estudio. Su propósito es analizar críticamente cada componente del estudio, no aceptar su conclusión sin cuestionamientos.

3. c) Opinión de expertos sobre el artículo. Aunque las opiniones pueden orientar, no reemplazan una evaluación estructurada basada en la metodología y resultados.

4. a) Porque permite entender si los hallazgos son clínicamente relevantes. La estadística debe vincularse al juicio clínico; no todo resultado significativo estadísticamente implica beneficio para el paciente.

5. d) Todas las anteriores. La Medicina Basada en la Evidencia se construye integrando experiencia clínica, evidencia científica y preferencias del paciente.

6. La lectura crítica es esencial porque permite al personal de salud distinguir entre estudios bien diseñados y aquellos con fallas metodológicas. Esto asegura que las decisiones clínicas se basen en evidencia válida, relevante y aplicable, optimizando así el cuidado del paciente.

7. Adopción de tratamientos ineficaces: Al no analizar la calidad del estudio, se pueden implementar intervenciones sin beneficio comprobado. Riesgos para el paciente: Se pueden indicar terapias con efectos adversos no bien documentados. Mala asignación de recursos: Se invierte tiempo y dinero en prácticas sin sustento científico sólido.

8. La lectura crítica es el mecanismo mediante el cual se evalúa la evidencia científica antes de aplicarla. Por ejemplo, ante un paciente con insuficiencia cardíaca, el médico puede encontrar un artículo sobre un nuevo fármaco. La lectura crítica le permite determinar si el estudio es válido y si sus resultados son aplicables a su paciente.

9. Un residente encontró un artículo que promovía el uso de un suplemento antioxidante para prevenir eventos cardiovasculares. Antes de recomendarlo, evaluó el estudio y descubrió que era observacional, con sesgo de selección y patrocinado por el fabricante. Gracias a la lectura crítica, evitó indicar un tratamiento no respaldado por evidencia sólida.

10. Falta de entrenamiento formal: Puede superarse con cursos estructurados de MBE y práctica guiada. Terminología estadística compleja: Se recomienda el uso de recursos didácticos y apoyo de tutores. Tiempo limitado: Integrar la lectura crítica en la práctica diaria y usar herramientas como inteligencia artificial puede optimizar el proceso.

https://doi.org/10.62486/978-9915-9928-3-9.Ch01

REFERENCIAS

Rathbone J, Carter M, Hoffmann T, Glasziou P. A comparison of the performance of seven key bibliographic databases in identifying all relevant systematic reviews of interventions for hypertension. Systematic Reviews. 2016 Feb 9;5(1):27. Available at: https://doi.org/10.1186/s13643-016-0197-5

Bramer WM, Rethlefsen ML, Kleijnen J, Franco OH. Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. Syst Rev. 2017 Dec;6(1):245. Available at: https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-017-0644-y

Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev. 2021 Jan 26;10(1):39. Available at: https://doi.org/10.1186/s13643-020-01542-z

Fontanive VN, Müller LP, Riva F, Seoane M, Celeste RK. Development and validation of high-sensitivity and high-specificity pubmed search filters for systematic and non-systematic reviews. BMC Med Res Methodol. 2025 Sep 29;25(1):216. doi:10.1186/s12874-025-02666-3. Available at: https://doi.org/10.1186/s12874-025-02666-3

DeMars MM, Perruso C. MeSH and text-word search strategies: precision, recall, and their implications for library instruction. jmla. 2022 Feb 11;110(1). Available at: https://doi.org/10.5195/jmla.2022.1283

McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement. Journal of Clinical Epidemiology. 2016 Jul;75:40–6. Available at: https://doi.org/10.1016/j.jclinepi.2016.01.021

Navarro-Ruan T, Haynes RB. Preliminary comparison of the performance of the National Library of Medicine's systematic review publication type and the sensitive clinical queries filter for systematic reviews in PubMed. jmla. 2022 Feb 11;110(1). Available at: https://doi.org/10.5195/jmla.2022.1286

Salvador-Oliván JA, Marco-Cuenca G, Arquero-Avilés R. Errors in search strategies used in systematic reviews and their effects on information retrieval. jmla. 2019 Apr 15;107(2). Available at: https://doi.org/10.5195/jmla.2019.567

Glasziou PP, Sanders S, Hoffmann T. Waste in covid-19 research. BMJ. 2020 May 12;m1847. Available at: https://doi.org/10.1136/bmj.m1847

Riganti P, Hoffmann TC. Shared decision-making and evidence-based medicine series: exploring contemporary challenges and future directions. BMJ EBM. 2024 Oct;29(5):283–4. Available at: https://doi.org/10.1136/bmjebm-2024-113024

Chapter 2 / Capítulo 2

Critical reading of scientific articles on health: A practical guide (Spanish Edition)
ISBN: 978-9915-9928-3-9
DOI: 10.62486/978-9915-9928-3-9.Ch02
Pages: 9-15

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Structure of a Scientific Article

Estructura de un artículo científico

Luis Mariano Tecuatl Gómez
Mauricio Alejandro Paz del Rio

1 ISSSTE San José del Cabo. Universidad De Los Cabos, México
2 Clínica Los Andes, La Paz, Bolivia

ABSTRACT

Introduction: The structure of a scientific article is fundamental for the transparent communication of knowledge and the critical appraisal of evidence. Understanding the organization of the IMRAD format and its complementary components allows for the proper interpretation of studies, the assessment of their methodological quality, and facilitates evidence-based decision-making. Development: This chapter describes the specific functions of each section of a scientific article, from the title and abstract to the methods, results, discussion, and references. It emphasizes that each section serves a methodological purpose aimed at ensuring the reproducibility, validity, and appropriate interpretation of the findings. It also analyzes the role of artificial intelligence as a support tool for optimizing the search, synthesis, and analysis of scientific literature, without replacing the researcher's critical judgment. Conclusions: A comprehensive understanding of the structure of a scientific article strengthens the ability to critically appraise and distinguish methodologically sound evidence. Reflectively, the development of these competencies represents an essential requirement for promoting clinical practice grounded in reliable evidence, fostering more objective and reproducible decisions oriented toward the continuous improvement of research and healthcare.

Keywords: Biomedical Research; Scientific Publications; Evidence-Based Medicine; Critical Appraisal; Reproducibility of Results

RESUMEN

Introducción. Estructura de un artículo científico constituye el fundamento para la comunicación transparente del conocimiento y la evaluación crítica de la evidencia. Comprender la organización del formato IMRyD y sus componentes complementarios permite interpretar adecuadamente los estudios, valorar su calidad metodológica y facilitar la toma de decisiones sustentadas en la Medicina Basada en la Evidencia. Desarrollo: El capítulo describe las funciones específicas de cada sección de un artículo científico, desde el título y el resumen hasta los métodos, resultados, discusión y referencias. Destaca que cada apartado cumple un propósito metodológico orientado a garantizar la reproducibilidad, la validez y la adecuada interpretación de los hallazgos. Asimismo, analiza el papel de la inteligencia artificial como herramienta de apoyo para optimizar la búsqueda, síntesis y análisis de la literatura científica, sin reemplazar el juicio crítico del investigador. Conclusiones: La comprensión integral de la estructura de un artículo científico fortalece la capacidad de realizar una lectura crítica y de distinguir evidencia metodológicamente sólida. El desarrollo de estas competencias representa un requisito indispensable para promover una práctica clínica fundamentada en evidencia confiable, favoreciendo decisiones más objetivas, reproducibles y orientadas al mejoramiento continuo de la investigación y la atención en salud.

Palabras clave: Investigación Biomédica; Publicaciones Científicas; Medicina Basada en la Evidencia; Lectura Crítica; Reproducibilidad de Resultados

1. CASO CLÍNICO

María, médico residente de primer año, busca información sobre la efectividad del ácido tranexámico para prevenir hemorragia postparto. Encuentra un artículo prometedor, pero al leerlo se siente confundida por los términos y la organización del contenido. Al tener poco tiempo para leer el artículo, no sabe qué parte le dará la información más relevante ni cómo juzgar si el estudio es confiable.

2. OBJETIVO DE ESTE APARTADO:

- Explicar cada una de las secciones principales de un artículo científico.
- Ofrecer una breve descripción de la información que el lector puede esperar encontrar en cada sección.
- Preparar al lector para abordar el análisis de artículos mediante la comprensión de su estructura y propósito.

3. INTRODUCCIÓN

La publicación de un artículo científico representa la culminación de la investigación, ya que permite divulgar, compartir y contrastar los resultados con la comunidad científica. Para realizar una lectura crítica, es fundamental que el lector comprenda la estructura y función de cada sección de un artículo, lo cual facilita evaluar la calidad metodológica y la aplicabilidad de los hallazgos. A continuación, se presentan las partes esenciales de un artículo de investigación original, comúnmente organizadas en el formato IMRyD (Introducción, Métodos, Resultados y Discusión), además de otros elementos adicionales.

Cabe recalcar que este no es un formato arbitrario, sino un reflejo directo del proceso de investigación, por ello cada revista científica tiene sus propias normas que deben ser respetadas por los autores para la publicación de su artículo, el objetivo de este tema es mencionar los principales acápites y lo que contiene cada uno.

4. ESTRUCTURA DE UN ARTÍCULO ORIGINAL

TÍTULO

El título es lo primero que el lector observa y entre sus características principales destacan el ser breve, claro y explicativo, permitiendo identificar de inmediato el problema de investigación. Es importante notar que no contiene abreviaturas ni explicaciones excesivas. Un buen título genera suficiente interés para motivar al lector a profundizar en el resumen y el artículo completo.

AUTORES Y AFILIACIONES

Los autores son los investigadores que contribuyeron significativamente al diseño, análisis e interpretación de los resultados y en la redacción o revisión crítica del contenido. Además se cuenta con los datos específicos de los autores como su afiliación específica, la institución a la que pertenecían al realizar el estudio, y es común que incluya información de contacto, correos electrónicos y datos de localización para facilitar la comunicación.

Consta de los nombres completos de los investigadores (sin iniciales), siendo el orden en el que aparecen el mero reflejo de la contribución individual que hicieron al trabajo, por ello el primero y el último son los más importantes (el primer autor es el investigador principal y el último es el experto que supervisó todo el trabajo de investigación).

AGRADECIMIENTOS O RECONOCIMIENTOS

Los agradecimientos permiten reconocer a colaboradores que ayudaron técnica, financiera o materialmente, sin alcanzar el nivel de autoría. También se suelen mencionar los posibles conflictos de interés en esta sección.

https://doi.org/10.62486/978-9915-9928-3-9.Ch02

RESUMEN (ABSTRACT)

El resumen es una sección breve que permite a los lectores evaluar rápidamente la relevancia del estudio. Incluye el objetivo, metodología, resultados clave y conclusiones principales, todo redactado en tercera persona en el idioma materno e inglés. Aunque el resumen es informativo, el lector debe profundizar en el artículo completo para una evaluación crítica detallada.

PALABRAS CLAVE

Es una lista de términos específicos relevantes para el artículo, que facilita la indexación y búsqueda del artículo en bases de datos. Generalmente, se utilizan términos de Medical Subject Headings (MeSH) en el idioma inglés o los Descriptores de Ciencias de la Salud (DeCS) en español.

INTRODUCCIÓN

La introducción establece el contexto de la investigación, identificando el problema y justificando la importancia del estudio usando fuentes bibliográficas de actualidad con información específica del problema en concreto, se sitúa y define los términos y abreviaturas especializadas que se vaya a utilizar en todo el artículo. También plantea el objetivo específico del estudio en su último párrafo, lo que facilita la comprensión del artículo y la conexión con los resultados. En términos generales, su redacción explica la necesidad del estudio y estimula la curiosidad del lector.

MÉTODOS (MATERIALES Y MÉTODOS)

La sección de métodos es clave para la reproducibilidad y validez del estudio, ya que describe cómo se realizó la investigación. Incluye el diseño del estudio, criterios de inclusión y exclusión, la muestra y entorno, las intervenciones, las variables, los procedimientos, y la estrategia de análisis estadístico. Este nivel de detalle permite que otros investigadores puedan replicar o mejorar el estudio.

RESULTADOS

Los resultados presentan los hallazgos sin interpretación adicional, utilizando textos, tablas y figuras para resaltar los datos más importantes. La presentación suele seguir un orden lógico: características descriptivas de los sujetos, análisis de la comparabilidad de grupos y resultados principales de acuerdo con el objetivo del estudio.

TABLAS Y FIGURAS:

Las tablas y figuras se numeran consecutivamente y ofrecen un resumen visual de los datos. Las tablas muestran datos en filas y columnas, mientras que las figuras representan datos numéricos de manera gráfica. Es común incluir un título breve y explicativo para cada tabla o figura y una descripción de las abreviaturas utilizadas.

DISCUSIÓN

La discusión analiza los resultados en relación con el objetivo del estudio y compara los hallazgos con otros estudios similares. Incluye una autocrítica honesta de los puntos fuertes y limitaciones, así como recomendaciones para investigaciones futuras. La discusión permite evaluar si los autores interpretan sus resultados de manera objetiva y reconocen posibles limitaciones en el diseño o la ejecución del estudio.

CONCLUSIONES

https://doi.org/10.62486/978-9915-9928-3-9.Ch02

Las conclusiones sintetizan los hallazgos clave en función de los objetivos del estudio y las respuestas a las preguntas planteadas. Las conclusiones sebasan en los resultados y no se extrapolan más allá de los datos obtenidos.

Dependiendo de los requerimientos de la revista, este acápite puede encontrarse al final de la discusión o como un apartado diferente.

REFERENCIAS

La bibliografía incluye todas las fuentes consultadas en el artículo. Constituye un grupo de datos precisos y detallados para la identificación de una fuente documental impresa o digital de la cual se obtuvo la información, testifican y autentifican los datos no originales del trabajo y proveen al lector la bibliografía referente. Las referencias se enumeran y presentan en el orden en el que se citan en el texto. Cada revista puede tener un estilo de referencia específico, como el estilo Vancouver o American PsychologicalAssociation (APA).

5. IMPORTANCIA DE COMPRENDER LA ESTRUCTURA DE UN ARTÍCULO CIENTÍFICO

Comprender la estructura de un artículo es fundamental para la lectura crítica, ya que facilita localizar la información relevante y evaluar la calidad de cada sección. Una revisión exhaustiva de todas las partes del artículo permite al lector determinar si el diseño y los hallazgos cumplen con los estándares metodológicos necesarios para ser aplicables en la práctica clínica.

5.1 APLICACIÓN DE LA INTELIGENCIA ARTIFICIAL EN LA LECTURA DE ARTÍCULOS CIENTÍFICOS

El uso de la inteligencia artificial ha revolucionado la forma en que los profesionales de la salud acceden, organizan y analizan la información científica. Herramientas como *ChatGPT*, *Elicit*, *Scite*, *Consensus*, *ResearchRabbit* y otras plataformas permiten a los médicos e investigadores analizar rápidamente la estructura, contenido y calidad de los artículos científicos, incluso sin ser expertos en estadística o metodología.

No obstante, el uso de la IA debe ser complementario, y no reemplaza el juicio clínico ni el análisis crítico humano. El lector debe confirmar que las interpretaciones generadas por la IA sean coherentes con el contenido original, y evitar el uso acrítico de estos resúmenes automatizados.

6. CONCLUSIONES

La estructura de un artículo científico no es meramente estética ni convencional, sino que responde a una lógica que permite la transparencia, reproducibilidad y comprensión del proceso de investigación. Conocer cada una de sus partes es el primer paso para desarrollar una lectura crítica, saber dónde buscar respuestas a nuestras preguntas clínicas, y evaluar si un estudio está bien diseñado y ejecutado. Dominar esta estructura facilitará al lector aplicar los hallazgos de manera segura y razonada en su práctica médica.

7. PREGUNTAS DE AUTOEVALUACIÓN

PREGUNTAS DE OPCIÓN MÚLTIPLE

1. ¿Cuál es el objetivo principal del resumen (abstract) de un artículo científico?a) Presentar las limitaciones del estudiob) Resumir los resultados de forma detalladac) Facilitar una evaluación rápida de la relevancia del estudio d) Introducir la hipótesis nula

2. ¿En qué parte del artículo se deben encontrar los criterios de inclusión y exclusión?a) Resultadosb) Discusiónc) Métodos d) Introducción

3. ¿Qué elemento no forma parte del formato IMRyD?a) Introducciónb) Métodosc) Resultadosd) Referencias

https://doi.org/10.62486/978-9915-9928-3-9.Ch02

4. ¿Cuál es la utilidad principal de las palabras clave?a) Evaluar el rigor científico del artículob) Resumir los resultadosc) Facilitar la búsqueda e indexación del artículo d) Indicar los hallazgos clínicos principales

5. ¿Cuál de los siguientes enunciados es verdadero sobre la discusión?a) Se deben evitar comparaciones con otros estudiosb) Se debe evitar mencionar las limitacionesc) Se permite interpretar los resultados en el contexto clínico d) Debe repetir todos los datos estadísticos

PREGUNTAS ABIERTAS:

6. ¿Por qué es importante que el lector conozca la estructura IMRyD antes de realizar una lectura crítica?

7. ¿Qué elementos deben incluirse en la sección de métodos para garantizar la reproducibilidad?

8. ¿Cómo se relaciona el objetivo del estudio con la redacción de la introducción y la discusión?

9. ¿Qué características debe tener un título científico para ser considerado adecuado?

10. ¿Por qué no es recomendable interpretar los resultados desde el resumen?

RESPUESTAS:

1. c) Facilitar una evaluación rápida de la relevancia del estudio: El resumen debe ofrecer una visión general del estudio (objetivo, métodos, resultados, conclusiones) que permita al lector decidir si es pertinente leer el artículo completo.
2. c) Métodos: Los criterios de selección son parte esencial del diseño metodológico y permiten juzgar la validez interna y la aplicabilidad de los resultados.
3. d) Referencias: IMRyD incluye Introducción, Métodos, Resultados y Discusión. Las referencias son un elemento adicional, aunque imprescindible.
4. c) Facilitar la búsqueda e indexación del artículo. Las palabras clave, especialmente si están en MeSH/DeCS, permiten que el artículo sea localizable en bases de datos científicas.
5. c) Se permite interpretar los resultados en el contexto clínico. La discusión es el espacio para interpretar los hallazgos, compararlos con la literatura, reconocer limitaciones y proponer implicancias clínicas o futuras líneas de investigación.
6. Porque esta estructura organiza la información de manera sistemática y lógica, permitiendo al lector saber dónde buscar los objetivos, los métodos, los resultados y las interpretaciones. Conocer esta organización facilita el análisis crítico, mejora la eficiencia de lectura y ayuda a detectar posibles sesgos o fallas metodológicas.
7. Deben incluirse: tipo de estudio, población y muestra, criterios de inclusión/exclusión, técnicas de recolección de datos, descripción de intervenciones (si las hay), definición operacional de variables, métodos de análisis estadístico, consideraciones éticas y cualquier software o instrumento utilizado. Cuanto más detallada sea esta sección, mayor posibilidad de reproducir el estudio en otro contexto.
8. En la introducción, el objetivo debe ser el resultado lógico de una revisión breve del problema y sus vacíos de conocimiento. En la discusión, el objetivo debe retomarse para valorar si fue respondido y cómo se interpretan los hallazgos a la luz de ese propósito. Ambos extremos del artículo deben alinearse con el objetivo central.
9. Debe ser breve, claro, específico, sin abreviaturas innecesarias, y reflejar con precisión el contenido del artículo (población, intervención, comparación, desenlace, diseño del estudio, según corresponda). Un buen título atrae al lector y mejora la visibilidad del artículo en buscadores.
10. Porque el resumen es una síntesis y puede omitir detalles importantes del diseño, limitaciones o contexto. Además, al no incluir toda la información metodológica o estadística, se corre el riesgo de una interpretación sesgada o incompleta. La evaluación crítica debe hacerse sobre el artículo completo.

https://doi.org/10.62486/978-9915-9928-3-9.Ch02

REFERENCIAS

Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021 Mar 29;n71. Available at: http://dx.doi.org/10.1136/bmj.n71

Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, The SPIRIT-AI and CONSORT-AI Working Group, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020 Sep;26(9):1364–74. Available at: https://doi.org/10.1038/s41591-020-1034-x

Greenhalgh T. How to Read a Paper: The Basics of Evidence-Based Medicine and Healthcare. 6th ed. Newark: John Wiley & Sons, Incorporated; 2019.https://content.e-bookshelf.de/media/reading/L-12549069-30fab85148.pdf

Guyatt G, Rennie D, Meade MO, Cook DJ, editors. Users' guides to the medical literature: a manual for evidence-based clinical practice. 3rd edition. New York: McGraw-Hill Education; 2015.

Straus SE, Glasziou P, Haynes RB, Richardson WS. Evidence-based medicine: how to practice and teach EBM. Fifth edition. Edinburgh London New York: Elsevier; 2019.

Day RA, Gastel B. How to write and publish a scientific paper. 8th ed. Santa Barbara: Greenwood; 2016.

EQUATOR Network. Enhancing the quality and transparency of health research [Internet]. Oxford: Centre for Statistics in Medicine, University of Oxford; 2008 [cited 2025 Jan 28]. Available at: https://www.equator-network.org/

Fletcher GS. Clinical epidemiology: the essentials. 6th edition. Philadelphia: Wolters Kluwer; 2021.

International Committee of Medical Journal Editors (ICMJE). Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals [Internet]. Updated 2023 [cited 2025 Jan 28]. Available at: http://www.icmje.org/

Peat J, Elliott E, Baur L, Keena V. Scientific Writing Easy when you know how [Internet]. 1st edn. Wiley; 2002 [cited 2026 Jan 10]. Available at: https://doi.org/10.1002/9781118708019

https://doi.org/10.62486/978-9915-9928-3-9.Ch02

Chapter 3 / Capítulo 3

Critical reading of scientific articles on health: A practical guide (Spanish Edition)
ISBN: 978-9915-9928-3-9
DOI: 10.62486/978-9915-9928-3-9.Ch03
Pages: 16-26

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Introduction to Evidence-Based Medicine

Introducción a la Medicina Basada en la Evidencia

Jaykel Evelio Gómez Triana
Paul Cardozo-Gil
Jhossmar Cristians Auza-Santivañez

1 Centro de Salud Nuestra Señora del Castillo. Sevilla, España.
2 Clinica de las Américas. Hospital Obrero Nro3, Caja Nacional de Salud. Santa Cruz. Bolivia.
3 Hospital del Gran Chaco "Fray Quebracho" de III nivel. Instituto Académico Científico Quispe-Cornejo. Tarija, Bolivia.

ABSTRACT

Introduction: Evidence-Based Medicine (EBM) is a clinical practice model that integrates the best available scientific evidence, the clinical expertise of the practitioner, and the patient's values and preferences to optimize healthcare decision-making. Its development arose in response to the need to reduce clinical variability and promote interventions supported by high-quality scientific information. Development: This chapter presents the conceptual and historical foundations of EBM, highlighting the contributions of Archie Cochrane and the consolidation of the movement at McMaster University. It describes its three essential pillars: scientific evidence, clinical expertise, and patient preferences. The five fundamental steps of EBM are analyzed: formulating structured clinical questions using the PICO model, systematically searching the scientific literature, critically appraising the evidence, applying the findings to the clinical context, and continuously evaluating the results obtained. The main advantages and limitations of EBM are also examined, as well as the hierarchy of scientific evidence and the GRADE system for assessing the quality of evidence and the strength of recommendations. Finally, the emerging role of artificial intelligence as a support tool for the search, selection, synthesis, and interpretation of scientific information is addressed. Conclusion: Evidence-based medicine (EBM) is an essential competency for healthcare professionals, as it allows them to transform scientific information into well-founded clinical decisions. More than a method, it represents a critical and reflective approach that requires questioning, analyzing, and applying available knowledge with scientific rigor, always aimed at providing safer, more effective care centered on the real needs of each patient.

Keywords: Evidence-Based Medicine; clinical decision-making; critical appraisal of scientific literature; levels of evidence; artificial intelligence

RESUMEN

Introducción. La Medicina Basada en la Evidencia (MBE) es un modelo de práctica clínica que integra la mejor evidencia científica disponible, la experiencia clínica del profesional y los valores y preferencias del paciente para optimizar la toma de decisiones sanitarias. Su desarrollo surge como respuesta a la necesidad de reducir la variabilidad clínica y promover intervenciones sustentadas en información científica de calidad. Desarrollo. El capítulo presenta los fundamentos conceptuales e históricos de la MBE, destacando las contribuciones de Archie Cochrane y la consolidación del movimiento en la Universidad McMaster. Se describen sus tres pilares esenciales: evidencia científica, experiencia clínica y preferencias del paciente. Se analizan los cinco pasos fundamentales de la MBE: formulación de preguntas clínicas estructuradas mediante el modelo PICO, búsqueda sistemática de la literatura científica, evaluación crítica de la evidencia, aplicación de los hallazgos al contexto clínico y evaluación continua de los resultados obtenidos. Se examinan además las principales ventajas y limitaciones de la MBE, así como la jerarquía de la evidencia científica y el sistema GRADE para valorar la calidad de la evidencia y la fuerza de las recomendaciones. Finalmente, se aborda el papel emergente de la inteligencia artificial como herramienta de apoyo para la búsqueda, selección, síntesis e interpretación de la información científica. Conclusión. La MBE constituye una competencia esencial para los profesionales de la salud, ya que permite transformar la información científica en decisiones clínicas fundamentadas. Más que un método, representa una actitud crítica y reflexiva que exige cuestionar, analizar y aplicar el conocimiento disponible con rigor científico, siempre orientado a brindar una atención más segura, efectiva y centrada en las necesidades reales de cada paciente.

Palabras clave: Medicina Basada en la Evidencia; toma de decisiones clínicas; lectura crítica de la literatura científica; niveles de evidencia; inteligencia artificial

1. CASO CLÍNICO

El Dr. Rojas, médico residente de Medicina Interna, realiza su guardia en uno de los hospitales de referencia de la ciudad de La Paz. Durante la evaluación clínica de la señora Martha, una paciente de 67 años con antecedentes de hipertensión arterial adecuadamente controlada y dislipidemia en tratamiento, identifica un diagnóstico reciente de fibrilación auricular no valvular. Tras completar la valoración clínica y estratificar el riesgo tromboembólico mediante la escala CHA_2DS_2-VASc, obtiene una puntuación de 3 puntos, lo que indica un riesgo elevado de evento cerebrovascular y una clara indicación de terapia anticoagulante. Asimismo, no identifica contraindicaciones mayores para el inicio de anticoagulación oral. Al revisar los protocolos terapéuticos habitualmente utilizados en el hospital, observa que la warfarina continúa siendo el anticoagulante prescrito con mayor frecuencia como tratamiento de primera línea. Sin embargo, recuerda haber leído recientemente diversos estudios que sugieren que los anticoagulantes orales directos (ACOD), como apixabán y rivaroxabán, podrían ofrecer ventajas en términos de eficacia, seguridad y facilidad de uso en comparación con los antagonistas de la vitamina K.

Ante esta situación, surge una incertidumbre clínica relevante: ¿debería mantener la conducta terapéutica tradicional basada en warfarina o considerar el uso de un anticoagulante oral directo como alternativa terapéutica? Durante la consulta, la paciente expresa una inquietud frecuente en la práctica clínica: “Doctor, ¿cuál es el tratamiento más efectivo para prevenir un accidente cerebrovascular y que, al mismo tiempo, tenga el menor riesgo posible de sangrado u otras complicaciones?”

Consciente de la importancia de fundamentar sus decisiones en la mejor evidencia científica disponible, el Dr. Rojas decide formular una pregunta clínica estructurada e iniciar una búsqueda sistemática de la literatura biomédica. Su objetivo es identificar y evaluar críticamente los estudios más relevantes que comparen la eficacia y seguridad de los anticoagulantes orales directos frente a la warfarina en pacientes con fibrilación auricular no valvular, con el fin de proporcionar una recomendación terapéutica informada, individualizada y centrada en las preferencias y necesidades de la paciente.

2. OBJETIVOS

- Definir los principios fundamentales de la Medicina Basada en la Evidencia (MBE) y reconocer sus tres pilares esenciales.
- Describir los principales antecedentes históricos y la evolución de la MBE en la práctica clínica.
- Aplicar los cinco pasos de la MBE, incluyendo la formulación de preguntas clínicas mediante el modelo PICO.
- Identificar los niveles de evidencia, así como las principales ventajas y limitaciones de la MBE.
- Reconocer el papel de la inteligencia artificial como herramienta de apoyo en la búsqueda y aplicación de la evidencia científica.

3. INTRODUCCIÓN
¿QUÉ ES LA MEDICINA BASADA EN LA EVIDENCIA?

La Medicina Basada en la Evidencia (MBE) se definió inicialmente como el uso consciente, explícito y juicioso de la mejor evidencia científica disponible para la toma de decisiones relacionadas con la atención de los pacientes. Con el tiempo, este concepto evolucionó hacia un modelo más integral que reconoce la importancia de combinar la evidencia científica de mayor calidad con la experiencia clínica del profesional de la salud y los valores, preferencias y circunstancias particulares del paciente.

En la actualidad, la MBE constituye un enfoque fundamental para la práctica clínica moderna, ya que promueve la toma de decisiones informadas, transparentes y centradas en el paciente. Su objetivo es optimizar la calidad de la atención sanitaria, mejorar los resultados clínicos y favorecer el uso racional de las intervenciones diagnósticas, terapéuticas y preventivas sustentadas en la mejor evidencia disponible.

4. DEFINICIÓN Y COMPONENTES DE LA MBE

En la práctica clínica, la MBE se apoya en:

https://doi.org/10.62486/978-9915-9928-3-9.Ch03

- La mejor evidencia científica, basada en investigaciones bien diseñadas y rigurosamente evaluadas.
- La experiencia clínica del profesional de la salud, que ayuda a interpretar la evidencia en el contexto de cada paciente.
- Las preferencias y valores del paciente, asegurando que las decisiones médicas sean acordes a sus expectativas y prioridades.

La MBE intenta evitar intervenciones basadas en suposiciones o tradiciones sin fundamento, promoviendo decisiones respaldadas por evidencia sólida.

4.1 PROCESO SISTEMÁTICO

La MBE sigue un proceso estructurado que permite identificar y aplicar la mejor evidencia disponible. Este proceso es esencial en un contexto donde la cantidad de literatura científica crece exponencialmente y la variabilidad en la práctica médica sigue siendo alta. Con la MBE, los médicos se convierten en "consumidores inteligentes" de la literatura científica, desarrollando habilidades para seleccionar y aplicar los estudios relevantes en la práctica clínica.

5. HISTORIA DE LA MBE

Los orígenes de la MBE están ligados al epidemiólogo británico Archie Cochrane, quien, en 1972, publicó su influyente libro "*Effectiveness and Efficiency: Random Reflectionson Health Services*". Cochrane abogaba por el uso de ensayos clínicos aleatorizados como base para evaluar la eficacia de las intervenciones médicas y mejorar las decisiones sanitarias. Sus ideas llevaron al desarrollo de revisiones sistemáticas, resúmenes críticos de los estudios más relevantes, que permiten a los profesionales tomar decisiones informadas.El término "Evidence-Based Medicine" (Medicina Basada en la Evidencia) se originó en Canadá en la década de 1980, impulsado por la McMaster Medical School. En 1992, el concepto se formalizó en la revista *JAMA*, donde se definió como el "uso consciente, explícito y juicioso de las mejores pruebas disponibles para la atención integral de cada paciente"

6. CINCO PASOS DE LA MBE

Implementar la MBE requiere una serie de pasos sistemáticos que permitan al médico integrar la mejor evidencia científica en la atención de los pacientes. Cada uno de estos pasos es crucial, ya que omitir o apresurar alguna etapa puede comprometer la calidad de la atención. A continuación, se describen los cinco pasos de la MBE, que constituyen la base para tomar decisiones clínicas informadas y fundamentadas en evidencia.

PASO 1: FORMULAR UNA PREGUNTA CLÍNICA ESTRUCTURADA

La primera etapa de la MBE implica transformar una necesidad de información en una pregunta específica y estructurada, lo cual facilita encontrar respuestas precisas en la literatura científica. El modelo PICO (Paciente, Intervención, Comparación, Outcome o Resultado) es una herramienta ampliamente utilizada para formular preguntas clínicas efectivas.

Para una mejor comprensión se brinda el siguiente ejemplo:

Necesidad de información: necesito saber si los anticoagulantes directos son mejores que el tratamiento con Warfarina para evitar eventos tromboembólicos y con menor riesgo de sangrado en pacientes adultos con fibrilación auricular no valvular.

- P (Paciente): Pacientes adultos con fibrilación auricular no valvular y riesgo moderado-alto de evento tromboembólico (CHADS2-VASc ≥ 2).
- I (Intervención): Anticoagulantes orales directos (apixaban o rivaroxaban).
- **C (Comparación)**: Tratamiento con Warfarina.
- **(Outcome)**: Eficacia de prevención eventos tromboembólicos y seguridad en términos de reducción de sangrado.

Pregunta PICO completa: En pacientes adultos con fibrilación auricular no valvular y riesgo moderado-alto de ACV, ¿los anticoagulantes orales directos (como apixabán o rivaroxabán) son más eficaces y seguros que la warfarina para prevenir eventos tromboembólicos y reducir el

https://doi.org/10.62486/978-9915-9928-3-9.Ch03

riesgo de sangrado mayor?

PASO 2: BUSCAR LA MEJOR EVIDENCIA DISPONIBLE

Una vez formulada la pregunta, el siguiente paso es rastrear la evidencia científica que puede responder, empleando bases de datos confiables como PubMed, Cochrane Library, Scopus, Web ofScience, entre otros, que son referencias a nivel mundial; también existen bases de datos como SciELO (Scientific Electronic Library Online), que incluye investigaciones de la región, y LILACS (Literatura Latinoamericana y del Caribe en Ciencias de la Salud), que recopila literatura científica relevante para la práctica médica en América Latina. Es importante utilizar estrategias de búsqueda eficientes y enfocarse en estudios de alta calidad.

PASO 3: EVALUAR CRÍTICAMENTE LA EVIDENCIA

Este paso implica la evaluación sistemática de los estudios encontrados para determinar su validez, relevancia y aplicabilidad. La lectura crítica es esencial aquí, ya que permite identificar sesgos, limitaciones y fortalezas de cada estudio, asegurando que los resultados sean confiables y útiles para la toma de decisiones clínicas.

PASO 4: APLICAR LA EVIDENCIA A LA PRÁCTICA CLÍNICA

En este paso, el médico combina la evidencia encontrada con su experiencia y con las preferencias y particularidades del paciente. Es aquí donde la MBE se convierte en una práctica centrada en el paciente, ya que la evidencia seleccionada se adapta al contexto específico del paciente, haciendo que la decisión médica sea no solo fundamentada, sino también personalizada.

PASO 5: EVALUAR LOS RESULTADOS Y BUSCAR FORMAS DE MEJORA

Finalmente, es crucial evaluar la efectividad y eficiencia de la aplicación de la MBE en la práctica clínica. Al hacer una revisión de los resultados obtenidos, el médico puede identificar áreas de mejora para futuros casos y perfeccionar continuamente su habilidad en la lectura crítica y en la toma de decisiones clínicas basadas en evidencia.

7. VENTAJAS Y DESVENTAJAS DE LA MBE

Ventajas

- Disminuye la variabilidad en la práctica clínica: La MBE reduce la disparidad en los tratamientos al basarse en evidencia sólida, promoviendo un enfoque estandarizado.
- Incluye las preferencias del paciente: Al considerar los valores del paciente, la MBE asegura decisiones clínicas más informadas y personalizadas.
- Fomenta el autoaprendizaje: La MBE impulsa a los médicos a mantenerse actualizados y a mejorar su capacidad de interpretación crítica.
- Promueve una medicina actualizada y eficiente: Evita el uso de tratamientos obsoletos y promueve la adopción de intervenciones basadas en los últimos avances científicos.

Desventajas

- Requiere infraestructura y tiempo: La implementación de la MBE depende de recursos como acceso a bases de datos y tiempo para la lectura crítica de la literatura.
- Demanda habilidades de investigación: Para la lectura crítica, los médicos necesitan conocimientos en metodología de investigación y estadística.
- Riesgo de excesivo énfasis en metodologías específicas: Existe el riesgo de depender demasiado de ciertos tipos de evidencia, como los ensayos clínicos, sin considerar otros factores contextuales importantes.

8. NIVELES DE EVIDENCIA Y JERARQUÍA DE ESTUDIOS

No toda la evidencia científica tiene el mismo valor para la toma de decisiones clínicas. La

https://doi.org/10.62486/978-9915-9928-3-9.Ch03

Medicina Basada en la Evidencia reconoce que los diferentes diseños de estudio aportan grados variables de certeza, por lo que se han desarrollado sistemas de clasificación que permiten jerarquizar la evidencia según su capacidad para responder de manera válida y confiable a una pregunta clínica específica.

Tradicionalmente, las revisiones sistemáticas y los metaanálisis de ensayos clínicos aleatorizados se sitúan en la cúspide de la jerarquía de evidencia, ya que integran resultados de múltiples estudios, reducen el error aleatorio y permiten estimar con mayor precisión el efecto de una intervención. Inmediatamente después se encuentran los ensayos clínicos aleatorizados individuales, considerados el estándar de referencia para evaluar la eficacia y seguridad de intervenciones terapéuticas debido a su capacidad para minimizar sesgos mediante la aleatorización y el control.

En niveles intermedios se ubican los estudios observacionales analíticos, como los estudios de cohortes y los estudios de casos y controles. Aunque estos diseños no asignan de manera aleatoria la exposición, son fundamentales para evaluar asociaciones, identificar factores de riesgo, estudiar desenlaces poco frecuentes o analizar efectos a largo plazo, especialmente cuando los ensayos clínicos no son éticos o factibles. Por debajo de ellos se encuentran los estudios descriptivos, incluidos los estudios transversales, las series de casos y los reportes de casos, que si bien tienen menor capacidad para establecer causalidad, cumplen un rol importante en la generación de hipótesis y en la descripción inicial de nuevas enfermedades o fenómenos clínicos.

En la base de la jerarquía se sitúan la opinión de expertos, los consensos no sistemáticos y la experiencia clínica aislada. Aunque estos elementos pueden ser útiles en contextos de escasez de evidencia, son los más susceptibles a sesgos y deben interpretarse con cautela.

Es importante destacar que la jerarquía de evidencia no debe interpretarse de forma rígida o automática. Un estudio ubicado en un nivel alto puede presentar limitaciones metodológicas significativas, mientras que un estudio observacional bien diseñado puede ofrecer información clínicamente relevante y de alta calidad. Por ello, la lectura crítica y la evaluación metodológica individual de cada estudio siguen siendo esenciales, independientemente de su posición teórica en la jerarquía.

Figura 1. Jerarquía de la evidencia científica en Medicina Basada en la Evidencia

Fuente: Elaboración propia

https://doi.org/10.62486/978-9915-9928-3-9.Ch03

Con el objetivo de superar las limitaciones de los esquemas tradicionales, el sistema GRADE (GradingofRecommendationsAssessment, Development and Evaluation) propone un enfoque más integral para evaluar la calidad de la evidencia y la fuerza de las recomendaciones. Este sistema no se basa únicamente en el diseño del estudio, sino que considera múltiples dominios, como el riesgo de sesgo, la consistencia de los resultados, la precisión de las estimaciones, la aplicabilidad clínica y la probabilidad de sesgo de publicación. De esta manera, GRADE permite que la calidad de la evidencia se clasifique como alta, moderada, baja o muy baja, ofreciendo una valoración más contextualizada y útil para la toma de decisiones clínicas.

Tabla 1. Clasificación de la evidencia según el sistema GRADE

Calidad de la evidencia (GRADE)	Definición	Diseños de estudio habituales	Interpretación clínica
Alta	Existe alta confianza en que el efecto estimado es cercano al efecto real. Es poco probable que nuevas investigaciones cambien la estimación.	Ensayos clínicos aleatorizados bien diseñados, revisiones sistemáticas de ECA sin limitaciones importantes	Las decisiones clínicas pueden basarse con seguridad en esta evidencia
Moderada	Existe confianza moderada en el efecto estimado; investigaciones futuras podrían modificar la estimación o la confianza en ella.	Ensayos clínicos con algunas limitaciones metodológicas, estudios observacionales con gran tamaño de efecto	Las recomendaciones son razonablemente firmes, pero deben revisarse con nueva evidencia
Baja	La confianza en el efecto estimado es limitada; el efecto real podría ser sustancialmente diferente.	Estudios observacionales bien realizados, ECA con limitaciones importantes	Las decisiones deben tomarse con cautela y considerando el contexto clínico
Muy baja	Existe muy poca confianza en el efecto estimado; el efecto real probablemente sea muy diferente.	Estudios observacionales con alto riesgo de sesgo, series de casos, opinión de expertos	La evidencia es insuficiente para respaldar recomendaciones firmes

Fuente: Elaboración propia

A diferencia de la jerarquía clásica de evidencia, el sistema GRADE no clasifica los estudios únicamente por su diseño, sino que evalúa de forma integral la calidad de la evidencia disponible para cada desenlace clínico. Esto permite que estudios observacionales bien realizados puedan alcanzar una calidad moderada o incluso alta, y que ensayos clínicos aleatorizados con limitaciones importantes sean reclasificados a niveles inferiores.

En síntesis, comprender los niveles de evidencia y la jerarquía de estudios permite al médico priorizar fuentes confiables, interpretar adecuadamente los resultados de la investigación y aplicar la evidencia de manera crítica y responsable en la práctica clínica, siempre considerando el contexto del paciente y las circunstancias específicas de atención.

9. MEDICINA BASADA EN LA EVIDENCIA E INTELIGENCIA ARTIFICIAL

La incorporación de la inteligencia artificial (IA) en el ámbito de la MBE constituye uno de los avances más relevantes de la práctica médica actual. Su valor principal radica en potenciar las capacidades del profesional, optimizando los procesos de búsqueda, análisis e interpretación de la literatura científica. En un contexto donde la producción de artículos biomédicos aumenta de manera exponencial, la IA se convierte en un aliado para gestionar grandes volúmenes de información y apoyar la toma de decisiones clínicas fundamentadas.

En la etapa inicial del proceso de MBE, la IA puede mejorar notablemente la formulación de preguntas clínicas estructuradas. A través de modelos de lenguaje como ChatGPT, el clínico puede transformar una cuestión narrativa en una pregunta PICO clara y orientada, lo cual facilita la búsqueda de información pertinente en bases de datos biomédicas. Asimismo, la IA es capaz de identificar automáticamente los términos MeSH y DeCS más apropiados para cada componente de la pregunta, evitando omisiones frecuentes durante la estrategia de búsqueda y mejorando la sensibilidad y especificidad de la misma.

En cuanto a la localización y selección de la evidencia, la IA permite automatizar tareas tradicionalmente demandantes. Herramientas como Elicit o plataformas de aprendizaje automático pueden cribar cientos de resúmenes y clasificarlos según relevancia, tipo de estudio, nivel de evidencia o población estudiada. Este filtrado preliminar, aunque siempre sujeto a validación humana, reduce de manera significativa el tiempo invertido en revisar literatura y permite enfocarse rápidamente en los estudios más útiles para responder la pregunta clínica.

La lectura crítica, una de las competencias centrales de la MBE, también puede beneficiarse de la IA. Muchos modelos son capaces de identificar de forma automática elementos clave del

https://doi.org/10.62486/978-9915-9928-3-9.Ch03

diseño metodológico, resaltar aspectos del análisis estadístico y señalar posibles fuentes de sesgo. Aunque estas funciones no reemplazan el juicio del profesional, sí actúan como una guía inicial que facilita la comprensión del artículo y orienta la revisión detallada. La IA también puede generar resúmenes estructurados que permiten obtener una visión general del estudio antes de profundizar en cada sección, además de comparar hallazgos entre distintos artículos cuando el clínico necesita evaluar la consistencia de los resultados.

Un aspecto cada vez más relevante es la capacidad de la IA para traducir los resultados estadísticos a un lenguaje clínico comprensible y útil. Por ejemplo, puede convertir medidas como el riesgo relativo en riesgo absoluto o en números necesarios a tratar (NNT), permitiendo una interpretación más práctica y orientada al paciente. De esta manera, la IA contribuye a mejorar la aplicabilidad clínica de los hallazgos, y ayuda al médico a valorar la magnitud real del beneficio o del daño asociado a una intervención.

A pesar de sus ventajas, el uso de IA en la MBE exige un enfoque crítico y ético. Los modelos dependen de la calidad de la información con la que fueron entrenados, y pueden reproducir sesgos presentes en los datos originales. Además, muchos algoritmos funcionan como "cajas negras", donde los criterios utilizados para generar una respuesta no siempre son transparentes ni verificables. Por ello, el profesional debe mantener una postura vigilante, validar la información proporcionada por la IA y recordar que ninguna herramienta tecnológica sustituye la experiencia clínica ni la comprensión metodológica del lector crítico.

En síntesis, la inteligencia artificial representa un recurso de alto valor para apoyar el proceso de la Medicina Basada en la Evidencia, siempre que se utilice de manera responsable y complementaria al pensamiento clínico. Su mayor fortaleza no es reemplazar al médico, sino potenciar su capacidad para acceder, analizar y aplicar la evidencia científica de forma más rápida, rigurosa y personalizada para cada paciente.

Tabla 2. Aplicación de la inteligencia artificial en el proceso de la Medicina Basada en la Evidencia

Etapa del proceso de la MBE	Aporte de la inteligencia artificial	Valor para la práctica clínica	Consideraciones críticas
Formulación de la pregunta clínica	Transformación de preguntas narrativas en preguntas estructuradas (PICO). Identificación automática de términos MeSH y DeCS.	Mejora la claridad de la pregunta clínica y optimiza la estrategia de búsqueda bibliográfica.	La pregunta debe ser validada por el clínico para asegurar relevancia clínica real.
Búsqueda de la evidencia	Construcción de estrategias de búsqueda, uso de operadores booleanos y filtrado por tipo de estudio o nivel de evidencia.	Aumenta la sensibilidad y especificidad de la búsqueda y reduce el riesgo de omitir estudios relevantes.	Puede depender de la calidad del prompt o de los datos de entrenamiento.
Selección y cribado de estudios	Clasificación automatizada de títulos y resúmenes según relevancia, población, intervención o diseño del estudio.	Reduce de forma significativa el tiempo invertido en revisar grandes volúmenes de literatura.	El cribado inicial no sustituye la selección final por el investigador.
Lectura crítica del artículo	Identificación de elementos metodológicos clave, posibles sesgos y aspectos relevantes del análisis estadístico. Generación de resúmenes estructurados.	Facilita la comprensión inicial del estudio y orienta una lectura crítica más eficiente.	No reemplaza el análisis metodológico profundo del profesional.
Síntesis y comparación de evidencia	Comparación de resultados entre estudios y generación de síntesis narrativas preliminares.	Ayuda a evaluar consistencia y dirección de los resultados entre diferentes investigaciones.	Riesgo de simplificación excesiva si no se revisa críticamente.
Interpretación estadística	Traducción de medidas estadísticas a indicadores clínicos (riesgo absoluto, NNT, NNH).	Mejora la aplicabilidad clínica y la comunicación con el paciente.	Los cálculos deben ser corroborados con los datos originales.
Aplicación clínica	Adaptación de la evidencia al contexto del paciente considerando variables clínicas relevantes.	Apoya decisiones más personalizadas y centradas en el paciente.	No sustituye el juicio clínico ni la experiencia médica.
Aspectos éticos y limitaciones	Identificación potencial de inconsistencias, sesgos o limitaciones en los datos analizados.	Promueve un uso más reflexivo y responsable de la evidencia.	Riesgo de sesgos algorítmicos y falta de transparencia ("caja negra").

Fuente: Elaboración propia

10. CONCLUSIONES

La MBE representa un avance importante en la medicina moderna, ya que permite a los médicos tomar decisiones informadas basadas en evidencia sólida. Su implementación

https://doi.org/10.62486/978-9915-9928-3-9.Ch03

sistemática en la práctica clínica garantiza una atención de calidad, reduce la variabilidad en los tratamientos y fomenta una medicina centrada en el paciente. Aunque existen desafíos y limitaciones, la MBE sigue siendo una herramienta esencial para todos los profesionales de la salud comprometidos con brindar la mejor atención posible.

11. PREGUNTAS DE AUTOEVALUACIÓN

PREGUNTAS DE OPCIÓN MÚLTIPLE:

1. ¿Cuál es el objetivo de la Medicina Basada en la Evidencia? a) Usar la experiencia del médico sin basarse en la evidencia. b) Usar solo la evidencia científica sin considerar la opinión del paciente. c) Integrar la mejor evidencia científica, experiencia clínica y preferencias del paciente. d) Basar todas las decisiones en métodos tradicionales.

2. ¿Cuál de las siguientes opciones representa un componente del modelo PICO? a) Investigación. b) Contexto. c) Comparación. d) Opinión.

3. ¿Cuál de las siguientes bases de datos es específica para literatura en ciencias de la salud en América Latina? a) PubMed. b) LILACS. c) Cochrane. d) Google Scholar.

4. ¿Qué paso de la MBE implica evaluar si los resultados de un estudio son válidos y aplicables? a) Buscar la evidencia.

b) Formular la pregunta. c) Evaluar críticamente la evidencia. d) Aplicar la experiencia clínica.

5. ¿Qué representa el sistema GRADE dentro de la MBE? a) Un sistema para clasificar pacientes. b) Una guía para aplicar tratamientos tradicionales. c) Un método para valorar la calidad de la evidencia y la fuerza de las recomendaciones. d) Una base de datos médica.

PREGUNTAS ABIERTAS

6. Explique brevemente en qué consiste el modelo PICO y cómo ayuda en la formulación de preguntas clínicas.

7. Enumere y describa los cinco pasos de la MBE.

8. Mencione dos ventajas y dos desventajas de la MBE en la práctica clínica y explique cómo impactan en la atención al paciente.

9. ¿Qué rol cumple la inteligencia artificial en la aplicación moderna de la MBE y cuáles son sus limitaciones éticas?

10. Explique qué es el sistema GRADE y cómo puede influir en la toma de decisiones clínicas.

RESPUESTAS

1. c) Integrar la mejor evidencia científica, experiencia clínica y preferencias del paciente. La MBE se basa en un modelo tridimensional: evidencia científica, experiencia clínica y valores del paciente. Ninguno de estos elementos por sí solo es suficiente; la toma de decisiones clínicas requiere la integración equilibrada de los tres.

2. c) Comparación. El modelo PICO incluye Paciente/Problema, Intervención, Comparación y Outcome. La comparación es esencial para determinar si una intervención es superior, equivalente o inferior a otra. Las otras opciones no corresponden a ningún componente del modelo.

3. b) LILACS. LILACS es la base de datos más importante de literatura biomédica en América Latina y el Caribe. PubMed y Cochrane son bases internacionales, y Google Scholar no es una base especializada, por lo que no cumplen el criterio de especificidad regional.

4. c) Evaluar críticamente la evidencia. La evaluación crítica es el paso de la MBE en el que se determina si los resultados de un estudio son válidos, relevantes y aplicables. Los otros pasos se relacionan con la formulación de la pregunta, la búsqueda o la aplicación de la evidencia, pero no con el análisis de su calidad.

5. c) Un método para valorar la calidad de la evidencia y la fuerza de las recomendaciones. El sistema GRADE es el estándar internacional para jerarquizar la evidencia y establecer recomendaciones clínicas. No clasifica pacientes ni actúa como base de datos, y no se relaciona con tratamientos tradicionales. Su enfoque es evaluar cuánta confianza puede depositarse en los hallazgos disponibles.

6. El modelo PICO es una herramienta que permite estructurar preguntas clínicas de manera clara y orientada a la evidencia. Sus componentes son: Paciente o problema (P), Intervención

https://doi.org/10.62486/978-9915-9928-3-9.Ch03

(I), Comparación (C) y Outcome o desenlace clínico (O). Este formato delimita los elementos esenciales de la duda clínica y facilita la búsqueda eficiente de estudios relevantes en bases de datos científicas, reduciendo la ambigüedad y mejorando la precisión de la investigación.

7. Los cinco pasos son:

- Formular una pregunta clínica estructurada basada en una necesidad real de información.
- Buscar la mejor evidencia disponible utilizando estrategias sistemáticas y bases de datos confiables.
- Evaluar críticamente la validez, relevancia y aplicabilidad de los estudios encontrados.
- Aplicar la evidencia al contexto clínico individual, integrándola con la experiencia del médico y las preferencias del paciente.
- Evaluar los resultados obtenidos y buscar oportunidades de mejora para optimizar el proceso de toma de decisiones en futuros casos.

8. Entre sus ventajas destacan la reducción de la variabilidad clínica, que promueve decisiones más consistentes y alineadas con evidencia sólida, y la inclusión de las preferencias del paciente, lo que mejora la calidad de la relación terapéutica y la adherencia. En cuanto a sus desventajas, la MBE requiere tiempo y recursos para la búsqueda y lectura crítica, y demanda habilidades metodológicas que muchos profesionales deben desarrollar. Estas limitaciones pueden dificultar su aplicación, aunque su incorporación adecuada mejora significativamente la seguridad y eficacia del cuidado clínico.

9. La inteligencia artificial puede acelerar la búsqueda de literatura, optimizar estrategias de búsqueda, sintetizar información y apoyar la lectura crítica mediante la identificación de elementos metodológicos clave. Asimismo, facilita la interpretación estadística y la comparación de estudios. Sin embargo, presenta limitaciones éticas, como la posibilidad de replicar sesgos presentes en los datos, la falta de transparencia en los algoritmos y el riesgo de que los clínicos confíen excesivamente en sus resultados. Por ello, la IA debe utilizarse como apoyo y siempre bajo supervisión crítica del profesional.

10. El sistema GRADE es un método estandarizado que clasifica la calidad de la evidencia científica y la fuerza de las recomendaciones en guías clínicas. Considera aspectos como el diseño del estudio, el riesgo de sesgo, la consistencia de los resultados y su aplicabilidad. Su influencia en la práctica clínica es significativa, ya que orienta al profesional sobre cuánta confianza puede depositarse en la evidencia disponible y qué intervenciones son más adecuadas para determinadas condiciones clínicas.

https://doi.org/10.62486/978-9915-9928-3-9.Ch03

REFERENCIAS

Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021 Mar 29;n71. Available at: https://doi.org/10.1136/bmj.n71

Wallach JD, Boyack KW, Ioannidis JPA. Reproducible research practices, transparency, and open access data in the biomedical literature, 2015–2017. Dirnagl U, editor. PLoS Biol. 2018 Nov 20;16(11):e2006930. Available at: https://doi.org/10.1371/journal.pbio.2006930

Iqbal SA, Wallach JD, Khoury MJ, Schully SD, Ioannidis JPA. Reproducible Research Practices and Transparency across the Biomedical Literature. Vaux DL, editor. PLoS Biol. 2016 Jan 4;14(1):e1002333. Available at: https://doi.org/10.1371/journal.pbio.1002333

Cochrane Collaboration, Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, et al. Cochrane Handbook for Systematic Reviews of Interventions [Internet]. Cochrane Collaboration; 2021. (Cochrane book series). Available at: https://books.google.com.bo/books?id=sxdxzgEACAAJ

Khan H. Comment on "Improving the reporting and use of trial results in clinical trials registries: global practices, barriers, and recommendations." Journal of Clinical Epidemiology. 2025 Dec;188:111975. Available at: doi:10.1016/j.jclinepi.2025.111975

Glasziou P, Chalmers I, Altman DG, Bastian H, Boutron I, Brice A, et al. Taking healthcare interventions from trial to practice. BMJ. 2010 Aug 13;341(aug13 1):c3852–c3852. Available at: doi:10.1136/bmj.c3852

Ioannidis JPA, Fanelli D, Dunne DD, Goodman SN. Meta-research: Evaluation and Improvement of Research Methods and Practices. PLoS Biol. 2015 Oct 2;13(10):e1002264. Available at: doi:10.1371/journal.pbio.1002264

Murad MH, Wang Z. Guidelines for reporting meta-epidemiological methodology research. Evid Based Med. 2017 Aug;22(4):139–42. Available at: doi:10.1136/ebmed-2017-110713

Reeves BC, Deeks JJ, Higgins JP, Shea B, Tugwell P, Wells GA, et al. Including non-randomized studies on intervention effects. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al., editors. Cochrane Handbook for Systematic Reviews of Interventions [Internet]. 1st ed. Wiley; 2019 [cited 2026 Apr 1]. p. 595–620. Available at: doi:10.1002/9781119536604.ch24

Ioannidis JPA. Why Most Clinical Research Is Not Useful. PLoS Med. 2016 Jun 21;13(6):e1002049. Available at: https://doi.org/10.1371/journal.pmed.1002049

Chapter 4 / Capítulo 4

Critical reading of scientific articles on health: A practical guide (Spanish Edition)
ISBN: 978-9915-9928-3-9
DOI: 10.62486/978-9915-9928-3-9.Ch04
Pages: 27-40

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

The research process in the context of Evidence-Based Medicine

El proceso de investigación en el contexto de la Medicina Basada en la Evidencia

Freddy Ednildon Bautista-Vanegas
Micaela Mariel Moncada Mercado
Carlos Alberto Paz-Roman

1 Neurokliniken Beelitz GmbH Neurologische Rehabilitationsklinik: Beelitz Heilstätten, Brandenburg, DE. Germany.
2 Neurologisches Fachkrankenhaus für Bewegungsstörungen/Parkinson: Beelitz Heilstätten, Brandenburg, DE. Germany.
3 Universidad Técnica de Oruro. La Paz, Bolivia.
4 Hospital de Niños Dr. Mario Ortíz Suárez, Santa cruz Bolivia.

ABSTRACT

Introduction: The research process is the foundation for generating reliable scientific evidence in health and represents an essential component of Evidence-Based Medicine (EBM). Its objective is to answer clinical questions using a systematic, rigorous, and reproducible method that produces knowlede useful for healthcare decision-making. Development: This chapter describes the main stages of the research process, from defining the problem and formulating the research question to publishing the results. It highlights the importance of literature review, selecting the appropriate methodological design, defining the population and sample, accurately measuring variables, and analyzing and interpreting the data. It also presents the main classifications of epidemiological studies, including descriptive, analytical, observational, and experimental designs. Furthermore, it analyzes fundamental concepts related to research variables, Type I and Type II errors, the most frequent biases, and the principles of internal and external validity—essential elements for evaluating the methodological quality of a study. Finally, the role of artificial intelligence as a complementary tool for optimizing literature searches, methodological design, data analysis, and scientific writing is addressed. Conclusion: Understanding the research process allows for the critical interpretation of scientific evidence and the differentiation of reliable results from those affected by errors or biases. Beyond producing knowledge, research involves developing a critical attitude, where each clinical question becomes an opportunity to generate evidence that contributes to safer, more effective, and patient-centered care.

Keywords: Health research; study designs; research variables; bias; publication of articles

RESUMEN

Introducción: El proceso de investigación constituye el fundamento para la generación de evidencia científica confiable en salud y representa un componente esencial de la Medicina Basada en la Evidencia (MBE). Su objetivo es responder preguntas clínicas mediante un método sistemático, riguroso y reproducible que permita producir conocimiento útil para la toma de decisiones sanitarias. Desarrollo: El capítulo describe las principales etapas del proceso de investigación, desde el planteamiento del problema y la formulación de la pregunta de investigación hasta la publicación de los resultados. Se destaca la importancia de la revisión de literatura, la selección del diseño metodológico apropiado, la definición de la población y muestra, la medición adecuada de variables y el análisis e interpretación de los datos. De esta manera, se presentan las principales clasificaciones de los estudios epidemiológicos, incluyendo diseños descriptivos, analíticos, observacionales y experimentales. Se analizan además los conceptos fundamentales relacionados con las variables de investigación, los errores tipo I y tipo II, los sesgos más frecuentes y los principios de validez interna y externa, elementos indispensables para evaluar la calidad metodológica de un estudio. Finalmente, se aborda el papel de la inteligencia artificial como herramienta complementaria para optimizar la búsqueda bibliográfica, el diseño metodológico, el análisis de datos y la redacción científica. Conclusión: Comprender el proceso de investigación permite interpretar críticamente la evidencia científica y diferenciar resultados confiables de aquellos afectados por errores o sesgos. Más allá de producir conocimiento, investigar implica desarrollar una actitud crítica, donde cada pregunta clínica se convierte en una oportunidad para generar evidencia que contribuya a una atención más segura, efectiva y centrada en el paciente.

Palabras clave: Investigación en salud; diseños de estudio; variables de investigación; sesgo; publicación de artículos

1. CASO CLÍNICO

Durante su internado rotatorio en un hospital de segundo nivel de atención en una población rural, Ana observa que un número considerable de pacientes con hipertensión arterial presenta un control inadecuado de las cifras de presión arterial, a pesar de encontrarse bajo tratamiento farmacológico antihipertensivo. Esta situación despierta su interés, ya que reconoce que el control insuficiente de la hipertensión constituye un importante factor de riesgo para el desarrollo de eventos cardiovasculares, cerebrovasculares y renales.

Ante esta observación, surge una interrogante clínica relevante: ¿el mal control de la presión arterial se encuentra asociado principalmente a una baja adherencia terapéutica o intervienen otros factores, como los hábitos alimentarios, el estrés, el nivel de actividad física o determinadas características sociodemográficas y clínicas de los pacientes?

Como parte de las actividades académicas de su rotación, Ana debe desarrollar un proyecto de investigación. Con el propósito de responder esta pregunta, decide investigar los factores asociados al control inadecuado de la hipertensión arterial en pacientes ambulatorios atendidos en el hospital. Para ello, comprende que es necesario conocer y aplicar de manera rigurosa las diferentes etapas del proceso de investigación en salud, desde la formulación de la pregunta de investigación hasta la interpretación crítica de los resultados obtenidos.

2. OBJETIVOS

- Comprender las etapas principales del proceso de investigación en salud y cómo cada una contribuye a la generación de evidencia científica confiable.
- Reconocer la importancia de un proceso de investigación estructurado y riguroso para asegurar la validez, precisión y aplicabilidad de los resultados en la práctica clínica.
- Destacar la relevancia del proceso de revisión por pares y la publicación como herramientas para que la evidencia científica esté disponible para consulta, contribuyendo a la práctica de la Medicina Basada en la Evidencia.
- Fomentar una visión crítica del lector para que identifique cómo la calidad de un estudio puede afectar su decisión clínica.

3. INTRODUCCIÓN

La investigación en salud surge de la necesidad de responder interrogantes relacionadas con el diagnóstico, tratamiento, pronóstico y prevención de las enfermedades. Estas interrogantes se transforman en preguntas de investigación que son abordadas mediante un proceso sistemático, estructurado y metodológicamente riguroso, cuyo propósito es generar conocimiento científico válido, reproducible y aplicable a la práctica clínica. El proceso de investigación constituye la base para la producción de evidencia científica de calidad, permitiendo obtener resultados confiables que contribuyen a la toma de decisiones en el contexto de la Medicina Basada en la Evidencia (MBE). Su objetivo fundamental es ampliar el conocimiento disponible mediante la formulación de nuevas preguntas, la generación de hipótesis y la búsqueda sistemática de respuestas sustentadas en métodos científicos apropiados. Por esta razón, resulta indispensable comprender cada una de las etapas que conforman el proceso investigativo, desde la identificación del problema y la formulación de la pregunta de investigación hasta el análisis, interpretación y difusión de los resultados. La omisión o ejecución inadecuada de cualquiera de estas etapas puede comprometer la validez y la utilidad de los hallazgos obtenidos. El conocimiento del proceso de investigación permite desarrollar una lectura crítica más profunda de la literatura científica. Más allá de la interpretación de los resultados estadísticos, el lector debe ser capaz de evaluar la calidad metodológica del estudio, identificar posibles fuentes de error o sesgo y valorar la validez y aplicabilidad de sus conclusiones. En este sentido, el análisis estadístico representa únicamente una de las múltiples etapas que determinan la calidad global de una investigación científica.

https://doi.org/10.62486/978-9915-9928-3-9.Ch04

4. ETAPAS DEL PROCESO DE INVESTIGACIÓN

Fuente: Elaboración propia

4.1 PLANTEAMIENTO DEL PROBLEMA Y PREGUNTA DE INVESTIGACIÓN.

La investigación comienza con la definición clara de una pregunta clínica o problema, basado en la observación de un fenómeno o una necesidad en la práctica médica. Esta pregunta puede expresarse en forma de hipótesis o de objetivo específico, y debe estar fundamentada en una revisión inicial de la literatura.

4.2 REVISIÓN DE LITERATURA Y MARCO TEÓRICO.

La revisión de estudios previos ayuda a construir un marco teórico y contextualiza la pregunta de investigación. Este paso permite identificar vacíos en el conocimiento y formular hipótesis basadas en evidencia existente, en artículos de investigación se añade la información más relevante producto de esta revisión en el acápite de introducción y posteriormente se contrasta los resultados del estudio con la literatura revisada en el acápite de discusión.

4.3 DISEÑO DEL ESTUDIO.

Dependiendo de la pregunta de investigación y los objetivos, se selecciona el tipo de estudio más adecuado: descriptivo (para describir fenómenos) o analítico (para explorar asociaciones o relaciones causales). Además, se elige entre diseños transversales, longitudinales, observacionales o experimentales, y se estructura el estudio para minimizar sesgos y errores.

4.4 SELECCIÓN DE LA POBLACIÓN Y MUESTRA.

Se definen los criterios de inclusión y exclusión para la población de estudio y se calcula el tamaño de la muestra necesaria para asegurar resultados representativos y válidos.

4.5 MEDICIÓN DE VARIABLES Y ESTRATEGIA DE ANÁLISIS.

Las variables de interés se miden con instrumentos válidos y confiables. La elección de las variables (independientes, dependientes, y confusoras) y los métodos estadísticos adecuados son esenciales para interpretar los resultados correctamente y reducir el riesgo de sesgos.

https://doi.org/10.62486/978-9915-9928-3-9.Ch04

4.6 RECOLECCIÓN Y ORGANIZACIÓN DE DATOS.

Los datos se recogen y organizan en un protocolo estructurado que detalla cada paso, lo que minimiza los errores y asegura que los resultados sean consistentes y precisos.

4.7 ANÁLISIS E INTERPRETACIÓN DE RESULTADOS.

Se aplican métodos estadísticos para analizar los datos y se interpretan los hallazgos en el contexto del problema planteado. Este paso permite extraer conclusiones significativas y contextualizadas.

4.8 COMUNICACIÓN Y PUBLICACIÓN DE LOS RESULTADOS.

Finalmente, los resultados se publican en revistas científicas revisadas por pares, contribuyendo a la base de evidencia en salud y evitando el "conocimiento gris" (trabajos no publicados que no están disponibles para consulta pública).

5. PRINCIPALES TIPOS DE ESTUDIO EN INVESTIGACIÓN EN SALUD.

Los estudios en salud pueden clasificarse de acuerdo a varios criterios, lo cual permite a los investigadores seleccionar el diseño más adecuado para responder a la pregunta de investigación planteada.

5.1 CLASIFICACIÓN SEGÚN EL PROPÓSITO:

Descriptivos: Su objetivo es observar y describir fenómenos o características de una población, también reportan sucesos inusuales o nuevos sucesos que posteriormente ayudan a la generación de hipótesis.

Ejemplos de estudios:

- **Prevalencia.** Un estudio que estima cuántos adultos mayores en La Paz presentan Hipertensión Arterial Sistémica en el año 2025.
- **Serie de casos.** Informe clínico sobre cinco pacientes con enfermedad de Chagas con afectación gastrointestinal poco común en un hospital rural de Tarija.
- **Evaluación de prueba diagnóstica.** Estudio transversal para determinar la sensibilidad y especificidad de una nueva tira reactiva para detectar infección urinaria.
- **Incidencia.** Cálculo del número de nuevos casos de tuberculosis pulmonar en adolescentes entre enero y diciembre de 2025 en la ciudad de La Paz a 3600 m.s.n.m.
- **Descripción natural de una enfermedad.** Estudio que documenta la evolución clínica de pacientes con COVID-19 leve sin tratamiento antiviral.

Analíticos: Investigan relaciones o asociaciones causales entre variables, como el riesgo o el efecto de una exposición.

Ejemplos de estudios:

- **Estudio de Cohorte:** Se realiza el seguimiento de 500 pacientes con diagnóstico reciente de hipertensión arterial durante un periodo de 5 años. Al inicio del estudio, los participantes se clasifican en dos grupos según su exposición al hábito tabáquico: Expuestos: hipertensos fumadores activos; No expuestos: hipertensos que nunca han fumado o exfumadores de más de 10 años. Durante los 5 años de se estima la incidencia acumulada de enfermedad cardiovascular (ECV) en ambos grupos y determinar si el tabaquismo se asocia a un mayor riesgo de ECV en pacientes hipertensos.
- **Casos y controles:** Se selecciona un grupo de mujeres con cáncer de cuello uterino confirmado por biopsia (casos) y se compara con un grupo de mujeres sin la enfermedad con citología normal (controles). En ambos grupos se evalúa la presencia del VPH tipo 16 mediante pruebas moleculares, con el objetivo de determinar si este subtipo viral está más presente en las mujeres con cáncer.
- **Ensayo clínico:** Se realiza un ensayo clínico aleatorizado en adultos con obesidad sin diabetes, asignándolos al azar a recibir semaglutida o placebo, ambos con el mismo acompañamiento nutricional. Durante 6 a 12 meses se mide la pérdida de peso como desenlace principal. Al

https://doi.org/10.62486/978-9915-9928-3-9.Ch04

comparar los grupos, si quienes reciben semaglutida presentan una reducción de peso significativamente mayor que el grupo placebo, se concluye que el fármaco es eficaz para el manejo del peso en esta población.

5.2 CLASIFICACIÓN SEGÚN LA INTERVENCIÓN:

Observacionales: El investigador no interviene en el fenómeno estudiado, sino que se limita a observar y registrar los resultados, el factor de estudio no es controlado por los investigadores, pueden ser descriptivos o analíticos.

Ejemplos de estudios:

- **Transversal:** los investigadores recogen información en un único momento para evaluar si existe relación entre el nivel socioeconómico y el grado de control glucémico en pacientes con diabetes atendidos en centros de salud. Se mide simultáneamente la categoría socioeconómica y la hemoglobina glicosilada (HbA1c), permitiendo describir la situación actual y explorar asociaciones, aunque sin establecer causalidad.
- **Estudio de Cohorte:** se realiza un seguimiento de trabajadores expuestos a productos de la minería y se compara su evolución con trabajadores no expuestos durante varios años para observar la aparición de enfermedades respiratorias crónicas. Se registra la incidencia de síntomas y diagnósticos respiratorios, evaluando si la exposición laboral incrementa el riesgo de desarrollar estas patologías.
- **Casos y controles:** se seleccionan mujeres que han tenido un parto prematuro (casos) y se comparan con mujeres que tuvieron un parto a término (controles), con el objetivo de identificar factores de riesgo asociados a este desenlace. Se analizan antecedentes como infecciones urinarias, control prenatal insuficiente o tabaquismo, determinando si alguno de estos factores aparece con mayor frecuencia en las mujeres con parto prematuro.

Experimentales: El investigador aplica una intervención para observar sus efectos en la variable de interés, el factor de estudio es controlado de forma deliberada y estudia la relación causa-efecto, evalúa intervenciones terapéuticas o preventivas. Ejemplo: ensayos clínicos aleatorizados.

Ejemplos de estudio:

- **Ensayo clínico aleatorizado:** en un estudio multicéntrico se evalúa la eficacia de dexametasona versus betametasona para la maduración pulmonar fetal en embarazos pretérmino. Las gestantes elegibles son asignadas al azar a uno de los dos tratamientos y se les realiza seguimiento hasta el nacimiento para comparar desenlaces como síndrome de dificultad respiratoria neonatal, necesidad de ventilación mecánica y mortalidad perinatal.

5.3 CLASIFICACIÓN SEGÚN EL TIEMPO:

Transversales: Recogen datos en un solo momento temporal, proporcionando una "fotografía" de la situación, permite evaluar la relación entre diferentes variables en una población definida, sin embargo, no permite abordar la relación causa- efecto. Todos los estudios transversales son por definición descriptivos, siendo los más representativos los estudios de prevalencia y evaluación de pruebas diagnósticas.

Ejemplo de estudio:

- Se realiza una encuesta a 600 pacientes con enfermedad renal crónica atendidos en consulta externa para evaluar su adherencia actual al tratamiento farmacológico y a las recomendaciones dietéticas. En esta única medición se registran variables como cumplimiento de medicación, restricciones alimentarias, asistencia a controles y niveles recientes de creatinina. Este diseño permite estimar la prevalencia de baja adherencia y explorar factores asociados, como nivel educativo o tiempo desde el diagnóstico, brindando una "fotografía" del estado actual de la población estudiada.

Longitudinales: Siguen a los mismos individuos a lo largo del tiempo para observar cambios y efectos a largo plazo. Establece una secuencia temporal entre las variables y pueden ser estudios tanto descriptivos como analíticos, es decir: estudios de incidencia, estudios de cohorte (de causa a efecto) y estudios de casos y controles (de efecto a causa).

Ejemplo de estudios:

- **Cohorte:** se realiza el seguimiento de un grupo de pacientes con obesidad mórbida durante un periodo de 10 años con el objetivo de evaluar la aparición de enfermedades

https://doi.org/10.62486/978-9915-9928-3-9.Ch04

cardiovasculares como infarto agudo de miocardio, insuficiencia cardíaca o accidente cerebrovascular. Los participantes son evaluados periódicamente para registrar cambios en peso, presión arterial, perfil lipídico y hábitos de vida. Permite estimar la incidencia de nuevos casos y establecer una secuencia temporal entre la exposición (obesidad) y el desenlace cardiovascular.

- **Casos y controles:** se revisan historias clínicas de pacientes que desarrollaron neumonía asociada a ventilador (casos) y se comparan con historias de pacientes que no desarrollaron esta complicación (controles). Se recopila información retrospectiva sobre factores de exposición como duración de la intubación, uso de sedación profunda, presencia de aspiración o colonización previa.

Mixtos: Combina mediciones transversales y longitudinales para obtener una visión más completa del fenómeno estudiado.

Ejemplo de estudio:

- Intervención educativa en MBE dirigida a internos de medicina. En una primera fase, se realiza un análisis transversal para medir sus conocimientos, habilidades y actitudes respecto a la MBE antes de iniciar la intervención. Posteriormente, se efectúan dos seguimientos longitudinales, a los 6 y 12 meses, con el objetivo de evaluar si los cambios obtenidos tras el programa educativo se mantienen en el tiempo. Este diseño permite no solo describir la situación inicial, sino también observar la evolución y el impacto sostenido de la intervención en la práctica clínica de los futuros médicos.

5.4 CLASIFICACIÓN SEGÚN LA CRONOLOGÍA DE ESTUDIO:

Prospectivo: los investigadores recogen datos hacia adelante en el tiempo a medida que van sucediendo, es decir, su inicio es anterior a los hechos estudiados. Es ideal para observar la incidencia de eventos y examinar relaciones causales, ya que permite controlar las variables desde el comienzo.

Ejemplo:

- Los investigadores comienzan la recolección de datos antes de que ocurra el desenlace de interés y siguen a los participantes hacia adelante en el tiempo. Por ejemplo, en una cohorte prospectiva, se recluta a un grupo de mujeres con diagnóstico de síndrome de ovario poliquístico (SOP) y se las compara con un grupo sin SOP, ambos inicialmente libres de hipertensión arterial. A lo largo de 10 años de seguimiento, se registran mediciones periódicas de presión arterial, antecedentes clínicos y factores de riesgo metabólico para determinar la incidencia de hipertensión arterial en cada grupo. Este diseño permite establecer la secuencia temporal entre exposición (SOP) y desenlace (hipertensión) y controlar variables relevantes desde el inicio, fortaleciendo la evaluación de una posible relación causal.

Retrospectivo: Los investigadores recogen los datos de eventos que ya ocurrieron, como datos registrados de archivos o historias clínicas, por lo que su diseño es posterior a los hechos estudiados. Es útil para estudiar asociaciones entre factores de riesgo y enfermedades cuando ya se ha producido un desenlace.

Ejemplo:

- En un estudio de casos y controles, se revisan historias clínicas de pacientes que fallecieron por COVID-19 (casos) y se comparan con historias de pacientes que sobrevivieron a la enfermedad (controles), todos atendidos durante el mismo periodo. Se recopilan antecedentes previos al desenlace, como índice de masa corporal, comorbilidades, parámetros inflamatorios y tratamientos recibidos, con el objetivo de determinar si la obesidad actuó como un factor de riesgo independiente para la mortalidad. Este diseño permite explorar asociaciones relevantes en situaciones donde el desenlace ya está definido y no es posible realizar estudios prospectivos.

Ambispectivo: es la combinación de ambos métodos previamente mencionados. Empiezan con datos históricos y continúan recolectando datos hacia el futuro. Permite a los investigadores analizar eventos pasados y al mismo tiempo observar desenlaces futuros, maximizando la información disponible.

Ejemplo:

- Estudio en una cohorte ambispectiva de pacientes con lupus eritematoso sistémico, se revisan retrospectivamente las historias clínicas para documentar brotes previos, actividad de la enfermedad, tratamientos recibidos y complicaciones acumuladas. Posteriormente, los

https://doi.org/10.62486/978-9915-9928-3-9.Ch04

mismos pacientes son seguidos prospectivamente para evaluar la eficacia y seguridad de nuevos tratamientos inmunosupresores, registrando brotes futuros, efectos adversos y cambios en biomarcadores. Este diseño permite integrar una visión completa del curso de la enfermedad, maximizando la información disponible y agilizando la obtención de resultados clínicos relevantes.

6. VARIABLES EN INVESTIGACIÓN

Las variables se definen como una característica, propiedad o cualidad de personas o cosas de un conjunto determinado que es susceptible a la modificación y se la puede estudiar, controlar o medir a lo largo de una investigación. Son esenciales para definir el enfoque y los resultados de la investigación.

6.1 DESDE EL PUNTO DE VISTA METODOLÓGICO:

las variables se clasifican en:

Variable independiente: Es la variable que el investigador manipula o cuya influencia quiere estudiar. Representa la intervención o el factor de exposición o de estudio.

- Ejemplo: En un estudio que analiza el efecto de la dieta en la reducción del colesterol, la dieta (alta en grasas versus baja en grasas) sería la variable independiente, ya que es la que el investigador manipula para ver su efecto sobre el colesterol.

Variable dependiente: Es la respuesta o resultado que se mide en el estudio. Depende de los cambios o la manipulación de la variable independiente.

- Ejemplo: En el mismo estudio de dieta y colesterol, el nivel de colesterol en sangre (normal o elevado) es la variable dependiente, ya que es el resultado que se mide para evaluar el efecto de la dieta. (Dieta con más contenido en grasas produce colesterol más elevado)

Variables confusoras: Son factores externos que pueden influir en los resultados y distorsionar la relación entre la variable independiente y la dependiente.

- Ejemplo: En el estudio de dieta y colesterol, el nivel de actividad física podría ser una variable confusora. Las personas que hacen más ejercicio pueden tener niveles de colesterol más bajos, independientemente de su dieta. Si no se controla la actividad física, podría parecer que la dieta tiene un efecto mayor o menor del que realmente tiene.

Variables de control: Son aquellas que se mantienen constantes durante el estudio para evitar que influyan en los resultados. Controlarlas ayuda a reducir la variabilidad en los datos y a mejorar la validez interna del estudio.

- Ejemplo: En el mismo estudio sobre el efecto de la dieta en la reducción del colesterol, la edad de los participantes se podría controlar, seleccionando solo personas dentro de un mismo rango etario (por ejemplo, de 40 a 50 años). La edad influye en los niveles de colesterol, ya que las personas mayores tienden a tener un metabolismo más lento y colesterol más elevado. Si el estudio incluye personas de diferentes edades sin controlar este factor, la variabilidad en los resultados podría deberse a la edad y no a la dieta en sí.

6.2 DESDE EL PUNTO DE VISTA ESTADÍSTICO:

las variables se clasifican en:

Variables cualitativas (o categóricas): Representan atributos o categorías que no tienen un orden numérico inherente, se subdividen en:

- **Nominales:** Las variables nominales son categóricas y no presentan un orden inherente entre sus categorías. Para su análisis estadístico se emplean principalmente tablas de frecuencia y proporciones (número y porcentaje), y cuando se requiere comparar grupos, se utilizan pruebas como chi-cuadrado.

Por ejemplo, en un estudio con 300 pacientes con hipertensión, se analiza la distribución del grupo sanguíneo: A (25%), B (30%), AB (10%) y O (35%). Esta información se presenta en una tabla de frecuencias y puede compararse entre subgrupos, como pacientes con hipertensión controlada versus no controlada, utilizando la prueba de chi-cuadrado para evaluar si la distribución del grupo sanguíneo difiere significativamente entre ellos.

- **Ordinales**: Poseen un orden jerárquico entre sus categorías, aunque no existe una distancia exacta entre ellas. Además de resumirse mediante tablas de frecuencia y porcentajes,

https://doi.org/10.62486/978-9915-9928-3-9.Ch04

su análisis puede incluir pruebas como chi-cuadrado para tendencias o pruebas no paramétricas (por ejemplo, Wilcoxon) cuando se comparan grupos.

Ejemplo: evaluación del grado de disnea en pacientes con EPOC mediante la escala mMRC, donde se registran los siguientes valores: mMRC 0 (15 pacientes), mMRC 1 (20 pacientes), mMRC 2 (40 pacientes), mMRC 3 (10 pacientes) y mMRC 4 (5 pacientes). Esta variable ordenada permite analizar si existe una distribución diferente de severidad entre grupos, como antes y después de una intervención terapéutica, utilizando pruebas estadísticas adecuadas para datos ordinales.

Variables cuantitativas (o numéricas): representan cantidades medibles y se expresan en valores numéricos, permitiendo establecer comparaciones en términos de magnitud. Se subdividen en:

Según su naturaleza numérica:

- **Discretas**: solo toman valores enteros y suelen representar conteos, como el número de consultas médicas realizadas en un mes o el número de hijos.
- **Continuas**: pueden asumir cualquier valor dentro de un rango (incluso decimales) y corresponden a mediciones, como el peso corporal en kilogramos, la estatura en centímetros o los niveles de glucosa en sangre en mg/dL.

Según su escala de medición:

Intervalo: Tienen valores numéricos con distancias iguales entre sí, pero su cero no representa ausencia real de la variable, sino un punto arbitrario. Esto impide interpretar proporciones.

Ejemplo clásico: la temperatura en grados Celsius o Fahrenheit, donde 0 °C no significa ausencia de temperatura, sino un valor definido convencionalmente dentro de la escala.

- Razón: Variables numéricas que poseen un cero real, lo que significa ausencia total de la característica medida. Permiten todas las operaciones matemáticas, incluidas proporciones (por ejemplo, “el doble”, “la mitad”).

Ejemplo: peso corporal, altura, glucosa en sangre, edad, creatinina, número de internaciones, número de hijos. En todas estas variables, un valor de cero indica ausencia real de lo medido.

Para su análisis estadístico, estas variables permiten calcular medidas de tendencia central (media, mediana, moda) y medidas de dispersión (desviación estándar, rango, varianza). La elección de la prueba estadística depende de la distribución de los datos: se emplean pruebas paramétricas como la *t de Student* cuando los datos siguen una distribución normal, y pruebas no paramétricas como la *Mann-Whitney U* cuando la distribución no es normal.

Por ejemplo, al comparar el número de internaciones en el último año entre pacientes con y sin diabetes, se obtiene una media de 1.1 internaciones en el grupo sin diabetes y 2.3 internaciones en el grupo con diabetes. Esta diferencia puede analizarse con una *t de Student* si los datos son normales, o con una prueba de Mann-Whitney U si no cumplen este supuesto.

Tabla 1. Clasificación de variables en investigación en salud

Tipo de variable	Subtipo	¿Orden?	¿Cero real?	Ejemplos
Cualitativa	Nominal	No	No aplica	Grupo sanguíneo, sexo
Cualitativa	Ordinal	Sí	No aplica	Dolor, NYHA, mMRC
Cuantitativa	Discreta	No	Sí (si es razón)	Número de hijos, internaciones
Cuantitativa	Continua	Sí	Sí (si es razón)	Peso, glucosa, creatinina
Cuantitativa	Intervalo	Sí	No	Temperatura °C
Cuantitativa	Razón	Sí	Sí	Edad, peso, número de consultas

Fuente: Elaboración propia

7. TIPOS DE ERROR EN INVESTIGACIÓN

Existen dos tipos principales de error en investigación que afectan la validez de los resultados:

7.1 ERROR TIPO I (FALSO POSITIVO):

Ocurre cuando se rechaza incorrectamente la hipótesis nula, es decir, se concluye que hay un efecto o relación cuando en realidad no lo hay.

- Ejemplo: Imaginemos un estudio que evalúa si un nuevo medicamento reduce la presión arterial comparado con un placebo. Supongamos que los resultados muestran una diferencia estadísticamente significativa, y el estudio concluye que el medicamento es eficaz para reducir la presión arterial. Sin embargo, en realidad, el medicamento no tiene ningún efecto sobre la presión arterial; la diferencia observada es sólo producto del azar.

https://doi.org/10.62486/978-9915-9928-3-9.Ch04

Este sería un Error Tipo I, ya que se ha encontrado un "efecto" que en realidad no existe.

Consecuencias: Un Error Tipo I puede llevar a introducir tratamientos ineficaces en la práctica clínica, lo que implica un gasto innecesario y posible riesgo para los pacientes.

7.2 ERROR TIPO II (FALSO NEGATIVO):

Ocurre cuando se acepta incorrectamente la hipótesis nula, es decir, se concluye que no hay efecto cuando en realidad sí lo hay.

- Ejemplo: Imaginemos un estudio que evalúa si una dieta baja en carbohidratos ayuda a reducir el nivel de glucosa en pacientes con diabetes tipo 2. Supongamos que los resultados del estudio no muestran una diferencia estadísticamente significativa entre el grupo de dieta baja en carbohidratos y el grupo control, y se concluye que la dieta no tiene efecto. En realidad, la dieta baja en carbohidratos sí reduce los niveles de glucosa, pero debido a un tamaño de muestra insuficiente o variabilidad en los datos, el estudio no logró detectar el efecto. Este sería un Error Tipo II.

Consecuencias: Un Error Tipo II puede llevar a descartar tratamientos o intervenciones que realmente son beneficiosos, lo que priva a los pacientes de una opción terapéutica efectiva.

8. SESGOS Y VALIDEZ EN LA INVESTIGACIÓN

8.1 SESGO:

Es una desviación sistemática que se introduce en un estudio y que provoca que los resultados obtenidos se alejen de la realidad. A diferencia del azar (que introduce variabilidad), el sesgo no ocurre al azar, sino que se repite de forma constante debido a errores en el diseño, la selección de participantes, la medición de variables o el análisis.

En otras palabras, un estudio sesgado tiende a sobreestimar o subestimar la verdadera relación entre una exposición y un desenlace. Esto afecta directamente la validez interna del estudio y, por lo tanto, la correcta interpretación de la evidencia.

El sesgo puede aparecer en cualquier etapa de una investigación, desde la selección de la muestra hasta la forma de recolectar datos o la interpretación de los resultados. Por ello, reconocerlo permite valorar la calidad del estudio y la confiabilidad de sus conclusiones.

Los tipos de sesgo más frecuentes son:

- **Sesgo de selección:** Ocurre cuando los individuos que participan en el estudio no representan adecuadamente a la población general. *Ejemplo:* Un estudio sobre tabaquismo sólo incluye pacientes que asisten a clínicas privadas. Esto puede sobreestimar la adherencia al abandono del tabaco.
- **Sesgo de información (clasificación):** Se produce cuando hay errores al medir o clasificar las variables. *Ejemplo:* Un estudio sobre obesidad utiliza autoinforme de peso y talla, lo que puede subestimar el IMC real.
- **Sesgo de recuerdo:** Frecuente en estudios retrospectivos, donde los participantes no recuerdan con precisión exposiciones pasadas. *Ejemplo:* En un estudio sobre factores de riesgo de aborto espontáneo, las mujeres que han tenido pérdida tienden a recordar más detalles de hábitos nocivos que las que no.
- **Sesgo del observador:** Sucede cuando el investigador influye en la medición, voluntaria o involuntariamente. *Ejemplo:* Un evaluador que conoce el grupo experimental puede interpretar sus resultados de forma más favorable.

Validez: La validez se refiere al grado en que los resultados de un estudio reflejan la realidad y pueden ser confiables para la toma de decisiones clínicas. Se divide en dos componentes fundamentales: validez interna y validez externa, cada una con un rol clave en la interpretación crítica de la evidencia.

- **Validez interna:**es la capacidad del estudio para demostrar que la relación observada entre una exposición y un desenlace es real y no producto de factores ajenos al diseño. Es decir, indica qué tan bien el estudio controla las fuentes de error que podrían distorsionar sus conclusiones. Cuanto mejor se controlen estos elementos, más confianza se puede tener en que los resultados representan la verdadera relación causa-efecto dentro de la población estudiada. Es afectada por sesgos, variables confusoras no controladas y errores en la ejecución del estudio.

https://doi.org/10.62486/978-9915-9928-3-9.Ch04

- **Validez externa (generalizabilidad):** Describe la capacidad de aplicar los resultados del estudio a otras poblaciones, lugares o situaciones distintas a las del estudio original. Un estudio puede tener excelente validez interna y aun así no ser aplicable a diferentes contextos si su población o condiciones no representan la realidad clínica del lector. Esta depende de la representatividad de la muestra, las características de la población, el entorno asistencial, las características culturales y epidemiológicas. *Ejemplo:* Un estudio realizado en varones de 40 a 60 años en Europa puede no ser generalizable a mujeres, adultos jóvenes, población latinoamericana.

Tabla 2: Comparación entre validez interna y externa

Característica	Validez interna	Validez externa
¿Qué evalúa?	Si los resultados reflejan una relación real entre exposición y desenlace dentro del estudio.	Si los resultados pueden aplicarse a otras poblaciones, entornos o contextos.
Pregunta clave	"¿Los resultados son creíbles dentro de esta investigación?"	"¿Puedo aplicar estos resultados a mis pacientes o a otras poblaciones?"
Factores que la afectan	Sesgos, confusores, errores de medición, mala selección de muestra, fallas en el diseño.	Tipo de población estudiada, contexto clínico, entorno cultural, criterios de inclusión muy restrictivos.
Cómo se mejora	Controlando sesgos, estandarizando mediciones, aleatorizando, enmascarando, ajustando confusores.	Seleccionando muestras representativas, utilizando criterios de inclusión amplios, replicando estudios en diversas poblaciones.
Limitaciones típicas	Un exceso de control puede limitar la generalización.	Una amplia generalización puede comprometer el control interno del estudio.
Ejemplo clínico	Un ensayo clínico con estricta aleatorización y doble ciego.	Un estudio cuyos resultados aplican a pacientes en otros países, edades o sistemas de salud.

Fuente: Elaboración propia

9. INTELIGENCIA ARTIFICIAL EN EL PROCESO DE INVESTIGACIÓN

La incorporación de la inteligencia artificial (IA) en la investigación en salud ha transformado la manera en que se formulan preguntas clínicas, se revisa la literatura y se estructuran los estudios. Aunque no reemplaza el rigor científico ni el juicio del investigador, la IA permite optimizar múltiples etapas del proceso investigativo descrito en este capítulo.

- **Planteamiento del problema:** Los modelos de lenguaje pueden ayudar a clarificar ideas iniciales y a formular preguntas clínicas estructuradas, como aquellas basadas en el marco PICO. Esto facilita definir con mayor precisión la población, la exposición, las variables y los posibles desenlaces, fortaleciendo desde el inicio la validez del estudio.
- **Revisión de literatura:** La IA ha demostrado ser especialmente útil gracias a herramientas que permiten realizar búsquedas semánticas más profundas y rápidas que las realizadas únicamente en bases tradicionales. Estas plataformas pueden sintetizar información, extraer datos relevantes y ayudar a identificar vacíos de conocimiento que justifican el inicio de una investigación. La IA no reemplaza la búsqueda en bases como PubMed o Scopus, pero la complementa, haciéndola más eficiente y estructurada.
- **Diseño metodologico de estudio:**Puede asistir al investigador proponiendo posibles tipos de estudio según el objetivo planteado, identificando potenciales fuentes de sesgo y sugiriendo estrategias para controlarlos. Algunos sistemas permiten incluso generar diagramas metodológicos o detectar variables confusoras que deben ser consideradas desde el diseño. Esto aporta a mejorar la validez interna del estudio.
- **Análisis de datos:** Las plataformas de IA pueden organizar bases de datos extensas, detectar patrones complejos y sugerir pruebas estadísticas adecuadas para cada tipo de variable. Asimismo, permiten explorar modelos predictivos basados en machine learning, siempre bajo interpretación responsable del investigador, quien debe evaluar su pertinencia clínica y metodológica.
- **Redacción de manuscritos:**Puede facilitar la elaboración de borradores, mejorar la claridad del lenguaje académico, generar resúmenes estructurados y apoyar en la organización de referencias en formato Vancouver. También puede contribuir en la creación de tablas, figuras y comparaciones didácticas que enriquecen la presentación de resultados. Sin embargo, el contenido final debe ser revisado rigurosamente y sustentado exclusivamente con literatura científica indexada.
- **Limitaciones éticas:**La IA no evalúa la calidad metodológica de un estudio ni puede

https://doi.org/10.62486/978-9915-9928-3-9.Ch04

reemplazar la revisión crítica del investigador. Tampoco debe emplearse para generar datos inexistentes o referencias no verificadas. Su uso debe ser transparente y complementario al trabajo científico, nunca sustitutivo.

10. CONCLUSIONES

Comprender el proceso de investigación y los conceptos básicos asociados es fundamental para interpretar y aplicar los resultados de estudios clínicos en la práctica de la MBE. Los tipos de estudio, la clasificación de variables y el conocimiento de los errores tipo I y tipo II permiten al lector desarrollar una lectura crítica y evaluativa de los artículos científicos. La MBE depende de la capacidad del médico para interpretar correctamente la evidencia científica, distinguiendo estudios sólidos de aquellos con limitaciones o sesgos. El dominio de estos conceptos permite que los profesionales de la salud puedan:

- Evaluar la validez y relevancia de la información obtenida en sus búsquedas.
- Reconocer el tipo de estudio y entender las variables y los errores que podrían influir en los resultados.
- Aplicar el conocimiento adquirido para tomar decisiones clínicas informadas y centradas en el paciente.

A medida que se avance en el libro, estos conceptos serán la base para evaluar críticamente distintos tipos de estudios y comprender cómo se relacionan con la práctica médica diaria.

12. PREGUNTAS DE AUTOEVALUACIÓN

PREGUNTAS DE OPCIÓN MÚLTIPLE:

1. ¿Cuál de las siguientes es una característica de un estudio transversal?
a) Observa eventos a lo largo del tiempo

b) Aplica una intervención terapéutica.

c) Recolecta datos en un solo momento temporal.
d) Evalúa exclusivamente relaciones causales.
2. ¿Qué tipo de variable es el nivel educativo (primaria, secundaria, universitaria, posgrado)?
a) Cuantitativa continua
b) Cualitativa nominal
c) Cualitativa ordinal
d) Cuantitativa discreta
3. En un estudio retrospectivo, ¿qué característica lo define?
a) Analiza datos previamente registrados en el pasado.
b) Recolecta datos hacia el futuro desde un punto inicial.
c) Requiere la aplicación de un tratamiento por parte del investigador.
d) Siempre es un ensayo clínico.
4. ¿Qué tipo de error se comete cuando se concluye que hay un efecto, pero en realidad no lo hay?
a) Error tipo I
b) Error tipo II
c) Sesgo de información
d) Falta de validez externa
5. ¿Cuál de las siguientes variables es cuantitativa continua?
a) Grupo sanguíneo
b) Número de hijo
c) Nivel de dolor (leve, moderado, severo)
d) Glucosa en sangre (mg/dL)

PREGUNTAS ABIERTAS:

6. Explique brevemente la diferencia entre una variable cualitativa nominal y una ordinal,

https://doi.org/10.62486/978-9915-9928-3-9.Ch04

incluyendo un ejemplo clínico de cada una.

7. ¿Qué ventajas y limitaciones tiene un estudio transversal comparado con uno longitudinal?

8. En el contexto de la Medicina Basada en la Evidencia, ¿por qué es importante la validez interna en un estudio?

9. ¿Qué tipo de sesgo ocurriría si un estudio sobre actividad física utiliza autoinforme de los pacientes, y cómo podría minimizarse?

10. Proponga una pregunta de investigación clínica relevante y mencione qué tipo de estudio utilizaría para responderla, justificando su elección.

RESPUESTAS:

1: c) Recolecta datos en un solo momento temporal. Un estudio transversal toma una "fotografía" de la población en un punto específico del tiempo. No permite establecer causalidad ni seguimiento temporal.

2: c) Cualitativa ordinal. El nivel educativo tiene categorías con un orden lógico, pero sin distancia numérica definida entre ellas, lo que lo convierte en una variable ordinal.

3: a) Analiza datos previamente registrados en el pasado. Los estudios retrospectivos utilizan información ya existente (historias clínicas, bases de datos) y analizan eventos que ya ocurrieron. No recolectan datos hacia adelante en el tiempo.

4: a) Error tipo I. El error tipo I es un falso positivo, es decir, concluir que existe un efecto cuando en realidad no lo hay (rechazar incorrectamente la hipótesis nula).

5: d) Glucosa en sangre (mg/dL). La glucosa es una variable cuantitativa continua, ya que puede tomar valores decimales dentro de un rango. Grupo sanguíneo y nivel de dolor son cualitativas; número de hijos es discreta.

6. Diferencia entre variable nominal y ordinal, con ejemplos.

Una variable nominal clasifica categorías sin orden entre ellas (ej.: grupo sanguíneo: A, B, AB, O).

Una variable ordinal tiene categorías ordenadas, pero sin una distancia numérica precisa (ej.: grado de dolor: leve, moderado, severo).

La clave es reconocer si existe jerarquía. Si hay orden, es ordinal; si no, es nominal.

7. Ventajas y limitaciones de un estudio transversal vs. longitudinal.

El estudio transversal es rápido, económico y útil para estimar prevalencias, pero no permite establecer causalidad ni secuencia temporal.

El estudio longitudinal permite observar cambios a lo largo del tiempo y evaluar causalidad, pero es más costoso y requiere seguimiento prolongado.

Todo radica en la temporalidad: el transversal describe; el longitudinal observa evolución y efectos.

8. Importancia de la validez interna en Medicina Basada en la Evidencia.

La validez interna asegura que los resultados del estudio reflejan una relación real entre la exposición y el desenlace, libres de sesgos y confusores. Sin validez interna, los hallazgos no son confiables, por lo que no deben aplicarse en la práctica clínica.

La validez interna determina si los resultados son creíbles; sin ella, la evidencia pierde valor.

9. Tipo de sesgo en autoinformes y cómo minimizarlo.

El uso de autoinforme puede introducir un sesgo de información, también llamado sesgo de recuerdo o medición, porque los participantes pueden reportar incorrectamente su actividad física.

Para minimizarlo, se pueden usar instrumentos validados, acelerómetros, registros directos o verificar datos con múltiples fuentes.

Siempre que el participante "recuerda" o "autoinforma", hay riesgo de error sistemático.

10. Pregunta de investigación propuesta y tipo de estudio apropiado.

Pregunta: "¿La obesidad aumenta el riesgo de desarrollar diabetes tipo 2 en adultos jóvenes?"

Estudio: Una cohorte prospectiva, porque permite seguir a los participantes a lo largo del tiempo para observar si la exposición (obesidad) precede al desenlace (diabetes).

Es fundamental que el estudiante justifique el tipo de estudio según temporalidad y causalidad.

https://doi.org/10.62486/978-9915-9928-3-9.Ch04

REFERENCIAS

Dewidar O, Al-Zubaidi A, Bondok M, Abdelrazeq L, Huang J, Jearvis A, et al. Reporting of equity in observational epidemiology: A methodological review. J Glob Health. 2024;14:04046. Available at: https://doi.org/10.7189/jogh.14.04046.
Goldkuhle M, Hirsch C, Iannizzi C, van Dalen EC, Hemkens LG, Monsef I, et al. Exploring the characteristics, methods and reporting of systematic reviews with meta-analyses of time-to-event outcomes: A meta-epidemiological study. BMC Med Res Methodol. 2024;24:291. Available at: https://doi.org/10.1186/s12874-024-02401-4.
Greenhalgh T. How to read a paper: The basics of evidence-based medicine and healthcare. 6th ed. Oxford: Wiley-Blackwell; 2019.
Harris AD, McGregor JC, Perencevich EN, Furuno JP, Zhu J, Peterson DE, et al. The use and interpretation of quasi-experimental studies in medical informatics. J Am Med Inform Assoc. 2006;13(1):16–23. Available at: https://doi.org/10.1197/jamia.M1749.
Hulley SB, Cummings SR, Browner WS, Grady D, Newman TB. Designing clinical research. 4th ed. Philadelphia: Lippincott Williams & Wilkins; 2013.
Kelsey JL, Whittemore AS, Evans AS, Thompson WD. Methods in observational epidemiology. 2nd ed. New York: Oxford University Press; 1996.
Lozada-Martínez ID, Hernandez-Paz DA, Fiorillo-Moreno O, Picón-Jaimes YA, Bermúdez V. Meta-research in biomedical investigation: Gaps and opportunities based on meta-research publications and global indicators in health, science, and human development. Publications. 2025;13(1):7. Available at: https://doi.org/10.3390/publications13010007.
Porta M, editor. A dictionary of epidemiology. 6th ed. New York: Oxford University Press; 2014.
Rodríguez-Suárez CA, González-de la Torre H. Tipologías y enfoques metodológicos de los estudios de revisiones en ciencias de la salud: recomendaciones para los investigadores. Rev Esp Salud Pública. 2024;98(2):e202402001.
Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: What it is and what it isn't. BMJ. 1996;312(7023):71–2. Available at: https://doi.org/10.1136/bmj.312.7023.71.
Szklo M, Nieto FJ. Epidemiology: Beyond the basics. 4th ed. Burlington: Jones & Bartlett Learning; 2018.
Toews I, Anglemyer A, Nyirenda JL, Alsaid D, Balduzzi S, Grummich K, et al. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials: A meta-epidemiological study. Cochrane Database Syst Rev. 2024;1:MR000034. Available at: https://doi.org/10.1002/14651858.MR000034.pub3.
Warriner D. How to read a paper: The basics of evidence-based medicine. BMJ. 2008;336(7657):1381. Available from: https://doi.org/10.1136/bmj.a186

Chapter 5 / Capítulo 5

Critical reading of scientific articles on health: A practical guide (Spanish Edition)
ISBN: 978-9915-9928-3-9
DOI: 10.62486/978-9915-9928-3-9.Ch05
Pages: 41-50

Check for updates

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

How to find a scientific article on health? A step-by-step strategy

¿Cómo buscar un artículo científico en salud? Estrategia paso a paso

Mauricio Alejandro Paz Del Rio
Blas Apaza-Huanca
Josué Elías Peca-Hoyos

1 Clínica Los Andes. La Paz, Bolivia.
2 Ministerio de Salud y Deportes. Instituto Académico Científico Quispe-Cornejo. La Paz, Bolivia.
3 Universidad Autónoma Juan Misael Saracho. Tarija, Bolivia.

ABSTRACT

Introduction: Literature searching is a fundamental competency within Evidence-Based Medicine (EBM), as it allows for the identification, retrieval, and selection of relevant scientific information to answer clinical questions systematically and with sound reasoning. An appropriate search strategy facilitates access to up-to-date, valid evidence applicable to clinical practice. Development: This chapter describes a structured methodology for conducting efficient scientific searches in health. The process begins with formulating clinical questions using the PICO strategy, a tool that allows for defining the population, intervention, comparison, and outcome of interest. Subsequently, the chapter addresses the identification of key concepts and their transformation into controlled terms using biomedical thesauri such as MeSH and DeCS, which improve the precision and reproducibility of searches. Furthermore, it explains the use of Boolean operators (AND, OR, and NOT) to optimize the sensitivity and specificity of the results obtained. The chapter also presents the main scientific information resources, including primary studies, systematic reviews, and evidence synopses. Therefore, criteria for evaluating the quality of the findings are analyzed, considering bibliometric indicators, clinical relevance, methodological quality, and applicability of the results. Furthermore, the role of artificial intelligence as a complementary tool to support the search and management of scientific information is explored. Conclusion: Searching for scientific evidence is not simply about finding articles, but about locating the best available evidence to answer a specific clinical question. The quality of clinical decisions depends, to a large extent, on the ability to formulate appropriate questions, conduct efficient searches, and critically evaluate the retrieved information. In an era of information overload, learning to search with scientific rigor is as important as learning to interpret the evidence found.

Keywords: Literature search; Evidence-Based Medicine; PICO strategy; MeSH and DeCS terms; Boolean operators

RESUMEN

Introducción: La búsqueda bibliográfica constituye una competencia fundamental dentro de la Medicina Basada en la Evidencia (MBE), ya que permite identificar, recuperar y seleccionar información científica relevante para responder preguntas clínicas de manera sistemática y fundamentada. Una estrategia de búsqueda adecuada facilita el acceso a evidencia actualizada, válida y aplicable a la práctica asistencial. Desarrollo: El capítulo describe una metodología estructurada para realizar búsquedas científicas eficientes en salud. El proceso inicia con la formulación de preguntas clínicas mediante la estrategia PICO, herramienta que permite delimitar la población, intervención, comparación y resultado de interés. Posteriormente, se aborda la identificación de conceptos clave y su transformación en términos controlados utilizando tesauros biomédicos como MeSH y DeCS, los cuales mejoran la precisión y reproducibilidad de las búsquedas. Además, se explica el uso de operadores booleanos (AND, OR y NOT) para optimizar la sensibilidad y especificidad de los resultados obtenidos. El capítulo también presenta los principales recursos de información científica, incluyendo estudios primarios, revisiones sistemáticas y sinopsis de evidencia. Por tanto, se analizan criterios para evaluar la calidad de los hallazgos encontrados, considerando indicadores bibliométricos, relevancia clínica, calidad metodológica y aplicabilidad de los resultados. Además, se explora el papel de la inteligencia artificial como herramienta complementaria para apoyar la búsqueda y gestión de información científica. Conclusión: Buscar evidencia científica no consiste únicamente en encontrar artículos, sino en localizar la mejor evidencia disponible para responder una pregunta clínica específica. La calidad de las decisiones clínicas depende, en gran medida, de la capacidad para formular preguntas adecuadas, realizar búsquedas eficientes y evaluar críticamente la información recuperada. En una era de sobreabundancia informativa, aprender a buscar con rigor científico es tan importante como aprender a interpretar la evidencia encontrada.

Palabras clave: Búsqueda bibliográfica; Medicina Basada en la Evidencia; Estrategia PICO; Términos MeSH y DeCS; Operadores booleanos

1. CASO CLÍNICO

Un médico residente de medicina interna, atiende a una paciente con hipertensión arterial resistente al tratamiento antihipertensivo convencional. Ante la persistencia de cifras tensionales elevadas, surge la interrogante clínica de si la adición de espironolactona al esquema terapéutico vigente podría mejorar el control de la presión arterial. Para responder esta pregunta de manera fundamentada, la residente reconoce la necesidad de consultar la evidencia científica disponible. Sin embargo, enfrenta dificultades relacionadas con el proceso de búsqueda bibliográfica: desconoce cuáles son las fuentes de información más apropiadas, si debe utilizar motores de búsqueda generales o bases de datos biomédicas especializadas, qué términos de búsqueda emplear y cómo identificar estudios pertinentes, metodológicamente sólidos y clínicamente relevantes.

Esta situación representa un escenario frecuente en la práctica clínica contemporánea, donde los profesionales de la salud deben tomar decisiones basadas en la mejor evidencia científica disponible. No obstante, la ausencia de competencias en búsqueda bibliográfica y gestión de la información puede limitar la capacidad para localizar, seleccionar y utilizar evidencia de calidad de manera eficiente y sistemática.

2. OBJETIVOS

- Formular preguntas clínicas estructuradas utilizando la estrategia PICO para orientar la búsqueda de evidencia científica.
- Identificar y seleccionar los conceptos clave de una pregunta de investigación para transformarlos en estrategias de búsqueda bibliográfica.
- Aplicar descriptores controlados de vocabularios biomédicos estandarizados, como MeSH y DeCS, para optimizar la sensibilidad y especificidad de las búsquedas.
- Construir estrategias de búsqueda avanzadas mediante el uso adecuado de operadores booleanos, truncamientos y filtros metodológicos.
- Seleccionar las bases de datos y fuentes de información más apropiadas según el tipo de pregunta clínica o de investigación.
- Integrar herramientas de inteligencia artificial como apoyo complementario en los procesos de búsqueda, síntesis y gestión de la información científica, reconociendo sus ventajas y limitaciones.

3. INTRODUCCIÓN

En el contexto de la Medicina Basada en la Evidencia (MBE), la capacidad para buscar información científica de manera eficiente constituye una competencia esencial para la práctica clínica. Buscar información no equivale necesariamente a encontrar la mejor evidencia disponible. Una estrategia de búsqueda inadecuada puede generar un volumen excesivo de resultados irrelevantes o, por el contrario, omitir estudios de alta calidad que resulten determinantes para la toma de decisiones clínicas. La búsqueda bibliográfica representa el proceso mediante el cual los profesionales de la salud identifican, recuperan y seleccionan información científica relevante para responder preguntas relacionadas con el tratamiento, el diagnóstico, el pronóstico, la prevención y otros aspectos del cuidado de los pacientes. Este proceso permite fundamentar las decisiones clínicas n evidencia científica actualizada, válida y aplicable al contexto asistencial. Para garantizar resultados pertinentes y reproducibles, la búsqueda debe realizarse de manera sistemática, estructurada y orientada por una pregunta clínica claramente definida. El método clásico de búsqueda científica comprende una secuencia ordenada de pasos que facilitan la localización eficiente de la mejor evidencia disponible:

- Definir una pregunta clara.
- Usar términos precisos (tesauros).
- Combinar términos (operadores booleanos).
- Evaluar la calidad y relevancia de lo encontrado.

Este capítulo explica cada paso de forma práctica, preparando el terreno para el Capítulo 7, donde se mostrará cómo aplicarlo en bases de datos específicas como PubMed, Cochrane o SciELO.

https://doi.org/10.62486/978-9915-9928-3-9.Ch05

4. PASOS PARA UNA BÚSQUEDA CIENTÍFICA EFICIENTE

4.1 FORMULAR UNA BUENA PREGUNTA

Antes de iniciar la búsqueda, es crucial definir qué queremos saber, utilizando una estructura clara. Existen dos tipos principales de preguntas:

- **Básicas o generales**: Son preguntas abiertas, útiles para aprender conceptos. Estas se estructuran en dos componentes: raíz (por ejemplo, quién, por qué, dónde, cuál, cómo) y verbo.
- Ejemplo: ¿Por qué los corticoides aumentan la glucemia?
- **Específicas o de actualización**: Se centran en un aspecto puntual como el diagnóstico, tratamiento, pronóstico o prevención, se deben formular mediante la estrategia PICO.

4.1.1 ESTRATEGIA PICO

La estrategia PICO permite formular la pregunta de investigación de manera clara y estructurada. Cada componente ayuda a delimitar la búsqueda y mejorar la precisión de los resultados:

Componente	Descripción
P = Población o paciente	La condición clínica específica que se estudiará y que delimita la búsqueda de información.
I = Intervención o exposición	El factor de estudio, como una intervención, exposición o procedimiento.
C = Comparación	Intervención alternativa o grupo de comparación (puede omitirse si no es aplicable)
O = Outcome o Resultado	El resultado que se espera observar, como mejoría de síntomas, reducción de incidencia, etc.

Según el tipo de información que se desee recabar, los componentes de esta estrategia podrían ser modificados. Por ejemplo, cuando deseamos conocer información descriptiva, es decir datos de incidencia, prevalencia o frecuencia, los datos de intervención y comparación pueden ser excluidos de esta estrategia.

Ejemplos de preguntas PICO:

Tipo de pregunta	Paciente	Intervención	Comparación	Resultado	Pregunta PICO
Tratamiento	Pacientes con diabetes mellitus tipo II	Sitagliptina	Metformina	Mejorar el control de la glucemia	¿El uso de sitagliptina es mejor en el control glucémico que la metformina en pacientes con diabetes mellitus tipo II?
Diagnóstico	Pacientes con apendicitis aguda	Ecografía	Tomografía	Diagnóstico temprano	¿La ecografía permite un diagnóstico temprano en pacientes con apendicitis aguda? (gold standard es tomografía)
Pronóstico	Pacientes adultos	Antecedente de diabetes gestacional	Sin antecedente de diabetes gestacional	Desarrollo de diabetes tipo II	¿La diabetes gestacional es un factor predictivo para el desarrollo de diabetes mellitus tipo II?

Fuente: Elaboración Propia

5. CLASIFICACIÓN DE LOS RECURSOS DE INFORMACIÓN

Existen diferentes recursos de información, los cuales pueden ser:

- Estudios primarios u originales: Corresponden a investigaciones individuales con diferentes tipos de diseños, tales como: series de casos, estudios de cohortes, estudios de casos y controles, ensayos clínicos, etc.
- Síntesis: Se basan en la búsqueda exhaustiva de la literatura para identificar, evaluar y sintetizar todos los estudios primarios sobre el efecto de alguna intervención, son denominadas revisiones sistemáticas, revisiones narrativas, etc.
- Sinopsis: Corresponden a resúmenes estructurados de la información principal de estudios relevantes con análisis crítico por parte de los autores, en este grupo encontramos guías de

https://doi.org/10.62486/978-9915-9928-3-9.Ch05

la práctica clínica, revisiones globales de la evidencia (overviews) y revisiones secundarias como UpToDate®.

6. USO DE TESAUROS: TÉRMINOS MESH Y DECS

¿QUÉ ES UN TESAURO?

Un tesauro es una herramienta que organiza términos y palabras clave en una estructura jerárquica y estandarizada. Su propósito es facilitar la búsqueda de información precisa y relevante al usar términos controlados, en lugar de palabras libres o sin estructura. Los tesauros se utilizan ampliamente en bases de datos científicas y bibliotecas, ya que permiten agrupar conceptos relacionados bajo términos unificados, lo que mejora la consistencia y la precisión de las búsquedas.

En el contexto de la salud, los tesauros como MeSH (Medical Subject Headings) y DeCS (Descriptores en Ciencias de la Salud) son especialmente útiles porque:

- **Estandarizan los términos**: Cada concepto tiene un término específico que representa de manera consistente un tema, independientemente de los sinónimos o variantes idiomáticas.
- Facilitan las búsquedas precisas: Al utilizar términos controlados, se reduce el riesgo de obtener resultados irrelevantes y se aumenta la probabilidad de encontrar estudios directamente relacionados con el tema de interés.
- Estructuran el conocimiento: Los tesauros están organizados en jerarquías que permiten navegar entre términos amplios y específicos, adaptando la búsqueda según el nivel de detalle requerido.

Estos términos estandarizan la búsqueda, lo cual permite encontrar artículos relevantes sin la ambigüedad que puede surgir al usar lenguaje natural.

- **MeSH (Medical Subject Headings)**: Términos utilizados en bases de datos como PubMed.
- **DeCS (Descriptores en Ciencias de la Salud)**: La versión en español y portugués de MeSH, utilizada en SciELO, LILACS, BVS (Biblioteca Virtual de Salud).
- Tanto los términos MeSH como DeCS pueden encontrarse en Pubmed (inglés) o en la Biblioteca Virtual de salud (BVS - español).

Ejemplo: Para una búsqueda sobre el control de la glucemia en diabetes tipo II, se pueden usar términos MeSH como “Diabetes Mellitus, Type 2” o los términos DeCS “Diabetes Mellitus Tipo 2”.

7. OPERADORES BOOLEANOS

Los operadores booleanos son conectores que permiten realizar combinaciones entre términos de búsqueda, refinando los resultados:

Operador Booleano	Descripción
AND	Combina dos términos y devuelve artículos con ambas características. Ejemplo: Diabetes AND Ejercicio; Esto devolverá artículos que mencionan tanto diabetes como ejercicio, asegurando que ambos temas se abordan en los resultados. Reduce el número de resultados, centrándose en artículos donde se discutan ambos conceptos.
OR	Incluye artículos que contienen al menos uno de los términos solicitados, ampliando la búsqueda. Es útil para encontrar sinónimos o términos relacionados. Ejemplo: Diabetes OR Actividad Física. Esto mostrará artículos que mencionan al menos uno de los dos términos Amplía los resultados para abarcar ambos conceptos.
NOT	Excluye artículos que contienen el término subsiguiente al NOT, reduciendo los resultados a artículos que contengan solo el primer término. Es útil para eliminar información irrelevante en búsquedas amplias Ejemplo: Diabetes NOT Niños. Esto devolverá artículos que hablan de diabetes, pero excluirá aquellos que mencionan específicamente a niños.
Búsqueda combinada	Útil para combinar operadores para una búsqueda compleja. Por ejemplo, si se quiere encontrar artículos sobre el impacto del ejercicio en diabetes, pero solo en adultos (excluyendo niños), puedes escribir: Diabetes AND Ejercicio AND Adultos NOT Niños Este tipo de búsqueda mostrará artículos donde se hable de diabetes y ejercicio en adultos, pero excluirá aquellos en los que se mencionan niños.

https://doi.org/10.62486/978-9915-9928-3-9.Ch05

8. EVALUACIÓN DE LOS RESULTADOS DE LA BÚSQUEDA

Una vez obtenidos los resultados, el siguiente paso es determinar cuál evidencia vale la pena leer y aplicar. La calidad no depende solo del contenido, sino también del lugar donde se publica, el diseño del estudio y su aplicabilidad clínica. (Grafico 1)

Fuente: Elaboración propia

La evaluación debe realizarse en dos fases:

Fase 1. Evaluación de la calidad de la revista

La revista es el primer filtro de calidad. Publicar en una revista seria y bien indexada implica haber pasado por procesos rigurosos de revisión por pares. Para evaluar su calidad, se consideran principalmente tres aspectos:

A) SCIMAGOJOURNAL RANK (SJR) Y LOS CUARTILES (Q1–Q4)

SCImago clasifica las revistas según su impacto y calidad dentro de cada especialidad. Divide las revistas en cuatro cuartiles, del más alto al más bajo:

Cuartil	Descripción	Recomendación
Q1	25% superior de revistas de mayor impacto e influencia en su área	Ideal. Priorizar por alta calidad, rigor y visibilidad
Q2	25% siguiente, revistas de buen nivel académico	Muy recomendable. Rigor adecuado y buena reputación
Q3	Nivel intermedio	Aceptable si no existe evidencia en Q1–Q2
Q4	Nivel más bajo en su área	Evitar salvo que se trate de temas extremadamente específicos

Fuente: Elaboración propia

Si un artículo está publicado en Q1 o Q2, la probabilidad de que sea metodológicamente sólido y bien revisado es mucho mayor.

b) Factor de impacto (FI). Indica cuántas veces, en promedio, los artículos de la revista fueron citados en los últimos dos años.

- FI ≥ 2: Buena relevancia en la mayoría de las áreas clínicas.
- FI ≥ 5: Muy alto impacto en medicina interna, cardiología, infectología, etc.
- FI < 1: Relevancia baja (puede ser aceptable en áreas muy específicas o locales).

Un alto FI no garantiza que todos los artículos dentro de esa revista sean buenos; solo es un indicador de la calidad general.

c) Altmetrics. Miden la influencia social y digital del artículo. Este sistema integra menciones

https://doi.org/10.62486/978-9915-9928-3-9.Ch05

en redes sociales científicas, blogs, noticias, Wikipedia, descargas del texto completo y guardados en plataformas académicas. Un puntaje alto en Altmetrics indica que un artículo está generando interés y discusión, lo cual puede ayudar al lector a priorizar qué revisar primero cuando existen múltiples publicaciones recientes. No obstante, la visibilidad digital no es sinónimo de calidad metodológica; un artículo puede ser muy difundido por razones ajenas a su rigor científico. Por ello, los Altmetrics son útiles como complemento, pero nunca deben reemplazar la evaluación crítica del contenido.

El indicador central de este sistema es el AltmetricAttention Score, que agrupa todas estas interacciones en un solo valor numérico. Generalmente, este puntaje se representa mediante un círculo multicolor –el llamado "donut Altmetric"– que resume visualmente la magnitud y la procedencia del impacto. Cuanto mayor es el puntaje, mayor ha sido la atención generada por el artículo.

Un puntaje bajo suele reflejar escasa actividad digital y poca difusión más allá del ámbito académico tradicional. En cambio, un puntaje intermedio indica que el artículo ha comenzado a circular en redes sociales y en algunos espacios de divulgación. Finalmente, un puntaje alto representa un artículo que está recibiendo considerable atención pública, profesional o mediática, ya sea por la importancia de sus hallazgos, por su relación con temas de salud emergentes o, en ocasiones, por la controversia que genera. Es importante enfatizar que unAltmetric elevado no implica necesariamente rigor metodológico ni validez clínica, solo señala que el artículo es visible y ampliamente discutido.

El valor del Altmetrics radica, precisamente, en su capacidad para mostrar qué investigaciones están influyendo en la conversación científica y social en tiempo real. Para el profesional de la salud, esto puede ser útil en varios sentidos: permite identificar rápidamente estudios emergentes, priorizar lecturas en áreas con abundante producción reciente y comprender qué temas están captando la atención de la comunidad médica y del público general. Sin embargo, su utilidad se limita a orientar la selección preliminar de artículos. El AltmetricAttention Score nunca debe utilizarse como sustituto de la evaluación científica.

Por tanto, interpretar Altmetrics requiere comprender lo que mide y lo que no mide. Mide alcance, impacto comunicacional y nivel de discusión; no mide calidad metodológica, validez interna, fuerza de asociación, riesgo de sesgo ni aplicabilidad clínica. Por ello, su uso óptimo consiste en integrarlo como una herramienta complementaria que ayuda a decidir qué leer primero, dejando la decisión de qué aplicar en la práctica clínica al análisis crítico riguroso del estudio.

Fase 2. Evaluación crítica del artículo: Desglose del artículo para analizar la metodología, los posibles sesgos y la aplicabilidad de los resultados al contexto clínico. Preguntas útiles incluyen:

- ¿Quiénes son los autores y presentan algún conflicto de intereses?
- ¿El tipo de estudio y el nivel de evidencia son adecuados?
- ¿Los pacientes del estudio son comparables a mis pacientes?
- ¿Las conclusiones son clínicamente significativas y aplicables?

9. LA IMPORTANCIA DE UNA BÚSQUEDA CIENTÍFICA EFICIENTE

Realizar una búsqueda científica eficaz es una habilidad esencial en la MBE. Este proceso va más allá de obtener información; se trata de acceder a la evidencia científica más relevante y confiable para tomar decisiones clínicas informadas y seguras. A través de una búsqueda estructurada y precisa, el médico puede:

- Actualizarse continuamente: La medicina evoluciona rápidamente, y la capacidad de encontrar artículos científicos de calidad permite a los profesionales mantenerse al tanto de las últimas investigaciones y tratamientos.
- Tomar decisiones fundamentadas: La evidencia encontrada en bases de datos como PubMed, SciELO y BVS respalda las decisiones médicas y reduce la dependencia de opiniones no fundamentadas, mejorando la calidad de la atención al paciente.
- **Reducir la variabilidad en la práctica clínica**: Al utilizar artículos revisados y de alta calidad, el médico minimiza las variaciones en los tratamientos, ofreciendo una atención basada en evidencia científica y no en prácticas obsoletas o subjetivas.
- **Optimizar el tiempo de búsqueda**: Con el dominio de herramientas como los tesauros MeSH y DeCS, así como el uso de operadores booleanos y búsqueda con inteligencia artificial, el médico puede acceder rápidamente a información específica, sin perder tiempo revisando artículos irrelevantes.

https://doi.org/10.62486/978-9915-9928-3-9.Ch05

Dominar estas técnicas de búsqueda científica empodera al médico para tomar decisiones clínicas sólidas y mejorar su capacidad de respuesta a las necesidades cambiantes de sus pacientes y de la práctica médica. Con esta habilidad, el médico no solo podrá acceder a información relevante de forma eficiente, sino que también podrá aplicar el MBE en su contexto clínico, aumentando la seguridad y la efectividad de sus decisiones.

10. BÚSQUEDA CON INTELIGENCIA ARTIFICIAL

La Inteligencia Artificial (IA) puede acelerar y orientar la búsqueda, pero no sustituye la revisión en bases de datos científicas.

Las herramientas recomendadas son:

- **ChatGPT:** Para estructurar preguntas PICO y sugerir términos de búsqueda.
- **Elicit:** Encuentra artículos relevantes a partir de una pregunta.
- **Consensus:** Resume hallazgos con referencias directas.

Día a día aparecen nuevos buscadores por lo que se debe estar al tanto, además de confirmar la existencia del artículo en PubMed, BVS o en la plataforma correspondiente antes de citar. En otro apartado se hablará a detalle sobre la búsqueda de artículos con IA

11. PREGUNTAS DE AUTOEVALUACIÓN

PREGUNTAS DE OPCIÓN MÚLTIPLE (SELECCIÓN MÚLTIPLE)

1. ¿Cuál es la principal función de la estrategia PICO en la búsqueda científica?a) Organizar la información encontradab) Definir de forma estructurada la pregunta de investigaciónc) Filtrar artículos por fecha de publicaciónd) Clasificar las revistas científicas

2. En la búsqueda de artículos, el operador booleano OR se utiliza para:a) Reducir los resultados a artículos que contengan ambos términosb) Excluir términos irrelevantesc) Ampliar la búsqueda incluyendo cualquiera de los términosd) Buscar únicamente en revistas de alto impacto

3. ¿Cuál de los siguientes es un ejemplo de término MeSH?a) “High bloodpressure”b) Hypertensionc) Hipertensión arteriald) “Hypertension in elderlypatients”

4. Si quieres excluir artículos sobre niños en una búsqueda de hipertensión, el operador correcto sería:a) ANDb) NOT c) ORd) NEAR

5. En la evaluación rápida de una revista científica, un buen indicador de alta relevancia es:a) Factor de impacto (FI) ≥ 2 b) Publicación en cualquier idiomac) Tener muchos autores en cada artículod) Publicar únicamente casos clínicos

PREGUNTAS ABIERTAS

6. Explique en sus propias palabras por qué la formulación de la pregunta con PICO mejora la eficiencia de una búsqueda científica.

7. Mencione un ejemplo real de búsqueda utilizando operadores booleanos y explique su lógica.

8. ¿Qué ventajas ofrecen los términos MeSH y DeCS frente al uso de palabras libres en las búsquedas?

9. Enumere tres criterios básicos para evaluar la calidad de un artículo científico encontrado en una base de datos.

10. Describa una situación clínica en la que considere útil complementar la búsqueda clásica con herramientas de inteligencia artificial.

RESPUESTAS

1: b) Definir de forma estructurada la pregunta de investigación. La estrategia PICO organiza los elementos esenciales de la pregunta clínica (Paciente, Intervención, Comparación y Resultado), lo que permite orientar y optimizar la búsqueda.

2: c) Ampliar la búsqueda incluyendo cualquiera de los términos. OR recupera artículos que contengan al menos uno de los términos relacionados, útil especialmente para sinónimos o

https://doi.org/10.62486/978-9915-9928-3-9.Ch05

conceptos relacionados.

3: b) Hypertension. Es el término estandarizado utilizado por PubMed para representar el concepto de hipertensión como encabezado de tema.

4: b) NOT. NOT elimina resultados que contengan el término especificado después del operador.

5: a) Factor de impacto (FI) ≥ 2. En la mayoría de las áreas de salud, este valor sugiere buena visibilidad y rigurosidad editorial.

6. La estrategia PICO mejora la eficiencia de la búsqueda porque permite definir con claridad qué paciente se está estudiando, qué intervención o exposición se desea analizar, con qué se compara y qué resultado se busca. Al estructurar la pregunta, se identifican términos más precisos y relevantes, lo que reduce la cantidad de resultados irrelevantes y aumenta la posibilidad de encontrar evidencia directamente aplicable al caso clínico.

7. Hypertension AND Spironolactone NOT Children

La lógica es la siguiente:

- "Hypertension AND Spironolactone" asegura que aparezcan artículos donde se discutan ambos conceptos.
- "NOT Children" excluye estudios pediátricos.

Con esta combinación se obtiene evidencia específica sobre espironolactona en hipertensión en población adulta.

8. Los términos MeSH y DeCS ofrecen varias ventajas: estandarizan el lenguaje, evitan ambigüedades, agrupan sinónimos bajo un mismo descriptor y facilitan que la búsqueda recupere artículos relevantes aunque los autores utilicen diferentes términos. Esto mejora la precisión, reduce la variabilidad y garantiza una búsqueda más completa y reproducible.

9. Tres criterios esenciales son:

- Validez metodológica: diseño apropiado para responder la pregunta clínica, adecuada selección de la muestra y control de sesgos.
- Calidad de la revista: indexación en bases reconocidas, cuartil Q1–Q2 o factor de impacto adecuado.
- Aplicabilidad clínica: si los participantes, el entorno y los resultados del estudio son comparables al contexto del lector y permiten contribuir a decisiones clínicas reales.

10. Una situación clínica útil sería ante un caso complejo donde el médico requiere actualizarse rápidamente sobre opciones terapéuticas emergentes, por ejemplo, el uso de nuevos fármacos en insuficiencia cardíaca. La inteligencia artificial puede sugerir términos clave, sintetizar evidencia reciente, identificar estudios relevantes y acelerar la búsqueda inicial, permitiendo que el profesional dedique más tiempo a la lectura crítica del material encontrado.

https://doi.org/10.62486/978-9915-9928-3-9.Ch05

REFERENCIAS

Schardt C, Adams MB, Owens T, Keitz S, Fontelo P. Utilization of the PICO framework to improve searching PubMed for clinical questions. BMC Med Inform Decis Mak. 2007 Dec;7(1):16. Available at: https://doi.org/10.1186/1472-6947-7-16
Booth A, Sutton A, Papaioannou D. Systematic Approaches to a Successful Literature Review. 3rd ed. Los Angeles: SAGE Publications; 2021.
National Library of Medicine. Introduction to MeSH 2024. Bethesda, MD: U.S. National Library of Medicine; 2024 [cited 2026 Feb 9]. Available at: https://www.nlm.nih.gov/mesh/meshhome.html
BIREME/OPS/OMS. DeCS – Descriptores en Ciencias de la Salud. São Paulo: Biblioteca Virtual en Salud; 2024 [citado 2026 Feb 9]. Disponible en: https://decs.bvsalud.org/
Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al., editors. Cochrane Handbook for Systematic Reviews of Interventions [Internet]. 1st ed. Wiley; 2019 [cited 2026 Jan 12]. Available at: https://doi.org/10.1002/9781119536604
Muka T, Glisic M, Milic J, Verhoog S, Bohlius J, Bramer WM, et al. A 24-step guide on how to design, conduct, and successfully publish a systematic review and meta-analysis in medical research. Eur J Epidemiol. 2020;35(1):49–60.Available at: https://doi.org/10.1007/s10654-019-00576-5.
Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet. 2019;393(10181):1577–1579. Available at: https://doi.org/10.1016/S0140-6736(19)30037-6.
Tricco AC, Langlois EV, Straus SE, eds. Rapid Reviews to Strengthen Health Policy and Systems: A Practical Guide. Geneva: World Health Organization; 2017. Available at: https://doi.org/10.5668/9789240006305
Bramer WM, Rethlefsen ML, Kleijnen J, Franco OH. Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. Syst Rev. 2019;8:67. Available at: https://doi.org/10.1186/s13643-019-0981-8.
Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst Rev. 2019;8:163. Available at: https://doi.org/10.1186/s13643-019-1089-6

Chapter 6 / Capítulo 6

Critical reading of scientific articles on health: A practical guide (Spanish Edition)
ISBN: 978-9915-9928-3-9
DOI: 10.62486/978-9915-9928-3-9.Ch06
Pages: 51-69

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Búsqueda de artículos de investigación con inteligencia artificial

Searching for Research Articles with Artificial Intelligence

Mauricio Alejandro Paz Del Rio
Blas Apaza-Huanca

1 Clínica Los Andes. La Paz, Bolivia
2 Ministerio de Salud y Deportes. Instituto Académico Científico Quispe-Cornejo. La Paz, Bolivia.

ABSTRACT

Introduction: Artificial intelligence (AI) has transformed the search for scientific literature by optimizing the retrieval, organization, and synthesis of biomedical evidence. Integrated with the principles of Evidence-Based Medicine, it facilitates the formulation of clinical questions using the PICO format, the identification of MeSH/DeCS terms, and the design of reproducible search strategies. Development: This chapter analyzes the use of language models and specialized tools such as Perplexity, Elicit, Scite AI, Consensus, Semantic Scholar, ResearchRabbit, and OpenEvidence to improve the location and preliminary appraisal of scientific studies. It also emphasizes the construction of structured prompts, mandatory reference validation, and the integration of AI with traditional biomedical databases to ensure methodological rigor and traceability of evidence. Conclusions: AI constitutes a methodological assistant that increases the efficiency of the literature search process, but it does not replace clinical judgment, critical appraisal, or the verification of original sources. From a reflective perspective, its responsible use requires methodological skills, critical thinking, and ethical commitment to ensure clinical decisions are based on reliable, up-to-date, and reproducible scientific evidence.

Keywords: Literature search; Artificial intelligence; PICO format; MeSH descriptors; Apocryphal citation; Active systematic review

RESUMEN

Introducción: La inteligencia artificial (IA) ha transformado la búsqueda de literatura científica al optimizar la recuperación, organización y síntesis de la evidencia biomédica. Integrada con los principios de la Medicina Basada en la Evidencia, facilita la formulación de preguntas clínicas mediante el formato PICO, la identificación de términos MeSH/DeCS y el diseño de estrategias de búsqueda reproducibles. Desarrollo: El capítulo analiza el empleo de modelos de lenguaje y herramientas especializadas como Perplexity, Elicit, Scite AI, Consensus, Semantic Scholar, Research Rabbit y OpenEvidence para mejorar la localización y evaluación preliminar de estudios científicos. Asimismo, enfatiza la construcción de prompts estructurados, la validación obligatoria de referencias y la integración de la IA con bases de datos biomédicas tradicionales para garantizar rigor metodológico y trazabilidad de la evidencia. Conclusiones: La IA constituye un asistente metodológico que incrementa la eficiencia del proceso de búsqueda bibliográfica, pero no sustituye el juicio clínico, la lectura crítica ni la verificación de las fuentes originales. Desde una perspectiva reflexiva, su utilización responsable exige competencias metodológicas, pensamiento crítico y compromiso ético para asegurar decisiones clínicas fundamentadas en evidencia científica confiable, actualizada y reproducible.

Palabras clave: Búsqueda bibliográfica; Inteligencia artificial; Formato PICO; Descriptores MeSH; Cita apócrifa; Revisión sistemática viva

1. CASO CLÍNICO

Un residente de Medicina Interna atiende a un paciente masculino de 58 años con hipertensión arterial de difícil control, obesidad y apnea obstructiva del sueño (AOS). El médico sospecha de hiperaldosteronismo primario y necesita la mejor evidencia disponible sobre: prevalencia en pacientes con AOS. pruebas diagnósticas más exactas, algoritmos diagnósticos recientes. En el hospital, el tiempo es limitado y las búsquedas tradicionales en PubMed necesitan tiempo para resultados exactos. El residente decide utilizar herramientas de inteligencia artificial para acelerar la búsqueda y filtrar información científica de calidad.

2. INTRODUCCIÓN

La búsqueda eficiente de información científica es un pilar de la Medicina Basada en la Evidencia (MBE). Tradicionalmente, se ha dependido de bases de datos como PubMed, Embase, SciELO o la Biblioteca Virtual en Salud (BVS), utilizando estrategias de búsqueda con términos MeSH, DeCS y operadores booleanos. Sin embargo, en la era digital, la inteligencia artificial (IA) ofrece nuevas herramientas que agilizan la búsqueda, resumen la evidencia y permiten formular preguntas más precisas, combinando el rigor de las bases de datos con la rapidez del procesamiento de lenguaje natural.

Estas herramientas no reemplazan la búsqueda manual ni la evaluación crítica del artículo, pero pueden ahorrar tiempo, sugerir términos de búsqueda más eficientes, organizar la evidencia y poder comunicarse con el mismo artículo si se tiene dudas.

3. TÉRMINOS MESH, DECS Y OPERADORES BOOLEANOS CON LLM

En la búsqueda de literatura médica, el uso de términos controlados es fundamental para recuperar artículos relevantes y evitar pérdidas de información. En medicina, dos sistemas destacan por su uso estandarizado:

MeSH (Medical Subject Headings): es el tesauro desarrollado por la Biblioteca Nacional de Medicina de los Estados Unidos, utilizado por PubMed/MEDLINE para clasificar artículos bajo descriptores estandarizados.

DeCS (Descriptores en Ciencias de la Salud): es la adaptación de MeSH al español y portugués, desarrollada por BIREME/OPS/OMS, utilizada en LILACS y en la BVS.

La importancia de estos términos radica en que los autores no siempre utilizan la misma palabra para describir un concepto. Por ejemplo, un estudio sobre “ataque cardíaco” puede estar indexado bajo el descriptor “Infarto de Miocardio”. Si solo se busca “ataque cardíaco” en el texto libre, se podrían perder artículos relevantes.Para optimizar la recuperación de evidencia, los términos MeSH y DeCS suelen combinarse con operadores booleanos, estas herramientas constituyen el lenguaje formal de la búsqueda científica y permiten establecer relaciones lógicas entre conceptos:

- AND, para combinar conceptos distintos y restringir la búsqueda.
- OR, para agrupar sinónimos o términos equivalentes y ampliar la búsqueda.
- NOT, para excluir términos no deseados, con uso cauteloso.
- Paréntesis y comillas: permite controlar el orden lógico de la búsqueda y recuperar frases exactas, respectivamente.

Tradicionalmente, la identificación de descriptores adecuados y su correcta combinación requería experiencia y tiempo (como se vio en el capítulo 5). En la actualidad, los modelos de lenguaje grande (LLM), como ChatGPT, Gemini y otras inteligencias artificiales conversacionales, pueden facilitar este proceso al:

- Sugerir rápidamente términos MeSH y DeCS pertinentes.
- Proponer sinónimos y términos de texto libre utilizados en la literatura.
- Construir estrategias preliminares de búsqueda adaptadas a distintas bases de datos.

No obstante, es importante recordar que los LLM no sustituyen la comprensión metodológica del investigador, sino que actúan como herramientas de apoyo para acelerar y optimizar la búsqueda.

https://doi.org/10.62486/978-9915-9928-3-9.Ch06

4. FORMATO PICO CON LLM

El formato PICO (Paciente, Intervención, Comparador, Outcome) es el más utilizado porque organiza los elementos clave de la pregunta de forma estructurada. PICO significa**:**

- **P (Paciente o Problema):** grupo, condición o población que nos interesa.
- **I (Intervención o Exposición):** tratamiento, prueba diagnóstica o factor de riesgo que queremos investigar.
- **C (Comparador):** alternativa frente a la cual comparamos la intervención. Puede ser placebo, estándar de cuidado o ausencia de exposición.
- **O (Outcome o Desenlace):** resultado clínico o biomédico que queremos medir.

La solicitud precisa de esto en los LLM nos brinda muchas ventajas: Ahorra tiempo en la estructuración inicial de la pregunta, puede generar variantes de la pregunta adaptadas a diagnósticos, terapias o pronóstico, sugiere palabras clave para usar en la búsqueda, puede formular la pregunta en distintos idiomas.

5. PREGUNTA DE INVESTIGACIÓN CON LLM

Formular una pregunta de investigación bien estructurada constituye un paso central en la MBE, ya que determina el enfoque de la búsqueda bibliográfica, la selección de los estudios relevantes y la posterior interpretación crítica de la literatura. Una pregunta mal definida conduce, con frecuencia, a búsquedas imprecisas, resultados heterogéneos y conclusiones de limitada aplicabilidad clínica.

En este contexto, el formato PICO es la herramienta ampliamente utilizada para organizar la pregunta clínica de manera sistemática. Su valor no radica únicamente en facilitar una búsqueda más eficiente, sino en obligar al clínico o investigador a explicitar desde el inicio los elementos clave del problema, promoviendo un análisis estructurado y orientado a la toma de decisiones.

Los LLM pueden apoyar este proceso al actuar como asistentes metodológicos. Su utilidad principal reside en ayudar a descomponer un problema clínico complejo en sus componentes PICO, proponer formulaciones alternativas de la pregunta de investigación y sugerir términos clave asociados a cada componente. De este modo, los LLM contribuyen a reducir errores frecuentes, como preguntas excesivamente amplias o ambiguas, sin reemplazar el razonamiento clínico ni la evaluación crítica del investigador.

Es importante destacar que el uso de LLM en esta etapa debe entenderse como un apoyo estructural y no como una automatización del proceso intelectual. La validez y relevancia de la pregunta de investigación siguen dependiendo del juicio clínico, del contexto asistencial y del conocimiento previo del investigador.

6. CONSTRUCCIÓN DE UN PROMPT PARA LA BÚSQUEDA DE LITERATURA CIENTÍFICA CON INTELIGENCIA ARTIFICIAL

Uno de los errores más frecuentes al utilizar inteligencia artificial en la búsqueda de artículos científicos es formular instrucciones demasiado breves o inespecíficas. Expresiones como “dame los términos MeSH de hiperaldosteronismo” suelen producir respuestas incompletas, desordenadas o poco útiles para una búsqueda real en bases de datos biomédicas.

En investigación en salud, la calidad de los resultados obtenidos mediante inteligencia artificial depende directamente de la calidad del prompt, es decir, de la instrucción que se le proporciona a la herramienta. En este sentido, el prompt cumple una función metodológica similar a la de una pregunta PICO bien formulada: orienta el proceso, delimita el alcance y mejora la reproducibilidad de la búsqueda.Para que la IA aporte verdadero valor a la MBE, el prompt debe construirse de forma estructurada, explícita y orientada a objetivos científicos, no como una consulta informal.

6.1 PRINCIPIOS PARA UN PROMPT CIENTÍFICAMENTE SÓLIDO.

Un prompt adecuado para búsqueda bibliográfica debe integrar varios elementos clave.

En primer lugar, es fundamental definir el rol que debe asumir la IA. Cuando no se especifica, la IA tiende a responder desde un enfoque general o divulgativo. En el contextode la investigación clínica, se debe indicar que actúe como experto en búsqueda de literatura biomédica y en el uso

https://doi.org/10.62486/978-9915-9928-3-9.Ch06

de tesauros controlados, como MeSH y DeCS.

En segundo lugar, debe establecerse con claridad el **objetivo de la tarea**. El propósito no es aprender qué son los términos MeSH, sino optimizar una búsqueda real en bases de datos científicas como PubMed y la BVS, con fines clínicos o investigativos.

El tercer elemento es el **contexto clínico**, que debe ser específico y bien delimitado. Incluir la condición de interés, la población (por ejemplo, adultos o pediátricos) y el enfoque clínico evita ambigüedades y mejora la selección de descriptores. Un contexto mal definido puede conducir a resultados irrelevantes o excesivamente amplios.

En cuarto lugar, el prompt debe describir de manera explícita la **tarea concreta** que se espera que la IA realice. En una búsqueda bibliográfica rigurosa, esto implica solicitar la identificación de términos MeSH oficiales, sus equivalentes en DeCS, sinónimos y términos de texto libre, así como la construcción de estrategias de búsqueda con operadores booleanos. Además, es recomendable especificar el tipo de estudios de interés (por ejemplo, prevalencia, diagnóstico, estudios observacionales) y el periodo temporal prioritario, especialmente en un contexto de evidencia actualizada.

Figura 1. Elementos clave para un buen prompt

Fuente: Elaboración propia con apoyo de Inteligencia artificial

Es esencial indicar el **formato de salida**. En investigación, la información estructurada en tablas facilita la lectura, la comparación y la reproducibilidad. Solicitar desde el inicio una salida tabular transforma la respuesta de la IA en una herramienta práctica, lista para ser utilizada o documentada en un protocolo de investigación.

6.2 PROMPT INTEGRAL PARA ESTRATEGIA DE BÚSQUEDA Y FORMULACIÓN DE LA PREGUNTA DE INVESTIGACIÓN CON INTELIGENCIA ARTIFICIAL

Una vez comprendidos los principios para construir prompts de calidad, es posible integrar en una sola instrucción todos los componentes clave del proceso de búsqueda en MBE: la formulación de la pregunta de investigación, la estructuración PICO, la identificación de términos MeSH y DeCS, y la construcción de estrategias de búsqueda reproducibles en bases de datos biomédicas.

Este enfoque integral permite que la IA funcione como un asistente metodológico, capaz de acompañar al investigador desde el planteamiento del problema clínico hasta la ejecución de la búsqueda bibliográfica, sin sustituir el razonamiento crítico ni la validación humana.A continuación, se presenta un modelo de prompt completo, diseñado para ser utilizado de forma

https://doi.org/10.62486/978-9915-9928-3-9.Ch06

reproducible en distintos contextos clínicos.

Ejemplo de Prompt:

Actúacomo un experto en búsqueda de literatura biomédica, Medicina Basada en la Evidencia y metodología de investigación clínica.

Objetivo: ayudarme a formular una pregunta de investigación clara y a diseñar una estrategia de búsqueda rigurosa, reproducible y optimizada para PubMed y la Biblioteca Virtual en Salud, con fines de investigación clínica.

Contexto clínico:

Hiperaldosteronismo primario en pacientes adultos con apnea obstructiva del sueño enfocado a estudios de prevalencia y estudios de exactitud diagnóstica.

Tareas específicas:

- Formular la pregunta de investigación utilizando el enfoque PICO.
- Desglosar explícitamente los componentes P (Paciente), I (Intervención o exposición), C (Comparador) y O (Outcome).
- Proponer al menos dos variantes completas de la pregunta PICO, orientadas a prevalencia y diagnóstico.
- Identificar únicamente términos MeSH oficiales utilizados en PubMed y sus equivalentes DeCS validados en español.
- Incluir sinónimos, acrónimos y términos de texto libre comúnmente utilizados por los autores en los artículos científicos.
- Proponer una estrategia de búsqueda estructurada con operadores booleanos optimizada para PubMed y otra para la BVS, listas para ser copiadas y ejecutadas.
- Indicar, cuando sea posible, el tipo de estudios que cada término suele recuperar (por ejemplo: estudios de prevalencia, estudios observacionales, estudios de exactitud diagnóstica).

Formato de salida:

Presenta la información de manera estructurada en tablas, incluyendo: Tabla PICO (P, I, C, O, términos MeSH/DeCS asociados). Tabla de descriptores con las siguientes columnas:

- Descriptor MeSH
- Descriptor DeCS
- Sinónimos o términos de texto libre
- Tipo de estudio asociado
- Estrategia de búsqueda sugerida para PubMed y BVS.

Figura2.Prompt para estrategia de búsqueda y formulación de la pregunta de investigación con Inteligencia artificial

Fuente: Elaboración propia con apoyo de inteligencia artificial

https://doi.org/10.62486/978-9915-9928-3-9.Ch06

Este tipo de prompt no solo mejora la calidad y pertinencia de los resultados obtenidos, sino que también forma al lector en pensamiento metodológico, permitiéndole comprender que la inteligencia artificial no reemplaza al investigador. Por el contrario, actúa como una herramienta que amplifica su capacidad para formular preguntas relevantes, estructurar búsquedas complejas y filtrar evidencia científica de manera eficiente.En el contexto de la MBE, aprender a construir prompts de este nivel constituye una competencia metodológica esencial, comparable a saber formular una pregunta PICO o diseñar una estrategia de búsqueda avanzada en PubMed.

6.3PROMPT INTEGRAL PARA BÚSQUEDA DE ARTÍCULOS DE INVESTIGACIÓN CON LLM

Además de ayudar a formular PICO, identificar MeSH/DeCS y construir cadenas booleanas, los LLM pueden emplearse de manera más "operativa": pedirles que propongan directamente un conjunto inicial de artículos relevantes a partir de una pregunta clínica o de investigación. En este rol, el LLM actúa como un asistente de cribado temprano, que sugiere literatura probable (revisiones sistemáticas, metaanálisis, ensayos clínicos, estudios de cohorte, etc.) y organiza la información en un formato útil para lectura rápida.La ventaja principal es pragmática: en escenarios clínicos con limitación de tiempo, un LLM puede acelerar tres tareas que suelen ser lentas cuando se realizan manualmente:

- Priorización inicial de evidencia: sugiere tipos de estudio y fuentes que suelen ofrecer mayor valor (p. ej., revisiones sistemáticas recientes y ensayos clínicos clave).
- Extracción de información "de primera pasada": resume pregunta, población, comparadores y desenlaces principales, permitiendo decidir rápidamente qué artículos vale la pena leer completos.
- Organización comparativa: presenta los estudios en tabla y facilita comparar resultados (eficacia, seguridad, subgrupos, limitaciones).

Sin embargo, este uso directo exige una advertencia metodológica: un LLM puede equivocarse en detalles bibliográficos (autores, año, DOI) o proponer artículos que no responden exactamente a la pregunta. Por tanto, el uso correcto en MBE es: LLM para generar una lista inicial más verificación obligatoria en la base de datos (PubMed/Scopus/Cochrane/otros) antes de citar o concluir. Dicho de otro modo, el LLM acelera el triage, pero la validez final depende de la comprobación en la fuente.

REDACCIÓN DEL PROMPT

Para que esta modalidad funcione, el prompt debe dejar claro que el objetivo no es solo una explicación, sino una curaduría de evidencia con trazabilidad. En particular, sumado a las características descritas previamente sobre el prompt, debe exigir:

- Delimitación explícita: periodo temporal (p. ej., últimos 5 años), población, comparadores, desenlaces.
- Jerarquización de evidencia: priorizar revisiones sistemáticas/metaanálisis, ensayos y luego estudios observacionales de alta calidad.
- Metadatos obligatorios: título, revista, año, tipo de estudio, hallazgos, y enlace/identificador verificable.
- Estructura de salida: tabla más síntesis crítica comparativa.
- Control de calidad: indicar que solo incluya artículos con enlace verificable (PubMed ID/URL, DOI), y que marque cualquier dato incierto.

FORMATO DE PROMPT

Actúa como un médico investigador experto en Medicina Interna, Endocrinología (dependiendo de la rama del trabajo que se está buscando) y Medicina Basada en la Evidencia, con experiencia en búsqueda avanzada y síntesis crítica de literatura científica.

Objetivo: Ayudarme a identificar y sintetizar la mejor evidencia científica disponible para orientar la toma de decisiones clínicas y diagnósticas.

Contexto clínico: Paciente masculino de 58 años con hipertensión arterial de difícil control, obesidad y apnea obstructiva del sueño. Se sospecha hiperaldosteronismo primario como causa secundaria de hipertensión.

Tareas: Necesito información sobre: Prevalencia del hiperaldosteronismo primario en

https://doi.org/10.62486/978-9915-9928-3-9.Ch06

pacientes con apnea obstructiva del sueño, pruebas diagnósticas con mayor exactitud para el diagnóstico de hiperaldosteronismo primario en este contexto, algoritmos diagnósticos recientes que integren hipertensión, apnea obstructiva del sueño e hiperaldosteronismo primario.

INSTRUCCIONES ESPECÍFICAS:

– Realiza una búsqueda dirigida de la mejor evidencia publicada en los últimos 5 a 10 años, priorizando revisiones sistemáticas, metaanálisis, ensayos clínicos y estudios observacionales de alta calidad.

– Selecciona entre 8 y 10 artículos relevantes provenientes de revistas indexadas en PubMed, Scopus o Cochrane.

– Para cada artículo incluye obligatoriamente:

- Título del artículo
- Autor principal
- Revista y año de publicación
- Tipo de estudio
- Principales hallazgos relacionados con prevalencia, exactitud diagnóstica o algoritmos diagnósticos
- Enlace verificable (PMID o DOI).

– Organiza la información en una tabla comparativa para facilitar la lectura.

– Finaliza con una síntesis crítica, integrando los hallazgos y destacando:

- Magnitud de la prevalencia reportada en pacientes con AOS.
- Pruebas diagnósticas con mejor rendimiento (sensibilidad, especificidad, valor clínico).
- Consensos actuales y limitaciones de la evidencia disponible.
- Indica explícitamente cualquier dato que requiera verificación adicional.

FORMATO DE SALIDA:

Figura 3.Prompt para búsqueda de artículos científicos con Inteligencia Artificial

Fuente: Elaboración propia con apoyo de inteligencia artificial

Tabla comparativa con las siguientes columnas: Título del artículo. Autor principal. Revista (año). Tipo de estudio. Población estudiada. Hallazgos clave (prevalencia / pruebas diagnósticas / algoritmo). Enlace verificable (PMID o DOI). Síntesis crítica final, estructurada en párrafos breves, que incluya:

- Magnitud de la prevalencia del hiperaldosteronismo primario reportada en pacientes con

https://doi.org/10.62486/978-9915-9928-3-9.Ch06

apnea obstructiva del sueño.

- Pruebas diagnósticas con mejor rendimiento (sensibilidad, especificidad y valor clínico).
- Consensos actuales sobre algoritmos diagnósticos y puntos de controversia.
- Limitaciones metodológicas de la evidencia disponible y brechas de conocimiento.

Nota de verificación: Indica explícitamente cualquier dato bibliográfico, resultado o afirmación que requiera verificación adicional en la fuente original.

6. USO DE OTRAS IA PARA ARTÍCULOS DE INVESTIGACIÓN

En los últimos años, han surgido diversas herramientas de IA diseñadas específicamente para agilizar la búsqueda, organización y síntesis de información científica. A diferencia de buscadores generales, estas plataformas están optimizadas para trabajar con literatura revisada por pares, bases de datos académicas y análisis de citaciones. En este acápite, analizaremos las más relevantes aplicadas a nuestro caso clínico y pregunta PICO.

A. PERPLEXITY AI

Perplexity integra un motor de búsqueda con un modelo de lenguaje que responde en formato conversacional y, de manera característica, cita explícitamente las fuentes consultadas. Esta funcionalidad lo convierte en una herramienta especialmente útil para obtener una visión rápida y orientativa del estado del arte, identificar revisiones recientes y acceder de forma directa a enlaces de PubMed y otros repositorios académicos.

En investigación clínica, Perplexity funciona mejor como herramienta de exploración inicial. Permite reconocer con rapidez qué líneas de evidencia existen, cuáles son los autores o grupos más citados y qué artículos actúan como referencias centrales del tema. No obstante, sus resultados deben considerarse un punto de partida, ya que la búsqueda final, exhaustiva y reproducible debe realizarse directamente en las bases de datos biomédicas. Es especialmente útil para orientarse rápidamente, identificar revisiones recientes y localizar artículos “semilla” que luego pueden ampliarse mediante búsqueda clásica y revisión de referencias. No debe utilizarse como único método cuando el objetivo es una revisión exhaustiva o sistemática.

Aplicación al caso clínico: Si se introduce una pregunta derivada del problema planteado al inicio del capítulo, por ejemplo:

“En adultos con apnea obstructiva del sueño, ¿cuál es la prevalencia de hiperaldosteronismo primario en comparación con adultos sin apnea obstructiva del sueño?”

Perplexity suele devolver un resumen narrativo acompañado de enlaces a PubMed y otras fuentes académicas. Una estrategia recomendable es solicitar explícitamente academicsources y pedir que los resultados se ordenen por fecha o relevancia, según el objetivo de la búsqueda.

Tabla 1 Ventajas y limitaciones de Perplexity AI para la búsqueda inicial de literatura científica

Ventajas	Limitaciones
Resultados rápidos, útiles en contextos clínicos con tiempo limitado.	No garantiza una búsqueda exhaustiva ni sistemática.
Enlaces directos a fuentes académicas (PubMed, editoriales).	Puede omitir artículos clave o relevantes.
Permite refinar la búsqueda mediante preguntas sucesivas.	Tiende a priorizar estudios más visibles o citados.
Facilita una visión global del estado del arte.	No evalúa la calidad metodológica de los estudios.
Útil para identificar artículos “semilla”.	Requiere verificación manual de datos y resultados.

Fuente: Elaboración propia con apoyo de inteligencia artificial

B. ELICIT

Elicit es una herramienta de inteligencia artificial diseñada específicamente para apoyar la búsqueda y síntesis de literatura científica, con un enfoque claro en investigación en salud y ciencias biomédicas. A diferencia de motores de búsqueda generales, Elicit trabaja directamente sobre bases de datos académicas (principalmente PubMed y Semantic Scholar) y estructura la

https://doi.org/10.62486/978-9915-9928-3-9.Ch06

información en tablas, lo que facilita el análisis comparativo.

En investigación clínica, Elicit resulta especialmente útil cuando el objetivo es identificar estudios relevantes de manera estructurada, extraer información clave de cada artículo y organizarla de forma sistemática. Su fortaleza principal radica en la capacidad de descomponer preguntas de investigación y vincularlas con evidencia primaria, reduciendo el tiempo invertido en la lectura inicial de múltiples resúmenes.

Al aplicar la pregunta mencionada del caso clínico, Elicit puede recuperar estudios observacionales, revisiones y ensayos relevantes, mostrando para cada uno información estructurada como población, variables evaluadas y principales resultados. Esto permite al personal de salud identificar rápidamente qué artículos ameritan una lectura completa y cuáles no aportan información directa a la pregunta clínica.

Se recomienda usarla cuando se necesita organizar evidencia primaria de forma estructurada, por ejemplo, en la fase inicial de un protocolo de investigación o al comparar resultados entre estudios observacionales. No sustituye la búsqueda exhaustiva en PubMed ni la lectura crítica completa, pero constituye un excelente complemento para priorizar artículos relevantes.

Tabla 2 Ventajas y limitaciones de Elicit para la identificación y organización inicial de estudios científicos

Ventajas	Limitaciones
Orientada específicamente a literatura científica.	No cubre todas las revistas ni bases de datos existentes.
Presenta resultados en tablas estructuradas.	Puede omitir artículos recientes aún no indexados.
Facilita la comparación rápida entre estudios.	Depende de la calidad de los metadatos disponibles.
Útil para identificar estudios observacionales y revisiones.	Menor utilidad para guías clínicas o documentos de consenso.
Ahorra tiempo en la lectura inicial de resúmenes.	Requiere verificación manual de resultados y conclusiones.

Fuente: Elaboración propia con apoyo de inteligencia artificial

C. SCITE AI

Scite AI es una herramienta de inteligencia artificial orientada a un aspecto específico y crítico de la investigación científica: cómo se citan los artículos en la literatura. A diferencia de otras plataformas que priorizan la recuperación o el resumen de estudios, Scite analiza el contexto de las citaciones y las clasifica según su función dentro del artículo citante. Las categoriza las citas como:

- Supporting (citan el artículo para respaldar un hallazgo),
- Contrasting (lo citan para cuestionar o contradecir resultados),
- Mentioning (lo mencionan sin posicionarse críticamente).

Esta característica aporta un valor metodológico importante, ya que permite al investigador identificar no solo qué artículos son citados con frecuencia, sino cómo es utilizada esa evidencia dentro del cuerpo de la literatura científica.

Al aplicarlo en el caso clínico, esta IA puede utilizarse después de identificar artículos clave (por ejemplo, estudios de prevalencia o exactitud diagnóstica). El personal de salud puede analizar si dichos artículos son mayoritariamente citados como evidencia de apoyo o si existen estudios que cuestionan sus conclusiones, lo cual ayuda a valorar la consistencia y robustez de los hallazgos.

Se recomienda utilizarlo después de haber identificado artículos relevantes mediante otras herramientas o búsquedas clásicas. Su mayor valor reside en apoyar la lectura crítica, ayudando a discernir si la evidencia es consistente, controvertida o débilmente sustentada. No debe utilizarse como única herramienta de búsqueda, sino como complemento para evaluar la solidez del cuerpo de evidencia.

https://doi.org/10.62486/978-9915-9928-3-9.Ch06

Tabla 3 Ventajas y limitaciones de Scite AI para el análisis del contexto de citación en literatura científica

Ventajas	Limitaciones
Analiza el contexto de las citaciones, no solo su número.	No es un motor de búsqueda primario de literatura.
Permite identificar rápidamente evidencia apoyada o controvertida.	Depende de la disponibilidad de textos completos indexados.
Útil para detectar artículos influyentes y debates científicos.	No evalúa directamente la calidad metodológica del estudio.
Facilita la lectura crítica al mostrar citas contrastantes.	Menor utilidad en temas con poca literatura publicada.
Aporta valor en fases avanzadas de análisis.	Requiere complementar con lectura directa del artículo.

Fuente: Elaboración propia con apoyo de inteligencia artificial

D. CONSENSUS

Consensus es una herramienta de inteligencia artificial orientada a responder preguntas clínicas o de investigación basándose exclusivamente en literatura científica revisada por pares. A diferencia de motores de búsqueda generales, Consensus no se apoya en contenido web amplio, sino que extrae respuestas a partir de artículos académicos, con énfasis en revisiones sistemáticas, metaanálisis y ensayos clínicos.

Su principal característica es que intenta sintetizar una respuesta basada en el "consenso" de la evidencia disponible, mostrando si los estudios tienden a apoyar, rechazar o no ser concluyentes respecto a una afirmación específica. Esto la convierte en una herramienta particularmente atractiva para clínicos que necesitan una respuesta rápida y basada en evidencia, sin perder el anclaje científico.

Ante la aplicación al caso clínico este puede ofrecer una respuesta resumida indicando si la mayoría de los estudios disponibles apoyan dicha asociación, junto con enlaces a los artículos que sustentan esa conclusión. Para el personal de salud, esto permite identificar rápidamente si existe respaldo científico consistente para la sospecha clínica y cuáles son los estudios clave que fundamentan dicha relación.

Se recomienda Consensus cuando se necesita una respuesta rápida y orientativa, basada en evidencia científica, para confirmar si una hipótesis clínica cuenta con respaldo en la literatura. Resulta ideal en etapas tempranas de toma de decisiones o como complemento a la lectura crítica, pero no debe reemplazar el análisis detallado de los estudios ni la construcción de una búsqueda exhaustiva cuando el objetivo es investigación formal.

Tabla 4 Ventajas y limitaciones de Consensus para la síntesis rápida de evidencia científica

Ventajas	Limitaciones
Respuestas basadas exclusivamente en literatura científica.	No sustituye una búsqueda exhaustiva y sistemática.
Enfoque en consenso de la evidencia disponible.	Puede simplificar en exceso temas complejos o heterogéneos.
Útil para responder preguntas clínicas directas.	Menor profundidad en detalles metodológicos de cada estudio.
Proporciona enlaces a los artículos que sustentan la respuesta.	Dependiente de la cantidad y calidad de estudios disponibles.
Ahorra tiempo en la toma de decisiones preliminares.	Requiere lectura crítica posterior de los artículos originales.

Fuente: Elaboración propia con apoyo de inteligencia artificial

E. SEMANTIC SCHOLAR

Semantic Scholar es un motor de búsqueda académica impulsado por inteligencia artificial cuyo objetivo principal es priorizar relevancia científica, más que coincidencia literal de palabras clave. A diferencia de PubMed, Semantic Scholar incorpora análisis de citaciones, influencia de los artículos y relaciones entre trabajos, lo que facilita identificar estudios fundamentales dentro de un campo.

En investigación clínica, su mayor valor reside en ayudar al investigador a detectar artículos "núcleo", es decir, aquellos que han tenido mayor impacto y que estructuran el conocimiento actual sobre un tema.

Al usar esta plataforma en el caso clínico, permitirá evidenciar estudios más citados sobre

https://doi.org/10.62486/978-9915-9928-3-9.Ch06

la asociación entre AOS e hiperaldosteronismo, además de revisiones clave que han influido en guías o algoritmos diagnósticos y redes de citación que conectan estudios endocrinológicos y cardiovasculares.

Es ideal para identificar artículos clave e influyentes y para ampliar la búsqueda mediante referencias y citaciones, pero no debe sustituir a PubMed cuando se requiere una búsqueda clínica sistemática.

Tabla 5 Ventajas y limitaciones de Semantic Scholar para la identificación de artículos influyentes y redes de citación

Ventajas	Limitaciones
Prioriza artículos influyentes y altamente citados.	No utiliza MeSH ni DeCS como PubMed.
Muestra relaciones entre artículos y citaciones clave.	Puede omitir literatura regional o latinoamericana.
Útil para identificar trabajos seminales.	No está orientado a búsquedas clínicas estructuradas.
Incluye métricas de impacto y relevancia.	Requiere complementar con PubMed/BVS para exhaustividad.

F. CONNECTEDPAPERS

ConnectedPapers es una herramienta de inteligencia artificial que genera mapas visuales de literatura científica, mostrando cómo se relacionan los artículos entre sí según patrones de citación y similitud temática. Su enfoque no es responder preguntas clínicas directas, sino ayudar a comprender la estructura y evolución del conocimiento dentro de un área específica.

Si aplicamos al caso clínico se podrá visualizar cómo han evolucionado las investigaciones sobre hipertensión secundaria, identificara artículos “puente” entre Endocrinología, Medicina del Sueño y Cardiología, finalmente detectara estudios relevantes que no aparecen en búsquedas tradicionales.

Es especialmente útil para revisiones narrativas, marcos teóricos e introducciones, pero no para síntesis clínica inmediata.

Tabla 6 Ventajas y limitaciones de ConnectedPapers para la visualización de relaciones entre artículos científicos

Ventajas	Limitaciones
Visualiza relaciones entre artículos de forma intuitiva.	No responde preguntas clínicas directas.
Facilita la comprensión de la evolución del tema.	No evalúa calidad metodológica.
Ayuda a identificar artículos poco visibles pero relevantes.	Menor utilidad en decisiones clínicas urgentes.
Útil para introducciones y revisiones narrativas.	Requiere partir de al menos un artículo inicial.

Fuente: Elaboración propia con apoyo de inteligencia artificial

G. RESEARCH RABBIT

Research Rabbit es una plataforma diseñada para el descubrimiento continuo de literatura científica. Permite crear colecciones de artículos y, a partir de ellas, sugiere nuevos estudios relacionados mediante algoritmos de recomendación. A diferencia de otras herramientas, Research Rabbit está pensada para el seguimiento longitudinal de un tema, más que para búsquedas puntuales.

Si aplicamos en el caso clinico nos permite mantenernos actualizados sobre nuevas publicaciones, identificar tendencias emergentes en diagnóstico o fisiopatología y descubrir estudios recientes relacionados con los ya seleccionados.

Research Rabbit es ideal para investigadores activos que desean mantenerse actualizados en un tema específico, más que para resolver preguntas clínicas urgentes.

https://doi.org/10.62486/978-9915-9928-3-9.Ch06

Tabla 7 Ventajas y limitaciones de Research Rabbit para el seguimiento continuo y descubrimiento de literatura científica

Ventajas	Limitaciones
Facilita el seguimiento continuo de un tema.	No está orientada a síntesis clínica inmediata.
Sugiere artículos relacionados automáticamente.	Depende de una selección inicial adecuada.
Útil para proyectos a mediano y largo plazo.	No sustituye una búsqueda estructurada.
Favorece el descubrimiento de literatura emergente.	Menor control metodológico de la búsqueda.

Fuente: Elaboración propia con apoyo de inteligencia artificial

7. RESOLUCIÓN DE PREGUNTAS COMPLEJAS CON OPENEVIDENCE

En la práctica clínica y en la investigación, no todas las preguntas son sencillas. Muchas veces el médico o el investigador necesita responder preguntas multidimensionales, que incluyen más de una variable o combinan varios aspectos: prevalencia, exactitud diagnóstica, pronóstico y tratamiento.

Las búsquedas tradicionales pueden requerir varias horas o días para recopilar, filtrar y analizar la literatura necesaria. Aquí es donde OpenEvidence se convierte en una herramienta valiosa: combina un modelo de lenguaje entrenado con literatura biomédica de alta calidad y la capacidad de devolver respuestas sintetizadas con referencias verificables.

¿QUÉ ES OPENEVIDENCE?

Es una plataforma de inteligencia artificial especializada en ciencias de la salud que:

- Accede a literatura científica indexada (PubMed, PMC, entre otras).
- Responde en lenguaje natural, pero con citas y enlaces directos a los artículos.
- Permite refinar la búsqueda con filtros (año, tipo de estudio, población, etc.).
- Integra información de guías clínicas, revisiones sistemáticas y ensayos.

Su mayor fortaleza está en responder preguntas clínicas complejas con soporte bibliográfico inmediato.

Aplicación al caso clínico paso a paso en OpenEvidence:

- **Ingreso de la pregunta completa:** podemos colocar la pregunta PICO tal cual está redactada, sin necesidad de abreviarla.
- **Revisión del resumen generado:**OpenEvidence devuelve un párrafo con la síntesis de resultados,
- **Validación de referencias:** Cada afirmación viene acompañada de un enlace directo a PubMed o PMC, lo que permite comprobar la fuente original.

Ejemplo de respuesta (simulada): En adultos con apnea obstructiva del sueño (AOS), la prevalencia de hiperaldosteronismo primario (PA) varía según la población estudiada, los criterios diagnósticos y la coexistencia de hipertensión. La literatura médica reciente muestra que la prevalencia de PA en pacientes con AOS e hipertensión oscila entre 11% y 27%.[1-4] Un metaanálisis reciente reporta una prevalencia agrupada de PA del 27% (IC 95%: 24–29%) en pacientes con AOS, significativamente mayor que en pacientes sin AOS (OR ≈ 2).[1] Otros estudios en cohortes de pacientes con AOS e hipertensión reportan prevalencias de PA entre 11% y 21%, con cifras más altas en hombres, pacientes jóvenes, con obesidad o con AOS moderada a severa.[2-4]

1. Primary Aldosteronism and Obstructive Sleep Apnea: A Meta-Analysis of Prevalence and Metabolic Characteristics.
Zhang R, Cai X, Lin C, et al.
Sleep Medicine. 2024;114:8-14. doi:10.1016/j.sleep.2023.12.007.

2. Prevalence of Primary Aldosteronism in Patients With Concomitant Hypertension and Obstructive Sleep Apnea, Baseline Data of a Cohort.
Heizhati M, Aierken X, Gan L, et al.
Hypertension Research : Official Journal of the Japanese Society of Hypertension. 2023;46(6):1385-1394. doi:10.1038/s41440-023-01226-w.

3. The Prevalence of Obstructive Sleep Apnea in Patients With Primary Aldosteronism.
Sheu JY, Chang LY, Chen JY, et al.
The Journal of Clinical Endocrinology and Metabolism. 2024;109(10):2681-2691. doi:10.1210/clinem/dgae415.

4. Implementation of a Formal Sleep Center-Based Screening Protocol for Primary Aldosteronism in Patients With Obstructive Sleep Apnea.
Wright K, Mahmoudi M, Agrawal N, et al.
Surgery. 2023;173(1):59-64. doi:10.1016/j.surg.2022.05.041.

Tabla 8 Ventajas y limitaciones de OpenEvidence para la respuesta rápida a preguntas clínicas basadas en evidencia

Ventajas	Limitaciones
Orientada específicamente a preguntas clínicas.	No sustituye revisiones sistemáticas formales.
Proporciona respuestas rápidas con respaldo bibliográfico.	Menor profundidad metodológica por estudio.
Útil en contextos asistenciales con poco tiempo.	Dependiente de la disponibilidad de evidencia publicada.
Refuerza la toma de decisiones basada en evidencia.	Requiere lectura crítica posterior.

Fuente:Elaboración propia con apoyo de inteligencia artificial

Recomendación: Usar OpenEvidence cuando la pregunta involucra varios componentes clínicos, el tiempo es limitado y se necesita una visión de conjunto y se requiere citar referencias concretas en un protocolo o artículo.

Evitar usarlo como única fuente; siempre complementar con búsqueda manual en bases de datos y evaluación crítica de cada estudio.

8. RECOMENDACIONES PARA LA BÚSQUEDA DE ARTÍCULOS CON INTELIGENCIA ARTIFICIAL

El uso de inteligencia artificial en la búsqueda de literatura científica puede ser una ventaja competitiva para médicos, investigadores y estudiantes. Sin embargo, su integración debe hacerse de forma planificada, crítica y complementaria a las estrategias tradicionales de búsqueda en bases de datos.

Las siguientes recomendaciones tienen como objetivo maximizar la utilidad y minimizar los riesgos asociados.

8.1. COMBINAR LA IA CON BÚSQUEDA TRADICIONAL

Las herramientas de IA no reemplazan el acceso directo a bases de datos como PubMed, Embase, SciELO o BVS. La estrategia más segura es:

- Usar la IA para definir la pregunta de investigación, identificar términos MeSH/DeCS y generar estrategias preliminares.
- Ejecutar la búsqueda directamente en las bases de datos para verificar la exhaustividad de los resultados.

8.2. VERIFICAR SIEMPRE LAS REFERENCIAS

Algunas IA pueden “alucinar” con referencias: inventar artículos, autores o DOI. Recomendación práctica:

- Revisar cada referencia en PubMed o la base de datos original.
- Confirmar fecha, autores, revista y DOI.
- Guardar las referencias validadas en un gestor bibliográfico (Zotero, Mendeley, EndNote).

8.3. DOCUMENTAR LA ESTRATEGIA DE BÚSQUEDA

La reproducibilidad es clave en MBE. Cuando uses IA:

- Guarda capturas de pantalla del prompt y la respuesta.
- Registra los términos y operadores utilizados.
- Documenta filtros aplicados (fecha, idioma, tipo de estudio).

https://doi.org/10.62486/978-9915-9928-3-9.Ch06

8.4. EVALUAR LA APLICABILIDAD CLÍNICA

La IA puede mostrar estudios de diferentes contextos geográficos o poblaciones. Antes de tomar decisiones clínicas o diseñar un protocolo, verificar:

- ¿La población estudiada es similar a la de mi contexto?
- ¿Las pruebas diagnósticas o intervenciones son viables en mi medio?
- ¿El nivel de atención y recursos disponibles son comparables?

8.5. RECONOCER LOS SESGOS PROPIOS DE LA IA

Las herramientas de IA reflejan las limitaciones de los datos con los que fueron entrenadas:

- Si prioriza estudios recientes, podría omitir evidencia histórica clave.
- Si el entrenamiento se basó mayormente en literatura anglosajona, podría subrepresentar investigaciones latinoamericanas o de países en desarrollo.

8.6. USAR IA DIFERENTE PARA OBJETIVOS DIFERENTES

Tabla 9 Herramientas de inteligencia artificial según el objetivo y momento de uso en la búsqueda de literatura científica

Objetivo clínico o investigativo	Herramienta de IA recomendada	Momento de uso en el proceso	Nivel de rigor esperado	Comentario clave para el lector
Obtener una visión rápida del estado del arte	Perplexity AI	Inicio de la búsqueda	Bajo–moderado	Útil para orientarse rápidamente y localizar artículos “semilla”; requiere verificación posterior.
Identificar y organizar estudios relevantes en tablas	Elicit	Búsqueda inicial–intermedia	Moderado	Facilita comparación entre estudios y priorización de lecturas; no sustituye PubMed.
Evaluar cómo se utiliza un artículo en la literatura	Scite AI	Análisis crítico	Moderado–alto	Permite identificar evidencia apoyada o controvertida; complementa la lectura crítica.
Conocer si existe consenso científico sobre una pregunta	Consensus	Exploración clínica rápida	Moderado	Ideal para confirmar respaldo general de una hipótesis; puede simplificar temas complejos.
Identificar artículos influyentes y trabajos seminales	Semantic Scholar	Exploración intermedia	Moderado	Útil para detectar estudios clave y redes de citación; no es búsqueda clínica estructurada.
Visualizar relaciones y evolución del conocimiento	Connected Papers	Revisión narrativa / marco teórico	Bajo–moderado	Excelente para comprender la evolución del tema y detectar artículos poco visibles.
Seguimiento continuo y descubrimiento de nueva literatura	Research Rabbit	Proyectos a mediano y largo plazo	Bajo–moderado	Ideal para mantenerse actualizado; depende de una buena selección inicial.
Responder rápidamente preguntas clínicas con respaldo científico	OpenEvidence	Práctica clínica / confirmación rápida	Moderado	Útil en contextos asistenciales; no reemplaza revisiones sistemáticas ni lectura crítica.
Búsqueda exhaustiva, reproducible y documentable	PubMed / BVS (método clásico)	Etapa final / investigación formal	Alto	Estándar de oro para protocolos, revisiones sistemáticas y publicaciones científicas.

Fuente: Elaboración propia con apoyo de inteligencia artificial

8.7. FOMENTAR EL PENSAMIENTO CRÍTICO

La IA es una herramienta de apoyo, no un sustituto del juicio clínico. Se debe asegurar de leer los artículos originales, evaluar la calidad metodológica de los estudios (usa guías CONSORT, PRISMA, STROBE), interpretar los resultados en el contexto del paciente o la población objetivo.

Como consejo final, la verdadera fortaleza de la inteligencia artificial no está en “buscar por ti”, sino en amplificar su capacidad de buscar, filtrar, sintetizar y aplicar la evidencia. Usada de forma estratégica y combinada con métodos tradicionales, la IA puede reducir drásticamente el tiempo de búsqueda y aumentar la precisión de los resultados, siempre que se mantenga la verificación humana y la lectura crítica.

https://doi.org/10.62486/978-9915-9928-3-9.Ch06

8.8 EL RIESGO ÉTICO Y LA FALSIFICACIÓN DE LA EVIDENCIA ASISTIDA POR IA

El término técnico "alucinación" (generación de texto sin base factual) es un término Estratégico utilizado por la cita Apócrifa (del griego, apókrūphos: oculto, falso, de origen incierto), es una referencia bibliográfica completa (autor, título, revista, DOI) generada por un LLM que parece auténtica, pero que no existe o remite a un contenido irrelevante.

Esto ocurre porque el LLM no busca la referencia; la predice basándose en patrones de lenguaje. Sabe que tras una afirmación metodológica debe venir una cita en formato APA, y la inventa siguiendo ese patrón. Este proceso es invisible para el clínico que confía. Esto conlleva a negligencia clínica.

Tabla 10 Riesgos profesionales y éticos asociados al uso de referencias falsas generadas por inteligencia artificial

Riesgo Profesional	Escenario de Falsificación	Consecuencia Ética
Error Tipo I por Datos Falsos	La IA alucina una referencia que afirma una tasa de efectividad de un 90% para un tratamiento.	El clínico prescribe el tratamiento basándose en un dato inflado y falso, exponiendo al paciente al riesgo sin el beneficio esperado.
Omisión de Alerta	La IA, al resumir un estudio, omite o tergiversa los eventos adversos graves, citando una referencia apócrifa.	El clínico no considera el riesgo de daño (Error Tipo S/Daño) porque la alerta de seguridad fue falsificada por la IA.
Dilución de la Evidencia	En un Meta-análisis, la IA sugiere una referencia para un ECA crítico. Si esa referencia es apócrifa, todo el análisis se basa en un supuesto erróneo.	Se viola el principio de Integridad Científica, base de la credibilidad profesional.

Fuente: Elaboración propia con apoyo de inteligencia artificial

El Protocolo de verificación: Este es el protocolo que se beneficia de seguir para la verificación:

Verificación Binaria: Toda referencia generada por IA (Perplexity, Elicit, Scite, o un LLM general) debe someterse a una verificación "existe/no existe" utilizando el DOI o el título en Google Scholar o PubMed.

Verificación de Contenido: Si la referencia existe, el clínico debe abrir el abstract (y, idealmente, el texto completo) para verificar la congruencia. No basta con que el artículo exista; debe confirmar que la conclusión que la IA le atribuyó (dosis, outcome, población) sea la real.

Registro de Excepción: El experto debe crear el hábito de documentar en sus notas clínicas o de investigación: "Evidencia obtenida con IA, verificada manualmente con el DOI XXXXX". Esta trazabilidad es su defensa ética y metodológica.

9. DE LA BÚSQUEDA PUNTUAL A LA MEDICINA BASADA EN LA EVIDENCIA DINÁMICA

El capítulo, al centrarse en una pregunta PICO y la búsqueda en bases de datos, representa el modelo tradicional y estático de la MBE: el clínico realiza una búsqueda en un momento dado para responder una pregunta.

El Desafío Estratégico: La evidencia clínica de alto impacto (ensayos de fase III) se publica a un ritmo acelerado. Una revisión sistemática (RS) está metodológicamente obsoleta en un promedio de 5.5 años, pero las guías de práctica clínica pueden tardar 18 meses en incorporar esa nueva información. Un experto no puede permitirse esta brecha.

La inteligencia artificial no solo debe agilizar la búsqueda; debe transformar el flujo de trabajo completo de la Revisión Sistemática (RS), permitiendo la Revisión Sistemática Viva (Living SystematicReview - LSR).

a. Las Revisiones Sistemáticas Vivas (LSR): Una LSR es una revisión sistemática en la que los autores se comprometen a monitorear continuamente la literatura relevante y a incorporar nuevos hallazgos en cuanto estén disponibles (generalmente cada 3 a 12 meses). Su implicación Clínica garantiza que las recomendaciones y las estimaciones del efecto (el Meta-análisis) reflejen la evidencia más actual, reduciendo el retraso en la implementación de guías clínicas.

b. La IA como Facilitador del Flujo de Trabajo (Workflow): El verdadero costo de una RS no está en la búsqueda, sino en el filtrado y la extracción de datos. La IA interviene en las siguientes etapas críticas para hacer viable una LSR:

https://doi.org/10.62486/978-9915-9928-3-9.Ch06

Tabla 11 Rol de la inteligencia artificial en las distintas etapas del workflow de una revisión sistemática

Etapa del Workflow (RS)	Rol Crítico de la IA	Herramientas Típicas
Monitoreo y Detección	Alertas Programadas y Búsqueda Automática: La IA mantiene las ecuaciones de búsqueda activas en múltiples bases de datos y envía alertas sobre nuevos papers que coinciden con la estrategia avanzada.	Servicios de alertas de PubMed, Scopus, Google Scholar.
Screening (Filtrado)	Clasificación por Machine Learning: Después de eliminar duplicados, la IA lee los títulos y resúmenes y les asigna un puntaje de relevancia (ej. "probabilidad de ser un ECA relevante"). Esto permite al equipo humano priorizar y descartar rápidamente miles de artículos irrelevantes.	Rayyan, Covidence (con funciones IA): Estas plataformas están diseñadas para la gestión colaborativa del screening.
Extracción de Datos	Procesamiento de Lenguaje Natural (PNL): La IA identifica y extrae semi-automáticamente los elementos PICO	Elicit, Scite AI, OpenEvidence: Estas herramientas son fundamentales aquí, ya que estructuran la respuesta a la pregunta PICO y extraen datos específicos.
Evaluación del Sesgo	Detección de Patrones de Reporte: La IA puede ayudar a identificar inconsistencias o faltantes en el reporte (ej. falta de descripción de la secuencia de aleatorización), lo que es un indicio de un posible Riesgo de Sesgo (RoB).	Herramientas experimentales de Natural Language Processing (NLP) sobre textos completos.

Fuente: Elaboración propia con apoyo de inteligencia artificial

La IA convierte al clínico de un simple buscador de evidencia en un gestor dinámico de la información. El valor no es la velocidad de la respuesta inicial, sino la garantía de que la evidencia que utilizas hoy es la más reciente y rigurosa.

11. PREGUNTAS DE AUTOEVALUACIÓN

PREGUNTAS DE OPCIÓN MÚLTIPLE

1. ¿Cuál es el principal beneficio del uso de modelos de lenguaje grande (LLM) en la búsqueda de literatura científica?
a) Reemplazar las bases de datos biomédicas tradicionales
b) Automatizar la lectura crítica de los artículos
c) Acelerar la identificación y organización inicial de evidencia relevante
D) Garantizar que la evidencia encontrada sea de alta calidad

2. ¿Qué característica distingue a herramientas como Perplexity o Consensus de una búsqueda clásica en PubMed?
a) Utilizan exclusivamente términos MeSH
b) Proporcionan respuestas narrativas basadas en múltiples fuentes
c) Permiten búsquedas exhaustivas y reproducibles
d) Evalúan directamente la calidad metodológica de los estudios

3. ¿Cuál de las siguientes herramientas es especialmente útil para analizar cómo un artículo es citado en la literatura científica (apoyo, contraste o mención)?
a) Elicit
b) Perplexity
c) Scite AI
d) Research Rabbit

4. ¿En qué escenario es más apropiado utilizar herramientas como OpenEvidence o Consensus?
a) Elaboración de una revisión sistemática
b) Construcción de una estrategia MeSH exhaustiva
c) Confirmación rápida de si existe evidencia que respalde una hipótesis clínica
d) Análisis estadístico de los resultados de un estudio

5. Según el enfoque del capítulo, ¿cuál es la relación correcta entre inteligencia artificial y Medicina Basada en la Evidencia?
a) La IA sustituye al método científico tradicional
b) La IA elimina la necesidad de lectura crítica
c) La IA complementa el método científico, mejorando la eficiencia
d) La IA garantiza conclusiones clínicas correctas

https://doi.org/10.62486/978-9915-9928-3-9.Ch06

PREGUNTAS ABIERTAS

6. Explique por qué las herramientas de inteligencia artificial no deben utilizarse como único método de búsqueda bibliográfica en investigación clínica.

7. Describa dos situaciones clínicas en las que el uso de LLM puede ser especialmente útil y dos en las que no lo sería.

8. Compare brevemente el rol de Elicit y Scite AI dentro del proceso de búsqueda y análisis de literatura científica.

9. ¿Qué se entiende por "curaduría inicial de evidencia" y cuál es el papel del médico en esta etapa cuando se utilizan LLM?

10. Explique cómo puede integrarse de forma responsable la inteligencia artificial en la toma de decisiones clínicas basadas en evidencia.

RESPUESTAS

1:c) Los LLM permiten acelerar la identificación y organización inicial de evidencia relevante, pero no sustituyen la búsqueda clásica ni la lectura crítica.

2: b) Estas herramientas ofrecen respuestas narrativas basadas en múltiples fuentes, lo que las diferencia de la búsqueda estructurada clásica.

3: c) Scite AI se centra en analizar el contexto de las citaciones (apoyo, contraste o mención).

4: c) OpenEvidence y Consensus son útiles para confirmar rápidamente si existe respaldo científico para una hipótesis clínica.

5: c) La IA complementa el método científico, mejorando la eficiencia, pero no garantiza calidad ni reemplaza el juicio clínico.

6. Debe mencionarse que la IA puede omitir artículos, cometer errores bibliográficos o simplificar evidencia compleja, por lo que siempre se requiere verificación en bases de datos originales.

7. Ejemplos útiles: guardias clínicas, toma de decisiones preliminares, orientación inicial de un tema. Ejemplos no útiles: revisiones sistemáticas formales, elaboración de guías clínicas.

8. Elicit organiza estudios y facilita comparación; Scite AI evalúa cómo esos estudios son utilizados en la literatura.

9. Curaduría inicial implica seleccionar y resumir evidencia preliminar; el médico debe verificar fuentes, evaluar calidad y decidir aplicabilidad clínica.

10. Integración responsable implica usar IA como apoyo, verificar fuentes originales, realizar lectura crítica y contextualizar la evidencia al paciente real.

https://doi.org/10.62486/978-9915-9928-3-9.Ch06

REFERENCIAS

Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, Ashrafian H, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. Lancet Digit Health. 2020;2(10):e537–e548. Available at: https://doi.org/10.1016/S2589-7500(20)30218-1.
Bolaños F, Salatino A, Osborne F, Motta E. Artificial intelligence for literature reviews: Opportunities and challenges. arXiv [Preprint]. 2024 Aug 6. Availableat: https://arxiv.org/abs/2402.08565.
Marshall IJ, Wallace BC. Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. Syst Rev. 2019;8:163. Available at: https://doi.org/10.1186/s13643-019-1089-6.
Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv [Preprint]. 2023 Apr 12. Availableat: https://arxiv.org/abs/2303.13375.
Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y, et al. Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. NPJ Digit Med. 2023;6:86. Available at: https://doi.org/10.1038/s41746-023-00800-7.
Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). New York: Association for Computing Machinery; 2021. p. 610–623. Available at: https://doi.org/10.1145/3442188.3445922.
Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ. 2020;368:l6927. Available at: https://doi.org/10.1136/bmj.l6927.
van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CLH. ChatGPT: Five priorities for research. Nature. 2023;614(7947):224–226. Available at: https://doi.org/10.1038/d41586-023-00288-7.
Perplexity AI. About Perplexity AI [Internet]. 2024 [cited 2025]. Available at: https://www.perplexity.ai.
Elicit. Elicit: The AI research assistant [Internet]. 2024 [cited 2025]. Available at: https://elicit.org.
Scite AI. How Scite works [Internet]. 2024 [cited 2025]. Available at: https://scite.ai.
Consensus. Scientific consensus search engine [Internet]. 2024 [cited 2025]. Available at: https://consensus.app.
Semantic Scholar. About Semantic Scholar [Internet]. 2024 [cited 2025]. Available at: https://www.semanticscholar.org.
OpenEvidence. Evidence-based clinical AI search platform [Internet]. 2024 [cited 2025]. Availableat: https://openevidence.com.

https://doi.org/10.62486/978-9915-9928-3-9.Ch06

Chapter 7 / Capítulo 7

Critical reading of scientific articles on health: A practical guide (Spanish Edition)
ISBN: 978-9915-9928-3-9
DOI: 10.62486/978-9915-9928-3-9.Ch07
Pages: 70-82

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Research Methodology in the Health Field

Metodología de la Investigación en el Área de la Salud

Jhossmar Cristians Auza-Santivañez
Antonio Viruez-Soto
Nadia Sandra Orozco Vargas

1 Hospital del Gran Chaco "Fray Quebracho" de III nivel. Instituto Académico Científico Quispe-Cornejo. Tarija, Bolivia.
2 Hospital del Norte. Unidad de Cuidado Intensivo. El Alto, Bolivia
3 Pediatric Intensive Care of Instituto da Criança do Hospital das Clinicas da Faculdade de Medicina - Sao Paulo University, San Paulo, Brazil.

ABSTRACT

Introduction: Health research methodology comprises the set of scientific principles, procedures, and strategies designed to generate valid, reliable knowledge applicable to clinical practice. Its importance lies in providing the foundation for producing high-quality evidence that contributes to decision-making within the framework of Evidence-Based Medicine (EBM). Development: This chapter addresses the conceptual foundations of health research, highlighting its essential characteristics: systematicity, objectivity, reproducibility, ethics, and orientation toward scientific evidence. The scientific method is described as the central axis of the research process, encompassing the observation of the problem, the formulation of structured questions using the PICO model, the formulation of hypotheses, the selection of the methodological design, data collection and analysis, and the communication of results. The main research designs are also presented, including observational studies, experimental studies, and systematic reviews with meta-analysis. Furthermore, the concepts of variables, internal and external validity, as well as the main biases that can affect the methodological quality of a study, are analyzed. Finally, the importance of ethical principles, respect for participants, and compliance with international standards to guarantee scientific integrity is emphasized. Conclusion: The quality of research depends not only on the results obtained but also on the methodological rigor applied at each stage of the scientific process. Understanding research methodology allows for the critical interpretation of biomedical literature and the distinction between solid evidence and potentially biased conclusions. In an environment where scientific information is growing exponentially, developing critical thinking and methodological skills is an essential professional responsibility for transforming data into knowledge and knowledge into better decisions for people's health.

Keywords: Keywords. Research methodology; Scientific method; Study designs; Clinical trial; Health research

RESUMEN

Introducción: La metodología de la investigación en salud constituye el conjunto de principios, procedimientos y estrategias científicas destinados a generar conocimiento válido, confiable y aplicable a la práctica clínica. Su importancia radica en proporcionar las bases para producir evidencia de calidad que contribuya a la toma de decisiones en el marco de la Medicina Basada en la Evidencia (MBE). Desarrollo: El capítulo aborda los fundamentos conceptuales de la investigación en salud, destacando sus características esenciales: sistematicidad, objetividad, reproducibilidad, ética y orientación hacia la evidencia científica. Se describe el método científico como el eje central del proceso investigativo, que comprende la observación del problema, la formulación de preguntas estructuradas mediante el modelo PICO, el planteamiento de hipótesis, la selección del diseño metodológico, la recolección y análisis de datos y la comunicación de resultados. Asimismo, se presentan los principales diseños de investigación, incluyendo estudios observacionales, estudios experimentales y revisiones sistemáticas con metaanálisis. Se analizan además los conceptos de variables, validez interna y externa, así como los principales sesgos que pueden afectar la calidad metodológica de un estudio. Se enfatiza la importancia de los principios éticos, el respeto a los participantes y el cumplimiento de normas internacionales para garantizar la integridad científica. Conclusión: La calidad de una investigación no depende exclusivamente de los resultados obtenidos, sino del rigor metodológico aplicado en cada etapa del proceso científico. Comprender la metodología de la investigación permite interpretar críticamente la literatura biomédica y distinguir evidencia sólida de conclusiones potencialmente sesgadas. En un entorno donde la información científica crece de forma exponencial, desarrollar pensamiento crítico y competencias metodológicas constituye una responsabilidad profesional indispensable para transformar datos en conocimiento y conocimiento en mejores decisiones para la salud de las personas.

Palabras clave: Palabras clave. Metodología de la investigación; Método científico; Diseños de estudio; Ensayo clínico; Investigación en salud

1. CASO CLÍNICO

Una residente de la especialidad Medicina Interna, observa que los pacientes con diabetes mellitus tipo 2 atendidos en su establecimiento de salud, presentan un elevado número de complicaciones cardiovasculares. Durante una reunión académica surge la pregunta: ¿los programas nacionales en Bolivia estructurados de educación en salud y conducta terapéutica reducen el riesgo de hospitalización en comparación con la atención convencional?

La residente decide desarrollar un proyecto de investigación. Sin embargo, antes de iniciar la recolección de datos, necesita comprender cómo formular una pregunta científica, seleccionar un diseño metodológico adecuado, identificar variables, minimizar sesgos y garantizar la validez de los resultados.

Este escenario refleja una situación frecuente en el ámbito sanitario: transformar un problema en salud cotidiano en una investigación científica capaz de generar evidencia útil para la toma de decisiones.

2. INTRODUCCIÓN

La investigación científica constituye uno de los pilares fundamentales sobre los cuales se sustenta el progreso de las ciencias de la salud. Los avances en la prevención de enfermedades, el desarrollo de nuevas estrategias diagnósticas, la introducción de tratamientos innovadores y la optimización de los sistemas sanitarios han sido posibles gracias a la generación sistemática de conocimiento mediante métodos científicos rigurosos. En un contexto caracterizado por el crecimiento exponencial de la información biomédica, la investigación se ha convertido en una herramienta indispensable para transformar datos y observaciones clínicas en evidencia capaz de orientar la toma de decisiones en beneficio de los pacientes y de la sociedad en su conjunto (1,2).La práctica clínica moderna se desarrolla en un entorno cada vez más complejo, donde los profesionales de la salud enfrentan diariamente la necesidad de interpretar grandes volúmenes de información científica, evaluar críticamente la calidad de la evidencia disponible y aplicar los hallazgos más relevantes a contextos clínicos específicos. Esta realidad ha consolidado el paradigma de la Medicina Basada en la Evidencia (MBE), entendida como la integración equilibrada entre la mejor evidencia científica disponible, la experiencia clínica del profesional y los valores y preferencias del paciente (3,4). Dentro de este marco conceptual, la metodología de la investigación adquiere una relevancia estratégica, ya que proporciona los principios y procedimientos necesarios para producir evidencia válida, confiable y aplicable.

Más allá de ser un conjunto de técnicas estadísticas o procedimientos operativos, la metodología de la investigación representa una forma estructurada de pensamiento crítico orientada a la resolución de problemas. Su propósito fundamental es reducir la incertidumbre inherente al conocimiento empírico mediante la aplicación sistemática del método científico. Desde la identificación de un problema de salud hasta la interpretación y difusión de los resultados, cada etapa metodológica contribuye a garantizar que las conclusiones obtenidas reflejen de la manera más fiel posible la realidad estudiada (5). Históricamente, el desarrollo de la investigación en salud ha evolucionado desde observaciones clínicas aisladas hacia diseños metodológicos cada vez más sofisticados. Durante el siglo XX, la introducción de los ensayos clínicos aleatorizados, los estudios epidemiológicos modernos y las revisiones sistemáticas transformó la forma de generar evidencia científica. Actualmente, el auge de las tecnologías digitales, los registros electrónicos de salud, la inteligencia artificial, el análisis de grandes bases de datos (Big Data) y la medicina de precisión ha ampliado significativamente las posibilidades de investigación, permitiendo responder preguntas que hace apenas unas décadas resultaban inaccesibles (6,7).

Sin embargo, el acceso a tecnologías avanzadas y a grandes volúmenes de información no garantiza por sí mismo la producción de conocimiento científico de calidad. La evidencia demuestra que errores metodológicos durante la formulación de la pregunta de investigación, la selección del diseño, la medición de variables o el análisis estadístico pueden comprometer seriamente la validez de los resultados y conducir a conclusiones erróneas. De hecho, diversos autores han señalado que una proporción importante de la investigación biomédica presenta limitaciones metodológicas que afectan su reproducibilidad y utilidad clínica, fenómeno que ha dado lugar al denominado problema del desperdicio de la investigación ("researchwaste") (8,9).

En este sentido, comprender los fundamentos metodológicos de la investigación no constituye una competencia exclusiva de investigadores o académicos, sino una necesidad para todos los profesionales de la salud. Médicos, enfermeras, odontólogos, bioquímicos, farmacéuticos,

https://doi.org/10.62486/978-9915-9928-3-9.Ch07

fisioterapeutas y otros integrantes del equipo sanitario requieren habilidades para interpretar críticamente la literatura científica, identificar fortalezas y limitaciones metodológicas de los estudios y utilizar adecuadamente la evidencia en la práctica clínica cotidiana (10).

La investigación en salud también desempeña un papel esencial en la formulación de políticas sanitarias y en la gestión de los sistemas de salud. Las decisiones relacionadas con la asignación de recursos, la implementación de programas preventivos, la introducción de nuevas tecnologías sanitarias y la evaluación de intervenciones poblacionales deben sustentarse en evidencia científica sólida. En consecuencia, la calidad metodológica de las investigaciones trasciende el ámbito académico y tiene repercusiones directas sobre la salud de las poblaciones y la eficiencia de los sistemas sanitarios (11). Desde una perspectiva ética, la investigación científica implica una responsabilidad significativa hacia los participantes, las comunidades y la sociedad. La generación de conocimiento debe desarrollarse bajo principios de respeto, beneficencia, no maleficencia y justicia, garantizando que los beneficios potenciales superen los riesgos y que los resultados obtenidos contribuyan efectivamente al bienestar colectivo. Por ello, el rigor metodológico y la integridad científica son dimensiones inseparables dentro de la investigación en salud (12).

El presente capítulo tiene como objetivo proporcionar una visión integral de los fundamentos de la metodología de la investigación en el área de la salud. Se abordarán los principios del método científico, la formulación de preguntas de investigación, los principales diseños epidemiológicos y clínicos, la definición y medición de variables, los conceptos de validez y sesgo, así como las consideraciones éticas esenciales para la producción de evidencia científica de calidad. Comprender estos elementos permitirá al lector desarrollar una base sólida para interpretar críticamente la literatura biomédica y participar activamente en la generación de conocimiento orientado a mejorar la salud individual y colectiva.

3. CONCEPTO Y CARACTERÍSTICAS DE LA INVESTIGACIÓN EN SALUD

La investigación en salud puede definirse como el conjunto de procedimientos científicos destinados a generar conocimientos nuevos o validar conocimientos existentes relacionados con los procesos de salud y enfermedad.

Entre sus principales características destacan:

- Sistemática: sigue una secuencia lógica y organizada.
- Objetiva: busca minimizar la influencia de opiniones personales.
- Reproducible: permite que otros investigadores obtengan resultados similares bajo condiciones equivalentes.
- Ética: protege los derechos, la seguridad y el bienestar de los participantes.
- Basada en evidencia: utiliza métodos científicos para obtener conclusiones válidas.

La investigación en salud constituye uno de los pilares fundamentales para el avance de la medicina moderna y para la formulación de políticas sanitarias basadas en evidencia científica.

4. EL MÉTODO CIENTÍFICO

El método científico es la estrategia general utilizada para generar conocimiento confiable. En ciencias de la salud suele desarrollarse mediante las siguientes etapas:

4.1 OBSERVACIÓN DEL PROBLEMA

Corresponde a la identificación de una situación clínica, epidemiológica o sanitaria que requiere explicación.**Ejemplo:** aumento de infecciones intrahospitalarias en una unidad de cuidados intensivos.

4.2 FORMULACIÓN DE LA PREGUNTA DE INVESTIGACIÓN

La pregunta debe ser clara, específica y relevante.Una estrategia ampliamente utilizada es el formato PICO:

- **P (Paciente o población)**
- I (Intervención o exposición)

https://doi.org/10.62486/978-9915-9928-3-9.Ch07

- C (Comparación)
- **(Outcome o desenlace)**

Ejemplo:

En pacientes diabéticos tipo 2 (P), ¿un programa de educación terapéutica (I), comparado con atención convencional (C), reduce las hospitalizaciones cardiovasculares (O)?

4.3 FORMULACIÓN DE HIPÓTESIS

La hipótesis representa una explicación tentativa que será sometida a prueba.

Hipótesis de investigación: los programas educativos reducen las hospitalizaciones cardiovasculares.

Hipótesis nula: no existen diferencias entre ambos grupos.

4.4 DISEÑO DEL ESTUDIO

Se selecciona el diseño más apropiado para responder la pregunta planteada.

4.5 RECOLECCIÓN DE DATOS

Se aplican instrumentos válidos y confiables para obtener información.

4.6 ANÁLISIS E INTERPRETACIÓN

Los datos son procesados mediante métodos estadísticos adecuados.

4.7 COMUNICACIÓN DE RESULTADOS

Los hallazgos se difunden mediante artículos científicos, tesis,confencias, informes técnicos o congresos.

5. DISEÑOS DE INVESTIGACIÓN EN SALUD

La selección del diseño metodológico constituye una de las decisiones más importantes del investigador.

Figura 1

Fuente: Elaboración propia

https://doi.org/10.62486/978-9915-9928-3-9.Ch07

5.1 ESTUDIOS OBSERVACIONALES.

Los estudios observacionales son diseños epidemiológicos en los cuales el investigador se limita a observar y registrar la ocurrencia natural de las exposiciones y los desenlaces, sin asignar intervenciones ni modificar las condiciones bajo las cuales ocurren los fenómenos de interés. Constituyen la base de gran parte de la investigación clínica y de salud pública, especialmente cuando los ensayos clínicos no son factibles por razones éticas, logísticas o económicas. Su principal utilidad radica en describir la frecuencia de las enfermedades, identificar factores asociados, generar hipótesis etiológicas y evaluar asociaciones entre exposiciones y desenlaces en condiciones reales de práctica clínica.

Desde el punto de vista metodológico, los estudios observacionales pueden clasificarse en descriptivos y analíticos. Entre los diseños analíticos más utilizados se encuentran los estudios transversales, los estudios de casos y controles y los estudios de cohortes.

ESTUDIOS TRANSVERSALES

Los estudios transversales (cross-sectionalstudies) evalúan simultáneamente la exposición y el desenlace en una población definida durante un único momento o período específico. Debido a que la información se obtiene de forma concurrente, permiten estimar la prevalencia de enfermedades, factores de riesgo o condiciones de salud, así como explorar asociaciones entre variables.Estos estudios son especialmente útiles para conocer la magnitud de un problema sanitario, planificar programas de salud pública, identificar grupos de riesgo y generar hipótesis para investigaciones posteriores. Sin embargo, una de sus principales limitaciones es la imposibilidad de establecer la secuencia temporal entre exposición y desenlace, lo que restringe la capacidad para realizar inferencias causales.

ESTUDIOS DE CASOS Y CONTROLES

Los estudios de casos y controles (case-control studies) comparan individuos que presentan un determinado desenlace o enfermedad (casos) con individuos libres de dicha condición (controles), analizando retrospectivamente la frecuencia de exposición a uno o varios factores de riesgo.

Este diseño resulta particularmente eficiente para investigar enfermedades poco frecuentes o con largos períodos de latencia, ya que requiere un menor número de participantes y un tiempo de ejecución relativamente corto. La principal medida de asociación obtenida es la razón de momios (odds ratio, OR)**,** la cual constituye una estimación válida del riesgo relativo cuando la enfermedad es infrecuente.Entre sus limitaciones destacan la susceptibilidad al sesgo de selección, el sesgo de recuerdo (recallbias) y la dificultad para establecer con precisión la temporalidad entre la exposición y el desarrollo del desenlace.

Estudios de cohortes

Los estudios de cohortes (cohortstudies) consisten en el seguimiento longitudinal de grupos de individuos clasificados según la presencia o ausencia de una exposición determinada, con el objetivo de comparar la ocurrencia futura del desenlace de interés.Las cohortes pueden ser prospectivas, cuando el seguimiento inicia antes de que ocurra el desenlace, o retrospectivas, cuando utilizan registros previamente existentes para reconstruir la secuencia temporal entre exposición y evento.

Este diseño permite calcular directamente la incidencia, el riesgo absoluto, el riesgo relativo (RR)**,** las tasas de incidencia y otras medidas epidemiológicas de asociación. Debido a que la exposición precede al desenlace, los estudios de cohortes ofrecen una mayor solidez para establecer relaciones temporales y fortalecer la inferencia causal, aunque continúan siendo susceptibles a factores de confusión propios de los estudios no experimentales.

REPORTE DE ESTUDIOS OBSERVACIONALES

La calidad metodológica y la transparencia en la comunicación de los estudios observacionales son fundamentales para su adecuada interpretación y utilización en la Medicina Basada en la Evidencia. Con este propósito, la iniciativa StrengtheningtheReportingofObservationalStudies in Epidemiology (STROBE) desarrolló una lista de verificación de 22 ítems que orienta el reporte

https://doi.org/10.62486/978-9915-9928-3-9.Ch07

completo y transparente de estudios de cohortes, casos y controles y estudios transversales. Las recomendaciones STROBE no constituyen una herramienta para evaluar el riesgo de sesgo ni la calidad metodológica del estudio; su finalidad es mejorar la calidad del reporte científico, facilitar la reproducibilidad de la investigación y permitir que lectores, revisores y clínicos valoren adecuadamente la validez y aplicabilidad de los resultados.

5.2 ESTUDIOS EXPERIMENTALES.

Los estudios experimentales son diseños de investigación en los cuales el investigador interviene activamente asignando una exposición o intervención a los participantes con el propósito de evaluar su efecto sobre uno o más desenlaces previamente definidos. A diferencia de los estudios observacionales, la exposición no ocurre de forma natural, sino que es controlada por el investigador mediante un protocolo previamente establecido. Esta característica permite reducir la influencia de factores de confusión y fortalecer la inferencia causal entre la intervención y los resultados observados.

Dentro de los estudios experimentales, el ensayo clínico aleatorizado(RandomizedControlled Trial, RCT) constituye el diseño metodológico de mayor rigor para evaluar la eficacia y la seguridad de intervenciones terapéuticas, preventivas o diagnósticas. Debido a su elevada validez interna, representa el estándar de referencia para establecer relaciones de causalidad y sustenta gran parte de las recomendaciones incluidas en las guías de práctica clínica basadas en evidencia.

ENSAYO CLÍNICO ALEATORIZADO

Un ensayo clínico aleatorizado consiste en la asignación de los participantes, mediante un proceso de aleatorización, a dos o más grupos de comparación, generalmente un grupo que recibe la intervención de interés y otro que recibe el tratamiento estándar, placebo o ninguna intervención, según corresponda al objetivo del estudio y a las consideraciones éticas. La asignación aleatoria busca equilibrar las características basales entre los grupos, minimizando el efecto de variables de confusión conocidas y desconocidas. Para preservar la validez metodológica del estudio, los ensayos clínicos suelen incorporar procedimientos adicionales como la ocultación de la secuencia de asignación (allocationconcealment)y el enmascaramiento o cegamiento (blinding) de participantes, investigadores y evaluadores de resultados. Estas estrategias disminuyen el riesgo de sesgo de selección, desempeño, detección y medición, incrementando la credibilidad de los hallazgos.

Las principales fortalezas metodológicas de los ensayos clínicos aleatorizados incluyen:

- Mayor control de los sesgos, mediante la aleatorización, la ocultación de la asignación y el cegamiento de los participantes y evaluadores.
- Elevada validez interna, al reducir la influencia de factores de confusión y garantizar que las diferencias observadas entre los grupos sean atribuibles principalmente a la intervención estudiada.
- Capacidad para establecer relaciones causales, al asegurar la secuencia temporal entre la intervención y el desenlace, constituyendo el diseño con mayor nivel de evidencia para evaluar eficacia y seguridad terapéutica.
- Estimación directa del efecto del tratamiento, mediante medidas como el riesgo relativo (RR), la reducción absoluta del riesgo (RAR), la reducción relativa del riesgo (RRR), el número necesario a tratar (NNT) y el número necesario para dañar (NNH), fundamentales para la toma de decisiones clínicas basadas en evidencia.

No obstante, los ensayos clínicos presentan algunas limitaciones. Su desarrollo suele requerir elevados recursos económicos, infraestructura especializada y tiempos prolongados de seguimiento. Además, los estrictos criterios de inclusión y exclusión pueden limitar la validez externa o capacidad de generalizar los resultados a poblaciones diferentes de aquellas incluidas en el estudio. Asimismo, existen situaciones en las que, por razones éticas, no es posible realizar un ensayo clínico, especialmente cuando la intervención pudiera ocasionar daño o privar a los participantes de un tratamiento eficaz.

REPORTE DE LOS ENSAYOS CLÍNICOS

La calidad metodológica de un ensayo clínico no depende únicamente de su diseño, sino

https://doi.org/10.62486/978-9915-9928-3-9.Ch07

también de la transparencia con la que se comunican sus métodos y resultados. Con este propósito, la iniciativa ConsolidatedStandardsofReportingTrials (CONSORT) desarrolló una guía internacional para el reporte de ensayos clínicos aleatorizados, integrada por una lista de verificación de 25 ítems y un diagrama de flujo que documenta el proceso de selección, asignación, seguimiento y análisis de los participantes. La declaración CONSORT favorece la reproducibilidad, facilita la evaluación crítica de la literatura científica y mejora la calidad de las publicaciones biomédicas. Aunque no constituye una herramienta para valorar el riesgo de sesgo, su utilización permite identificar si un ensayo clínico reporta adecuadamente aspectos esenciales como la generación de la secuencia aleatoria, la ocultación de la asignación, el cegamiento, el cálculo del tamaño muestral, el análisis por intención de tratar y el registro previo del protocolo, elementos indispensables para interpretar correctamente la validez de los resultados.

5.3 REVISIONES SISTEMÁTICAS Y METAANÁLISIS

Las revisiones sistemáticas constituyen el diseño de investigación secundaria con mayor nivel de evidencia para responder preguntas clínicas cuando integran estudios primarios de alta calidad metodológica. Su objetivo es identificar, seleccionar, evaluar críticamente y sintetizar de manera explícita, transparente y reproducible toda la evidencia disponible relacionada con una pregunta de investigación previamente formulada. A diferencia de las revisiones narrativas, las revisiones sistemáticas siguen un protocolo metodológico predefinido que minimiza el riesgo de sesgo durante todas las etapas del proceso de investigación, desde la búsqueda bibliográfica hasta la interpretación de los resultados.

El proceso metodológico comprende la formulación de una pregunta clínicaestructurada generalmente mediante el formato PICO, la elaboración y registro de un protocolo, la búsqueda exhaustiva en múltiples bases de datos, la selección de estudios según criterios de elegibilidad previamente establecidos, la evaluación del riesgo de sesgo, la extracción estandarizada de los datos y la síntesis cualitativa o cuantitativa de la evidencia. Este enfoque permite obtener conclusiones más robustas y reproducibles que las derivadas de estudios individuales, constituyendo la base para la elaboración de guías de práctica clínica y recomendaciones fundamentadas en evidencia (2,3).

Cuando los estudios incluidos presentan suficiente homogeneidad clínica, metodológica y estadística, la revisión sistemática puede incorporar un metaanálisis**,** entendido como una técnica estadística destinada a combinar cuantitativamente los resultados de investigaciones independientes para obtener una estimación global del efecto. El metaanálisis incrementa la precisión de las estimaciones, mejora la potencia estadística y facilita la exploración de la heterogeneidad entre los estudios mediante análisis específicos, como el estadístico I^2, las pruebas de heterogeneidad y los análisis de sensibilidad.

Es importante destacar que no todas las revisiones sistemáticas incluyen un metaanálisis, ya que la combinación estadística únicamente es apropiada cuando existe una adecuada similitud entre los estudios en cuanto a población, intervenciones, comparadores, desenlaces y metodología. Del mismo modo, la presencia de un metaanálisis no garantiza automáticamente una evidencia de alta calidad, puesto que la validez de sus conclusiones depende de la calidad metodológica de los estudios primarios incluidos, del riesgo de sesgo, de la consistencia de los resultados y de la certeza global de la evidencia, aspectos que actualmente se valoran mediante la metodología GRADE (GradingofRecommendationsAssessment, Development and Evaluation).

En la jerarquía tradicional de la Medicina Basada en la Evidencia, las revisiones sistemáticas con metaanálisis de ensayos clínicos aleatorizados suelen ocupar el nivel más alto de evidencia para preguntas relacionadas con intervenciones terapéuticas. No obstante, la fortaleza de sus conclusiones siempre debe interpretarse considerando la calidad metodológica de los estudios incluidos, la heterogeneidad observada y la aplicabilidad clínica de los resultados a la población de interés.

REPORTE DE REVISIONES SISTEMÁTICAS

La elaboración y publicación de revisiones sistemáticas debe seguir las recomendaciones de la guía PreferredReportingItemsforSystematicReviews and Meta-Analyses(PRISMA 2020)**,** considerada actualmente el estándar internacional para el reporte transparente de este tipo de investigaciones. La declaración PRISMA incluye una lista de verificación de 27 ítems y un diagrama

https://doi.org/10.62486/978-9915-9928-3-9.Ch07

de flujo que documenta de forma detallada el proceso de identificación, selección, elegibilidad e inclusión de los estudios, favoreciendo la reproducibilidad, la transparencia y la evaluación crítica de la evidencia científica.

Además del cumplimiento de PRISMA, las revisiones sistemáticas de alta calidad deben registrar previamente su protocolo en plataformas como PROSPERO**,** describir explícitamente la estrategia de búsqueda, evaluar el riesgo de sesgo mediante herramientas validadas (como RoB 2 **o** ROBINS-I, según el diseño de los estudios incluidos) y valorar la certeza de la evidencia utilizando el sistema GRADE. Estas prácticas fortalecen la credibilidad de los resultados y reducen el riesgo de sesgos metodológicos y de reporte.

6. VARIABLES Y MEDICIÓN

Las variables representan las características que serán observadas o medidas.

Según su naturaleza

Cualitativas

- Sexo
- Estado civil
- Presencia de enfermedad

Cuantitativas

- Edad
- Presión arterial
- Índice de masa corporal

Según su función

Variable independiente

- Factor de exposición o intervención.

Variable dependiente

- Resultado o desenlace de interés.

La correcta definición operacional de las variables garantiza la reproducibilidad y validez de la investigación.

7. VALIDEZ, SESGOS Y CALIDAD METODOLÓGICA.

La validez hace referencia al grado en que los resultados reflejan la realidad.

Validez interna. Evalúa si los resultados obtenidos son atribuibles a la intervención o exposición estudiada.

Validez externa. Determina la posibilidad de generalizar los resultados a otras poblaciones.

PRINCIPALES SESGOS

Sesgo de selección. Ocurre cuando los participantes no representan adecuadamente a la población objetivo.

Sesgo de información. Se produce por errores en la medición o registro de datos.

Sesgo de confusión. Aparece cuando una tercera variable influye simultáneamente sobre exposición y desenlace.La identificación y control de estos sesgos constituye una de las funciones centrales del investigador.

8. CONSIDERACIONES ÉTICAS

La investigación en salud constituye una actividad científica orientada a generar conocimiento para mejorar la atención de los pacientes y fortalecer los sistemas sanitarios. Sin embargo, el avance del conocimiento nunca debe prevalecer sobre la dignidad, los derechos, la seguridad y el bienestar de las personas participantes. Por ello, toda investigación que involucre seres humanos, muestras biológicas identificables o información clínica debe desarrollarse bajo un marco ético riguroso que garantice el respeto de los derechos fundamentales y el cumplimiento de las normas nacionales e internacionales de investigación biomédica.

Los principios bioéticos que sustentan la investigación clínica fueron sistematizados inicialmente en el Informe Belmont y posteriormente incorporados en múltiples normativas

https://doi.org/10.62486/978-9915-9928-3-9.Ch07

internacionales, constituyendo actualmente la base de la ética en investigación. Estos principios son:

8.1 AUTONOMÍA

El principio de autonomía reconoce el derecho de toda persona a decidir libremente si desea participar en una investigación, sin coerción, manipulación ni influencia indebida. Su aplicación práctica se materializa mediante el proceso de consentimiento informado, el cual debe garantizar que el participante reciba información suficiente, comprensible y veraz sobre los objetivos del estudio, los procedimientos, los riesgos previsibles, los beneficios potenciales, las alternativas existentes, la confidencialidad de los datos y la posibilidad de retirarse del estudio en cualquier momento sin que ello afecte la calidad de su atención médica.

En poblaciones vulnerables como menores de edad, personas con discapacidad cognitiva, adultos mayores con deterioro neurocognitivo o pacientes en situación de emergencia deben implementarse medidas adicionales de protección que aseguren el respeto de sus derechos y la representación legal cuando corresponda.

8.2 BENEFICENCIA

El principio de beneficencia establece la obligación ética del investigador de maximizar los beneficios potenciales derivados de la investigación, tanto para los participantes como para la sociedad. Antes del inicio del estudio debe realizarse una evaluación rigurosa de la relación riesgo-beneficio, procurando que el conocimiento generado tenga relevancia científica, utilidad clínica y potencial impacto en la salud pública.

La beneficencia también implica diseñar investigaciones metodológicamente sólidas. Un estudio con deficiencias metodológicas carece de justificación ética, ya que expone innecesariamente a los participantes a riesgos sin posibilidad real de generar conocimiento válido o aplicable.

8.3 NO MALEFICENCIA

El principio de no maleficencia exige minimizar todos los riesgos físicos, psicológicos, sociales, económicos y legales asociados con la investigación. Los investigadores deben implementar procedimientos destinados a prevenir eventos adversos, establecer mecanismos de vigilancia durante el estudio y garantizar una atención adecuada en caso de presentarse daños relacionados con la investigación.

De esta manera, este principio comprende la protección de la privacidad, la confidencialidad de la información clínica y el manejo seguro de los datos personales conforme a las normas vigentes sobre protección de datos y seguridad de la información.

8.4 JUSTICIA

La justicia exige una distribución equitativa de los beneficios y las cargas derivadas de la investigación. La selección de los participantes debe basarse exclusivamente en criterios científicos previamente establecidos y nunca en razones de conveniencia, vulnerabilidad económica, discriminación o facilidad de acceso.De igual manera, los beneficios derivados del conocimiento generado deben contribuir al bienestar colectivo y favorecer el acceso equitativo a intervenciones eficaces, especialmente en poblaciones históricamente subrepresentadas o con mayores necesidades sanitarias.

8.5 EVALUACIÓN POR LOS COMITÉS DE ÉTICA EN INVESTIGACIÓN

Toda investigación que involucre seres humanos debe ser evaluada y aprobada previamente por un Comité de Ética en Investigación (CEI) independiente, competente y legalmente constituido antes del inicio del reclutamiento de participantes. Estos comités analizan la pertinencia científica del protocolo, la relación riesgo-beneficio, la protección de los derechos de los participantes, la

https://doi.org/10.62486/978-9915-9928-3-9.Ch07

idoneidad del consentimiento informado, la confidencialidad de la información y el cumplimiento de las normas éticas nacionales e internacionales.

La aprobación ética constituye un requisito indispensable para la ejecución del estudio y para la posterior publicación de los resultados en revistas científicas indexadas, las cuales exigen la declaración explícita de la aprobación por un CEI y el cumplimiento de las normas internacionales de investigación clínica.

8.6 DECLARACIÓN DE HELSINKI Y ESTÁNDARES INTERNACIONALES

La Declaración de Helsinki**,** elaborada por la Asociación Médica Mundial y actualizada periódicamente, representa el principal referente internacional para la investigación médica en seres humanos. Este documento establece principios fundamentales relacionados con la protección de los participantes, la evaluación independiente de los protocolos, el consentimiento informado, la privacidad, la transparencia en el registro y publicación de los estudios, así como la obligación de difundir los resultados, independientemente de que estos sean positivos, negativos o inconclusos.

Complementariamente, las Buenas Prácticas Clínicas (Good ClinicalPractice, GCP) y las directrices del Council for International Organizationsof Medical Sciences (CIOMS)proporcionan recomendaciones específicas para garantizar la integridad científica, la calidad metodológica y la protección de los participantes durante todo el proceso investigativo. En el contexto de la Medicina Basada en la Evidencia, el cumplimiento de los principios éticos no solo constituye una obligación moral y legal, sino también un requisito indispensable para valorar la credibilidad, validez y aplicabilidad de la evidencia científica. Un estudio metodológicamente robusto pero desarrollado sin apego a los principios éticos fundamentales pierde legitimidad científica y compromete la confianza de la comunidad académica y de la sociedad en los resultados obtenidos.

9. PREGUNTAS DE AUTOEVALUACIÓN

PREGUNTAS DE OPCIÓN MÚLTIPLE

1. ¿Cuál es el objetivo principal de la metodología de la investigación en salud?

a) Publicar artículos científicos b) Generar conocimiento válido y confiable c) Realizar análisis estadísticos complejos d) Elaborar revisiones narrativas

2. ¿Cuál es el diseño más apropiado para evaluar una intervención terapéutica?

a) Estudio transversal b) Casos y controles c) Ensayo clínico aleatorizado d) Serie de casos

3. ¿Cuál es el propósito principal del formato PICO?

a) Analizar datos estadísticos b) Formular preguntas de investigación estructuradas c) Redactar artículos científicos d) Elaborar referencias bibliográficas

4. ¿Cuál de los siguientes corresponde a un sesgo de selección?

a) Error de digitación de datos b) Participantes no representativos de la población objetivo c) Error estadístico aleatorio d) Confusión residual

5. ¿Qué guía se utiliza para reportar revisiones sistemáticas?

a) CONSORT b) STROBE c) PRISMA d) STARD

RESPUESTAS

- b
- c
- b
- b
- c

https://doi.org/10.62486/978-9915-9928-3-9.Ch07

REFERENCIAS

Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB. Designing Clinical Research. 5th ed. Philadelphia: Wolters Kluwer; 2022. Disponible en: https://apn.lwwhealthlibrary.com/book.aspx?bookid=3483
Gordis L. Epidemiology. 6th ed. Philadelphia: Elsevier; 2019. Disponible en: https://www.elsevier.com/books/epidemiology/gordis/978-0-323-55229-5
Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. Evidence-Based Medicine: How to Practice and Teach EBM. 2nd ed. Edinburgh: Churchill Livingstone; 2000. Disponible en: https://journals.lww.com/plasreconsurg/fulltext/2010/07000/how_to_practice_evidence_based_medicine.37.aspx
Guyatt G, Rennie D, Meade MO, Cook DJ. Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice. 3rd ed. New York: McGraw-Hill Education; 2015. Disponible en: https://jamaevidence.mhmedical.com/book.aspx?bookID=847
Hernández Sampieri R, Mendoza C. Metodología de la investigación: las rutas cuantitativa, cualitativa y mixta. Ciudad de México: McGraw-Hill Education; 2018.
Disponible en: https://artes.upn.edu.co/wp-content/uploads/2024/11/METODOLOGIA-DE-LA-INVESTIGACION-Sampieri-Mendoza-2018.pdf
Frieden TR. Evidence for Health Decision Making—Beyond Randomized, Controlled Trials. N Engl J Med. 2017;377(5):465-475. https://www.nejm.org/doi/full/10.1056/NEJMra1614394
Auza-SantiváñezJhossmar Cristians, Santiváñez-Cabezas María V., Dorta-Contreras Alberto Juan. Transferencia de conocimientos, clave para el desarrollo de la ciencia en Bolivia. Rev CubanaInvestBioméd [Internet]. 2021 Mar [citado 2026Jun 07] ;40(1): . Disponible en: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0864-03002021000100023
Ioannidis JPA. Why Most Published Research Findings Are False. PLoS Med. 2005;2(8):e124. https://doi.org/10.1371/journal.pmed.0020124
Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. Lancet. 2009;374(9683):86-89. Disponible en: https://doi.org/10.1016/S0140-6736(09)60329-9
Greenhalgh T. How to Read a Paper: The Basics of Evidence-Based Medicine and Healthcare. 7th ed. Hoboken: Wiley-Blackwell; 2024. Disponible en: https://www.wiley.com/en-us/shop/general-introductory-medical-science/how-to-read-a-paper-the-basics-of-evidence-based-healthcare-7th-edition-p-9781394206926
Brownson RC, Baker EA, Deshpande AD, Gillespie KN. Evidence-Based Public Health. 4th ed. New York: Oxford University Press; 2024. Disponible en: https://repository.unar.ac.id/jspui/bitstream/123456789/5288/1/Evidence-based%20Public%20Health%20by%20Ross%20C.%20Brownson%2C%20Elizabeth%20A.%20Baker%2C%20Anjali%20D.%20Deshpande%2C%20Kathleen%20N.%20Gillespie%20%28z-lib.org%29.pdf
World Medical Association. World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. https://www.wma.net/policies-post/wma-declaration-of-helsinki/
von Elm Erik, Altman Douglas G., Egger Matthias, Pocock Stuart J., Gøtzsche Peter C., Vandenbroucke Jan P..Declaración de la iniciativa STROBE (StrengtheningtheReportingofObservationalStudies in Epidemiology): directrices para la comunicación de estudios observacionales. Rev. Esp. Salud Publica [Internet]. 2008 Jun [citado 2026Jun 07] ;82(3): 251-259. Disponible en: http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1135-57272008000300002&lng=es.
Grech V, Eldawlatly AA. STROBE, CONSORT, PRISMA, MOOSE, STARD, SPIRIT, and other guidelines - Overview and application. Saudi J Anaesth. 2024 Jan-Mar;18(1):137-141. https://doi.org/10.4103/sja.sja_545_23
International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). ICH Harmonised Guideline E6(R3): Guideline for Good Clinical Practice [Internet]. Geneva: ICH; 2025 [citado 2026 Jun 28]. Disponible en: https://www.ich.org/page/efficacy-guidelines
World Medical Association. World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human participants. JAMA. 2024;332(4):281-286. doi:10.1001/jama.2024.12508
National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research [Internet]. Washington (DC): U.S. Department of Health, Education, and Welfare; 1979 [citado 2026 Jun 28]. Disponible en: https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html
Beauchamp TL, Childress JF. Principles of Biomedical Ethics. 8th ed. New York: Oxford University Press; 2019. Disponible en: https://global.oup.com/academic/product/principles-of-biomedical-ethics-9780190640873
Emanuel EJ, Wendler D, Grady C. What makes clinical research ethical? JAMA. 2000;283(20):2701-2711. doi:10.1001/jama.283.20.2701
International Committee of Medical Journal Editors (ICMJE). Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals [Internet]. Updated 2024 [citado 2026 Jun 28]. Disponible en: https://www.icmje.org/icmje-recommendations.pdf
Council for International Organizations of Medical Sciences (CIOMS). International Ethical Guidelines for Health-related Research Involving Humans [Internet]. 4th ed. Geneva: CIOMS; 2021 [citado 2026 Jun 28]. Disponible en: https://cioms.ch/publications/product/international-ethical-guidelines-for-health-related-research-involving-humans/
World Health Organization. Standards and Operational Guidance for Ethics Review of Health-Related Research with Human Participants [Internet]. Geneva: World Health Organization; 2011 [citado 2026 Jun 28]. Disponible en: https://iris.who.int/handle/10665/44783
Fletcher RH, Fletcher GS. Clinical Epidemiology: The Essentials. 6th ed. Philadelphia: Wolters Kluwer; 2021. Disponible en: https://shop.lww.com/Clinical-Epidemiology/p/9781975140984

Rothman KJ, Lash TL, Greenland S. Modern Epidemiology. 4th ed. Philadelphia: Wolters Kluwer; 2021. Disponible en: https://shop.lww.com/Modern-Epidemiology/p/9781975166311
Grimes DA, Schulz KF. Descriptive studies: what they can and cannot do. Lancet. 2002;359(9301):145-149. doi:10.1016/S0140-6736(02)07373-7
Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, for the STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for reporting observational studies. PLoS Med. 2007;4(10):e296. doi:10.1371/journal.pmed.0040296
Friedman LM, Furberg CD, DeMets DL, Reboussin DM, Granger CB. Fundamentals of Clinical Trials. 5th ed. Cham: Springer; 2015. doi:10.1007/978-3-319-18539-2
Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. Cochrane Handbook for Systematic Reviews of Interventions. Version 6.5 [Internet]. Cochrane; 2024. Disponible en: https://training.cochrane.org/handbook
Guyatt G, Rennie D, Meade MO, Cook DJ, editors. Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice. 3rd ed. New York: McGraw-Hill Education; 2015. Disponible en: https://jamaevidence.mhmedical.com
Schulz KF, Altman DG, Moher D, for the CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c332. doi:10.1136/bmj.c332
Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c869. doi:10.1136/bmj.c869
Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71
Greenhalgh T. How to Read a Paper: The Basics of Evidence-Based Medicine and Healthcare. 6th ed. Hoboken: Wiley-Blackwell; 2019. Disponible en: https://www.wiley.com/en-us/How+to+Read+a+Paper%3A+The+Basics+of+Evidence+Based+Medicine+and+Healthcare%2C+6th+Edition-p-9781119484745
Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to Meta-Analysis. Chichester: John Wiley & Sons; 2009. doi:10.1002/9780470743386
Schünemann HJ, Brożek J, Guyatt G, Oxman AD, editors. GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations [Internet]. Updated 2024. Disponible en: https://gdt.gradepro.org/app/handbook/handbook.html.

Chapter 8 / Capítulo 8

Critical reading of scientific articles on health: A practical guide (Spanish Edition)
ISBN: 978-9915-9928-3-9
DOI: 10.62486/978-9915-9928-3-9.Ch08
Pages: 83-97

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Practical Guide for Searching Scientific Articles

Guía práctica para búsqueda de artículos científicos

Timothé Schenker
Jhossmar Cristians Auza-Santivañez
Ariel Sosa-Remón
Carmen Julia Salvatierra-Rocha

1 Cabinet Similia Medica à Genève. Suisse.
2 Hospital del Gran Chaco "Fray Quebracho" de III nivel. Instituto Académico Científico Quispe-Cornejo. Tarija, Bolivia.
3 Instituto Nacional de Oncología y Radiobiología. La Habana, Cuba.
4 Clínica Foainini. Santa Cruz, Bolivia.

ABSTRACT

Introduction: The search for scientific evidence is an essential competency for modern clinical practice and for the application of Evidence-Based Medicine (EBM). This chapter uses the case of a patient with chronic obstructive pulmonary disease (COPD) who inquires about the effectiveness of home-based pulmonary rehabilitation or telehealth versus in-person rehabilitation to demonstrate how to transform a clinical need into a systematic literature search strategy. Development: The chapter describes, in a practical way, the process of searching for scientific literature in the main biomedical databases: Virtual Health Library (VHL), PubMed, SciELO, and the Cochrane Library. Initially, it emphasizes the identification of controlled terms using the DeCS and MeSH thesauri, which allow for the standardization of concepts and increase the precision and sensitivity of searches. Subsequently, it explains the construction of strategies using Boolean operators (AND, OR) and the use of filters related to language, publication date, population, and methodological design. The specific characteristics of each platform are presented. The VHL facilitates access to regional and Spanish-language literature; PubMed expands the search to include high-impact international evidence; SciELO provides open-access Latin American studies, and the Cochrane Library offers systematic reviews and syntheses of evidence of the highest methodological quality. Conclusions: Effective literature searching is a methodological process that goes beyond the simple use of keywords. No single database is sufficient on its own; therefore, combined searches maximize the retrieval of relevant information. The ability to locate, evaluate, and critically use scientific evidence is an indispensable skill for contemporary healthcare professionals, as it strengthens clinical decision-making, promotes patient-centered care, and contributes to the development of a more rigorous, transparent, and evidence-based medical practice.

Keywords: Keywords. Evidence-Based Medicine; Literature search; DeCS and MeSH; Biomedical databases; PICO strategy

RESUMEN

Introducción: La búsqueda de evidencia científica constituye una competencia esencial para la práctica clínica moderna y para la aplicación de la Medicina Basada en la Evidencia (MBE). Este capítulo utiliza el caso de un paciente con enfermedad pulmonar obstructiva crónica (EPOC) que consulta sobre la efectividad de la rehabilitación pulmonar domiciliaria o mediante TELESALUD frente a la rehabilitación presencial, para demostrar cómo transformar una necesidad clínica en una estrategia sistemática de búsqueda bibliográfica. Desarrollo: El capítulo describe de forma práctica el proceso de búsqueda de literatura científica en las principales bases de datos biomédicas: Biblioteca Virtual en Salud (BVS), PubMed, SciELO y Cochrane Library. Inicialmente, enfatiza la identificación de términos controlados mediante los tesauros DeCS y MeSH, los cuales permiten estandarizar conceptos y aumentar la precisión y sensibilidad de las búsquedas. Posteriormente, explica la construcción de estrategias mediante operadores booleanos (AND, OR) y el uso de filtros relacionados con idioma, fecha de publicación, población y diseño metodológico. Se presentan las características específicas de cada plataforma. La BVS facilita el acceso a literatura regional y en español; PubMed amplía la búsqueda hacia evidencia internacional de alto impacto; SciELO aporta estudios latinoamericanos de libre acceso y Cochrane Library ofrece revisiones sistemáticas y síntesis de evidencia de máxima calidad metodológica. Conclusiones: La búsqueda bibliográfica efectiva es un proceso metodológico que trasciende la simple utilización de palabras clave. Ninguna base de datos es suficiente por sí sola; por ello, la búsqueda combinada maximiza la recuperación de información pertinente, la capacidad de localizar, evaluar y utilizar críticamente la evidencia científica es una habilidad indispensable para el profesional de la salud contemporáneo, ya que fortalece la toma de decisiones clínicas, promueve una atención centrada en el paciente y contribuye al desarrollo de una práctica médica más rigurosa, transparente y fundamentada en el conocimiento científico.

Palabras clave: Medicina Basada en la Evidencia; Búsqueda bibliográfica; DeCS y MeSH; Bases de datos biomédicas; Estrategia PICO

1. CASO CLÍNICO

Un médico residente de primer año de Medicina Interna en un hospital de segundo nivel de atención de la ciudad de La Paz, Bolivia (3.600 msnm). Durante la consulta ambulatoria evalúa a un paciente de 62 años con antecedente de tabaquismo de 45 paquetes año, actualmente exfumador, diagnosticado con enfermedad pulmonar obstructiva crónica (EPOC) clasificada como GOLD grupo E, con alto riesgo de exacerbaciones. En los últimos 12 meses, el paciente ha presentado tres exacerbaciones agudas de la enfermedad, dos manejadas con corticosteroides sistémicos de forma ambulatoria y una que requirió hospitalización. Actualmente recibe tratamiento inhalado combinado con broncodilatadores de acción prolongada (LAMA/LABA) y oxigenoterapia domiciliaria nocturna. El paciente reside en la ciudad de El Alto, ubicada a más de 20 kilómetros del centro hospitalario. Refiere dificultades para asistir regularmente a los controles y terapias presenciales debido a las limitaciones de movilidad asociadas a su enfermedad, la distancia geográfica, los costos de transporte y las condiciones propias de la altitud. Durante la evaluación clínica presenta una disnea grado 3 según la escala Modified Medical Research Council (mMRC) y una puntuación de 22 puntos en el COPD Assessment Test (CAT), hallazgos compatibles con una importante carga sintomática y deterioro funcional. Asimismo, manifiesta fatiga durante actividades habituales de la vida diaria, con repercusión significativa sobre su calidad de vida.

El neumólogo tratante ha recomendado el inicio de un programa de rehabilitación pulmonar supervisada tres veces por semana en el hospital. Sin embargo, el paciente expresa interés en alternativas que le permitan realizar la rehabilitación en su domicilio o mediante estrategias de TELESALUD. En particular, desea conocer si estas modalidades ofrecen beneficios comparables a los programas presenciales en relación con:

- Mejoría de la capacidad funcional (prueba de caminata de 6 minutos e Incremental Shuttle Walking Test [ISWT]).
- Calidad de vida relacionada con la salud (COPD Assessment Test [CAT] y St. George'sRespiratoryQuestionnaire [SGRQ]).
- Control de síntomas respiratorios y disnea.
- Reducción de exacerbaciones y hospitalizaciones.
- Adherencia al tratamiento y al programa de rehabilitación.

Solicita que las recomendaciones estén sustentadas en evidencia científica reciente, preferentemente proveniente de revisiones sistemáticas, metaanálisis o ensayos clínicos de alta calidad metodológica, e idealmente disponibles en idioma español, con el fin de discutir las opciones terapéuticas junto con su familia. Ante esta situación clínica, surge la siguiente pregunta: ¿la rehabilitación pulmonar domiciliaria o basada en TELESALUD es tan efectiva como la rehabilitación pulmonar presencial para mejorar los desenlaces clínicos relevantes en pacientes con EPOC de alto riesgo?

2. OBJETIVOS

- Realizar búsquedas reproducibles y eficientes en las principales plataformas de literatura biomédica:BVS, PubMed, SciELO, Cochrane Library
- Aplicar en la práctica los conceptos aprendidos previamente sobre: Preguntas con PICO, uso de tesauros DeCS y MeSH, operadores booleanos y filtros avanzados
- Diseñar estrategias de búsqueda adaptadas al contexto clínico, considerando idioma, región geográfica, tipo de estudio y fecha de publicación.
- Evaluar la relevancia de los resultados para su paciente o pregunta clínica, priorizando artículos de mayor calidad y pertinencia.
- Documentar y guardar la estrategia de búsqueda, de modo que sea verificable, repetible y útil para informes, protocolos o publicaciones.

3. INTRODUCCIÓN

En la práctica clínica actual, el conocimiento adquirido durante la formación profesional y la experiencia acumulada, aunque fundamentales, no son suficientes para responder de manera adecuada a todas las interrogantes que surgen durante la atención de los pacientes. La Medicina Basada en la Evidencia (MBE) exige que las decisiones diagnósticas, terapéuticas y preventivas se sustenten en la mejor evidencia científica disponible, integrada con la experiencia clínica y las preferencias del paciente.En el caso presentado, un paciente con enfermedad pulmonar

https://doi.org/10.62486/978-9915-9928-3-9.Ch08

obstructiva crónica (EPOC) avanzada plantea una pregunta clínicamente relevante sobre la efectividad de la rehabilitación pulmonar domiciliaria o mediante telesalud en comparación con los programas presenciales tradicionales. Responder esta interrogante requiere mucho más que opiniones o experiencias individuales; demanda la búsqueda, identificación y análisis crítico de la evidencia científica más actualizada y metodológicamente sólida.

La capacidad para localizar información científica relevante constituye una competencia esencial para todo profesional de la salud. Una búsqueda bibliográfica eficiente permite acceder a estudios de alta calidad, minimizar el riesgo de tomar decisiones basadas en información incompleta y favorecer una atención centrada en evidencia confiable y aplicable al contexto clínico.

Con este propósito, el presente capítulo proporciona una guía práctica para la búsqueda sistemática y reproducible de literatura científica en las principales bases de datos biomédicas. Se abordará el uso de la Biblioteca Virtual en Salud (BVS), empleando los descriptores DeCS; PubMed, mediante los términos MeSH; SciELO, como fuente fundamental de literatura científica latinoamericana; y Cochrane Library, considerada uno de los principales referentes internacionales en revisiones sistemáticas y síntesis de evidencia de alta calidad. El dominio de estas herramientas permitirá al lector optimizar sus estrategias de búsqueda y acceder de manera eficiente a la mejor evidencia disponible para la toma de decisiones clínicas fundamentadas.

4. MEDIANTE LA BIBLIOTECA VIRTUAL DE SALUD

La Biblioteca Virtual en Salud (BVS) es la puerta de entrada a múltiples bases de datos latinoamericanas y globales, como Literatura Latinoamericana y del Caribe en Ciencias de la Salud (LILACS), Scientific Electronic Library Online (SciELO), MEDLINE y otras. Es especialmente útil cuando buscamos literatura en español o estudios realizados en nuestra región.

En este ejercicio, aplicaremos directamente el caso del paciente para obtener resultados relevantes en la BVS.

4.1: ACCEDER A BIREME Y AL TESAURO DECS

Accede al sitio de BIREME: Ve a la Biblioteca Virtual en Salud (BVS) enhttps://bvsalud.org.

Fuente: https://bvsalud.org.

4.2. IDENTIFICACIÓN DE TÉRMINOS DECS

Aunque ya trabajamos en capítulos anteriores cómo encontrar y seleccionar términos DeCS, aquí lo haremos de forma breve para la práctica:Haz clic en DeCS/MeSH o entra a**https://decs.bvsalud.org.**

https://doi.org/10.62486/978-9915-9928-3-9.Ch08

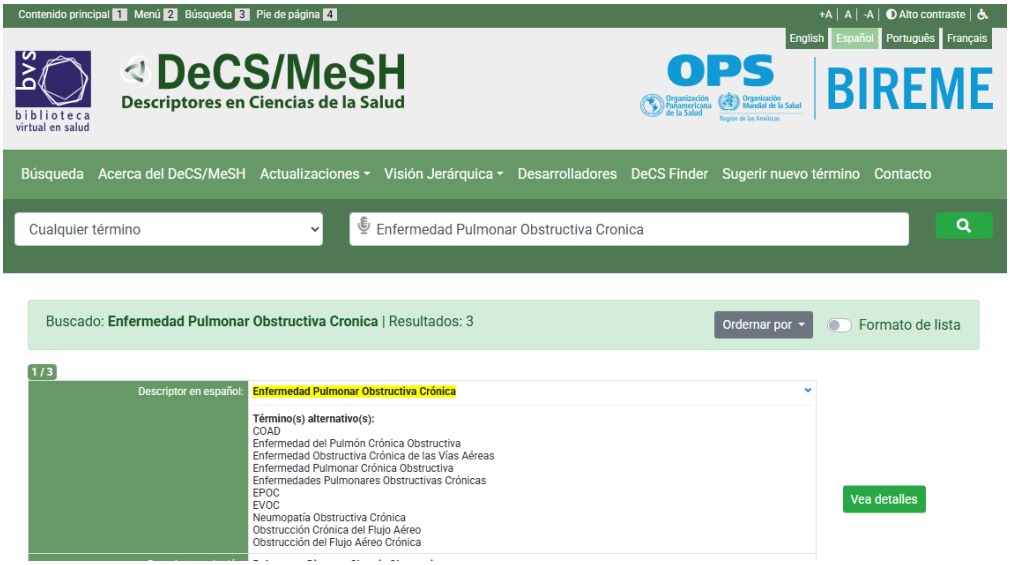
Buscar:

- "Enfermedad Pulmonar Obstructiva Crónica" en término DeCS oficial: Enfermedad Pulmonar Obstructiva Crónica. Sinónimos: EPOC, Chronic Obstructive PulmonaryDisease.
- "Rehabilitación pulmonar" → término DeCS oficial: Rehabilitación Pulmonar.

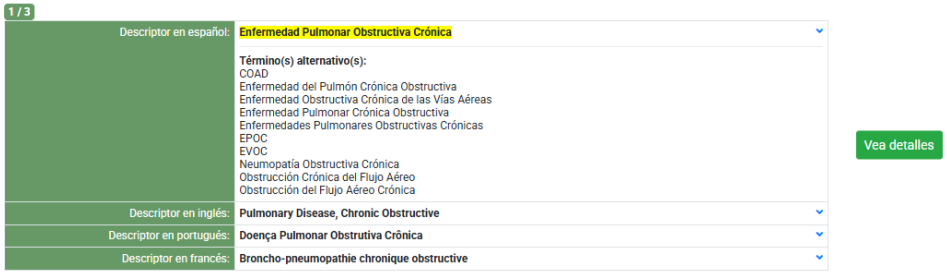
Puedes añadir términos relacionados como:

- Telemedicina
- Atención Domiciliaria de Salud

Se debe anotar los términos tal cual aparecen para usarlos en la búsqueda.



AAAA



BBBBDDDDCCCCEEEE

https://doi.org/10.62486/978-9915-9928-3-9.Ch08

A. Encabezado de búsqueda: "Buscado: Enfermedad Pulmonar Obstructiva Crónica, Resultados: 3": Indica el término que escribiste en el buscador y cuántos resultados (entradas) encontró en el tesauro que coinciden con ese término.

B. Descriptores: Descriptor en español: Aquí aparece el término oficial que debes usar en búsquedas dentro de la BVS: Enfermedad Pulmonar Obstructiva Crónica. Usar el descriptor oficial es clave porque las bases de datos lo reconocen y te devuelven resultados más precisos, incluso si los autores usan sinónimos en sus textos.

C. Término(s) alternativo(s): Lista de sinónimos y variantes que otros podrían usar para referirse a lo mismo: Ej.: COAD, EPOC, Enfermedad del Pulmón Crónica Obstructiva, etc. Estos no son los términos que se colocan directamente como descriptor, pero sirven para identificar cómo puede aparecer el concepto en artículos y para ampliar búsquedas con términos libres.

D. Traducciones del descriptor

- En inglés: "PulmonaryDisease, Chronic Obstructive": este es el término MeSH equivalente, para usar en PubMed.
- En portugués: "Doença Pulmonar ObstrutivaCrônica": útil para SciELO o búsquedas en Brasil/ Portugal.
- En francés: "Broncho-pneumopathiechronique obstructive": útil si quieres incluir literatura en ese idioma.

Código(s) jeráquico(s):	C08.381.495.389 C23.550.291.500.875
Identificador Único RDF:	https://id.nlm.nih.gov/mesh/D029424
Nota de alcance:	Una enfermedad causada por obstrucción crónica, difusa e irreversible del flujo aéreo pulmonar. Las subcategorías de EPOC comprenden la BRONQUITIS CRÓNICA y el ENFISEMA PULMONAR.
Nota de indización:	ENFERMEDADES PULMONARES OBSTRUCTIVAS también está disponible
Calificadores permitidos:	BL sangre CF líquido cefalorraquídeo CI inducido químicamente CL clasificación CN congénito CO complicaciones DG diagnóstico por imagen DH dietoterapia DI diagnóstico DT tratamiento farmacológico EC economía EH etnología EM embriología EN enzimología EP epidemiología ET etiología GE genética HI historia IM inmunología ME metabolismo

E. Botón "Vea detalles": Al hacer clic, podrás acceder a: Código jerárquico: muestra cómo se relaciona con otros términos más amplios o específicos dentro del árbol MeSH/DeCS. Identificador único RDF: Enlace único a la ficha oficial en MeSH. Nota de alcance: define con precisión el concepto. Nota de indización: menciona otros descriptores relacionados disponibles. Calificadores permitidos: Lista de abreviaturas que indica subaspectos del tema para búsquedas más precisas

4.3. UTILIDAD PRÁCTICA.

En la búsqueda del caso del paciente:

- El descriptor correcto es "Enfermedad Pulmonar Obstructiva Crónica".
- En la estrategia de búsqueda en BVS, debe ser escrito tal cual y, si es posible, combinado con otros descriptores como "Rehabilitación Pulmonar" y "Telemedicina".
- Las traducciones permiten replicar la misma búsqueda en PubMed (MeSH), SciELO en portugués y otras bases.

El Tesauro DeCS es la "gramática oficial" de las bases de datos de la BVS. Usar los descriptores correctos aumenta la precisión y evita que se pierdan artículos relevantes porque un autor usó un término diferente.

4.3. CONSTRUCCIÓN DE LA ESTRATEGIA

En la BVS, usaremos operadores booleanos para combinar términos:

"Enfermedad Pulmonar Obstructiva Crónica" [DeCS] AND "Rehabilitación Pulmonar" [DeCS]

https://doi.org/10.62486/978-9915-9928-3-9.Ch08

AND ("Telemedicina" [DeCS] OR "Atención Domiciliaria de Salud" [DeCS])

- Enfermedad Pulmonar Obstructiva Crónica: Garantiza que la base de datos recupere solo artículos indexados bajo este concepto exacto, incluso si en el texto los autores usan sinónimos como "EPOC" o "bronquitis crónica obstructiva". Lo usamos primero porque es la población de nuestro PICO: adultos con EPOC moderado–severo.
- **"Rehabilitación Pulmonar" [DeCS]:** También es un descriptor DeCS, representa la intervención principal de nuestro PICO, incluirlo asegura que los resultados se centren en programas estructurados de rehabilitación pulmonar, no en fisioterapia genérica o ejercicios no dirigidos a EPOC.
- **("Telemedicina" [DeCS] OR "Atención Domiciliaria de Salud" [DeCS]):** Aquí tenemos dos términos alternativos para describir la modalidad de la intervención. Telemedicina: cubre programas de rehabilitación a distancia por medios digitales, Atención Domiciliaria de Salud: incluye programas presenciales en el hogar, supervisados o guiados por profesionales.
- Usamos OR porque queremos que aparezca cualquiera de los dos en los resultados. Esto amplía la búsqueda para incluir todas las variantes relevantes.
- **AND**: exige que ambos términos estén presentes en los registros recuperados.
- **OR**: incluye artículos que contengan cualquiera de los términos listados.
- **Paréntesis**: agrupan la parte del OR para que la base de datos entienda que es una sola condición y que se combine en bloque con los otros términos.

Se interpretaría como: Dame artículos que: "sean sobre Enfermedad Pulmonar Obstructiva Crónica **Y** que incluyan Rehabilitación Pulmonar **Y** que hablen de Telemedicina **O** Atención Domiciliaria de Salud.

4.4 EJECUCIÓN DE LA BÚSQUEDA

- Vuelve a la página principal de la BVS
- Ingresa a "Búsqueda Avanzada"
- Copia la estrategia en la barra de búsqueda.
- Clic en "Buscar"

4.5 APLICACIÓN DE FILTROS:

Se encuentran en la barra lateral izquierda, se puede filtrar:

- Texto completo: Si están disponibles
- Base de datos: incluir LILACS, MEDLINE y otros que le interese.
- Asunto principal: Según criterio
- Tipo de estudio: revisiones sistemáticas, ensayos clínicos, estudios observacionales.
- Idioma: español, inglés, portugués.
- Intervalo de año de publicación: Últimos 5 años

https://doi.org/10.62486/978-9915-9928-3-9.Ch08

4.6. EVALUACIÓN DE LOS RESULTADOS

- Revisa títulos y resúmenes.
- Identifica si los estudios miden: Capacidad funcional (6MWT, ISWT), calidad de vida (CAT, SGRQ), exacerbaciones y hospitalizaciones.
- Descarta artículos no relacionados con EPOC o que no comparen modalidades de rehabilitación.

4.7. EXPORTACIÓN Y GUARDADO

- Marca los artículos relevantes.
- Usa la opción Enviar a / Guardar en RIS para importarlos luego a un gestor bibliográfico como Mendeley o Zotero.
- Guarda la estrategia de búsqueda para que puedas repetirla o actualizarla.

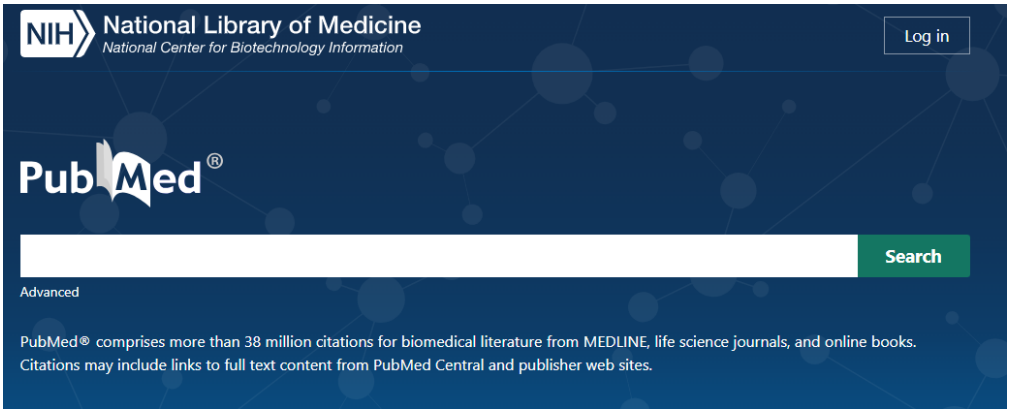
5. BÚSQUEDA DE ARTÍCULOS CIENTÍFICOS EN PUBMED

PubMed es una de las bases de datos biomédicas más completas, con más de 37 millones de registros. Su fortaleza está en la cobertura internacional, artículos de alto impacto y actualizaciones diarias. Es mantenida por la Biblioteca Nacional de Medicina de los Estados Unidos.

En este acápite, replicaremos la búsqueda de la BVS pero usando MeSH debido a que los términos deben ser en inglés.

5.1: ACCEDER A PUBMED

Visita el sitio web oficial de PubMed:https://pubmed.ncbi.nlm.nih.gov.



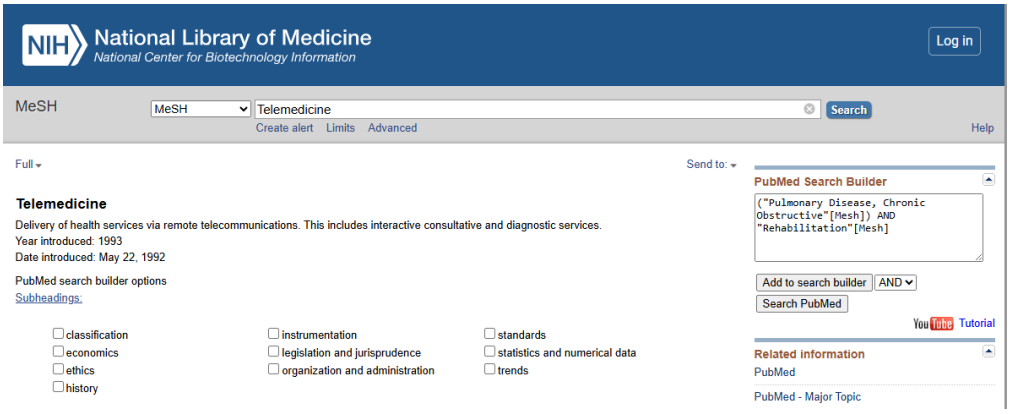
5.2 IDENTIFICACIÓN DE TÉRMINOS MESH

En el menú principal de PubMed:

- Ingresa a MeSH Database.
- Busca los términos equivalentes:
- Enfermedad Pulmonar Obstructiva Crónica: MeSH: “PulmonaryDisease, Chronic Obstructive” y haz clic en “Add tosearchbuilder”
- Rehabilitación Pulmonar: MeSH: “PulmonaryRehabilitation”
- Telemedicina: MeSH: *“Telemedicine”*
- Atención Domiciliaria de Salud: MeSH: *“Home Care Services”*

Alabrir cada término MeSH, puedes marcar la casilla “Restrictto MeSH Major Topic” si quieres que el término sea el foco principal del artículo.

https://doi.org/10.62486/978-9915-9928-3-9.Ch08



5.3 CONSTRUCCIÓN DE LA ESTRATEGIA CON BOOLEANOS

La estrategia equivalente en PubMed sería:

"Pulmonary Disease, Chronic Obstructive"[MeSH] AND "Pulmonary Rehabilitation"[MeSH] AND ("Telemedicine"[MeSH] OR "Home Care Services"[MeSH])

5.4 EJECUCIÓN DE LA BÚSQUEDA

- Haz clicken "Search PubMed".
- También se puede realizar copiar y pegar la estrategia de búsqueda mencionada previamente.
- Clic en "Search".

5.5. Aplicación de filtros: En el caso del paciente, filtra por:

- Publication dates: últimos 10 años.
- Languages: English, Spanish, Portuguese.
- Article types: Clinical Trial, Randomized Controlled Trial, Systematic Review, Meta-Analysis.
- Species: Humans.

5.6. EVALUACIÓN DE LOS RESULTADOS

- Revisa títulos y abstracts buscando: Medición de capacidad funcional (6MWT, ISWT), Calidad de vida (CAT, SGRQ), Frecuencia de exacerbaciones - hospitalizaciones.
- Prioriza estudios con población similar a la de don Jorge.

5.7. GUARDADO Y EXPORTACIÓN

- Usa Send to > Citation manager para exportar a Mendeley, Zotero o EndNote.
- Guarda la estrategia de búsqueda en tu cuenta de PubMed para repetirla más adelante.

5.8 ERRORES FRECUENTES A EVITAR EN PUBMED

- No mezclar términos MeSH con texto libre sin usar correctamente los campos ([MeSH] vs. [tiab]).
- No usar paréntesis al combinar OR con AND (la búsqueda se interpreta mal).
- Olvidar filtros de idioma y tipo de estudio cuando se necesita evidencia aplicable.

5.9 BÚSQUEDA DE ARTÍCULOS EN SCIELO

SciELO (Scientific Electronic Library Online) es una biblioteca científica que abarca una colección seleccionada de revistas científicas de América Latina, el Caribe, España y Portugal. Es una fuente valiosa para acceder a literatura científica en español y portugués. Es muy útil para:

- Encontrar literatura en español y portugués.

https://doi.org/10.62486/978-9915-9928-3-9.Ch08

- Acceder a investigaciones en contextos similares a los de nuestros pacientes.
- Localizar estudios que muchas veces no están indexados en PubMed.

En el caso del paciente, SciELO puede aportarnos evidencia regional sobre programas de rehabilitación pulmonar domiciliaria o a distancia en pacientes con EPOC.

6.1. ACCESO A SCIELO

- Entra a https://scielo.org y selecciona **SciELO Regional** o directamente la colección de un país específico (por ejemplo, SciELO Bolivia, Brasil, Chile).
- En la barra de búsqueda, encontrarás la opción de "Búsqueda avanzada" para aplicar filtros desde el inicio.

6.2. SELECCIÓN DE TÉRMINOS

Aunque SciELO no indexa usando MeSH, sí podemos usar términos DeCS para alinear la búsqueda, los mismos que utilizamos en la BVS

En SciELO, no es necesario escribir [DeCS], pero es recomendable colocar el término exacto y combinarlo con operadores booleanos.

6.3. ESTRATEGIA DE BÚSQUEDA

Para SciELO, nuestra estrategia sería:

"Enfermedad Pulmonar Obstructiva Crónica" AND "Rehabilitación Pulmonar" AND ("Telemedicina" OR "Atención Domiciliaria de Salud")

6.4. APLICACIÓN DE FILTROS

En la búsqueda avanzada de SciELO puedes filtrar por:

- Año de publicación: últimos 10 años.
- Idioma: Español y portugués (si se quiere mayor cercanía cultural y lingüística).
- Tipo de literatura: Artículos científicos, revisiones.
- País: Por ejemplo, Bolivia, Brasil, Chile o un conjunto de países latinoamericanos.

6.5. Revisión de resultados: En nuestro caso:

- Priorizar estudios que incluyan población latinoamericana y, si es posible, en altitudes similares (útil para la aplicabilidad clínica).
- Verifica que reporten resultados como 6MWT, CAT, SGRQ, exacerbaciones o adherencia.

6.6. GUARDADO DE ARTÍCULOS

- SciELO permite descargar PDF completos de forma gratuita.
- Guardar los archivos y referencias completas en un gestor bibliográfico.

https://doi.org/10.62486/978-9915-9928-3-9.Ch08

6.7 ERRORES FRECUENTES EN SCIELO

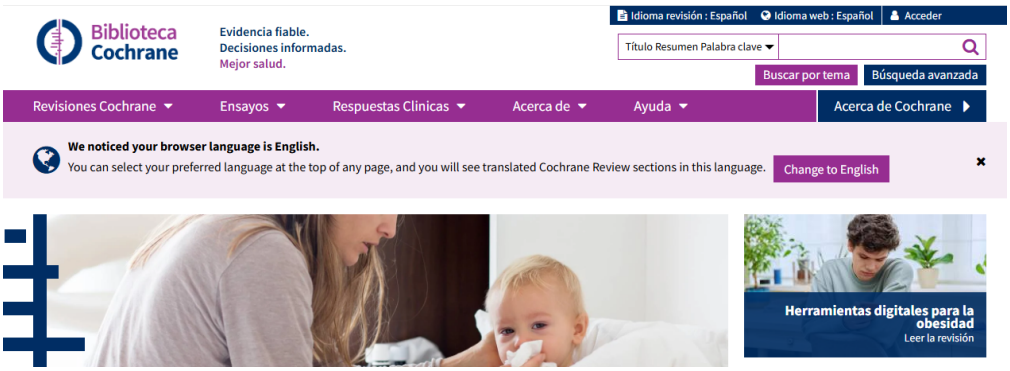
- Usar sólo términos libres sin probar los oficiales DeCS.
- No aplicar filtros de país o idioma, lo que lleva a resultados menos relevantes.
- No aprovechar que SciELO permite combinar términos en español y portugués en la misma búsqueda.

7. BÚSQUEDA EN COCHRANE

La Cochrane Library es una de las fuentes más reconocidas a nivel mundial para encontrar revisiones sistemáticas y, en algunos casos, ensayos clínicos de alta calidad. Es ideal cuando necesitamos una síntesis crítica de la evidencia disponible, como en el caso del paciente, donde queremos una respuesta rápida y confiable.

7.1. ACCESO A COCHRANE

- Ingresa a https://www.cochranelibrary.com
- Se puede usar la barra de búsqueda principal o la opción Advanced Search para mayor precisión.



7.2. SELECCIÓN DE TÉRMINOS

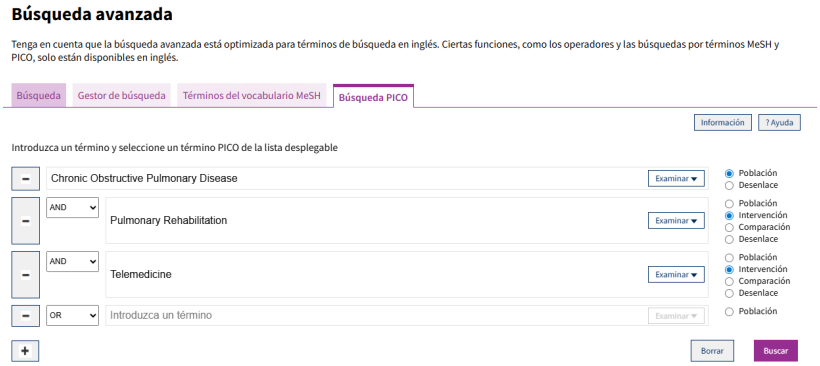
En Cochrane, la indexación está en inglés, por lo que debemos usar términos MeSH ya buscados cuando utilizamos PubMed

7.3. ESTRATEGIA DE BÚSQUEDA

En Cochrane podemos utilizar la misma búsqueda que se realizó en PubMed:

(Chronic Obstructive Pulmonary Disease" AND "Pulmonary Rehabilitation" AND (Telemedicine OR "Home Care Services")

Así mismo, se puede realizar una búsqueda más específica, utilizando la Búsqueda avanzada de Cochrane, en Búsqueda PICO para separar términos y combinar con booleanos:



https://doi.org/10.62486/978-9915-9928-3-9.Ch08

7.4. APLICACIÓN DE FILTROS

- Cochrane Reviews: revisiones sistemáticas Cochrane (evidencia más sólida).
- Other Reviews: revisiones no Cochrane (útiles si no hay revisiones Cochrane recientes).
- Trials: ensayos clínicos.
- Date: últimos 5–10 años para información actualizada.

7.5. REVISIÓN DE RESULTADOS

- En revisiones Cochrane, leer primero el Plain LanguageSummary para tener un resumen claro y comprensible para el paciente.
- Revisar el Abstract y las tablas de resultados para ver magnitud del efecto, calidad de la evidencia y aplicabilidad.

7.6. GUARDADO

- Puedes exportar la cita en formato RIS o copiar la referencia.
- Si la revisión es relevante, guarda también el PDF del resumen y las tablas.

8. CONCLUSIONES

La búsqueda de evidencia científica no se limita a escribir unas palabras clave en un buscador. Requiere un método estructurado, que permita localizar información relevante, de calidad y aplicable al contexto clínico del paciente.

En el caso del paciente, la aplicación de la estrategia en diferentes plataformas demostró que cada base de datos aporta un valor distinto:

- Biblioteca Virtual en Salud (BVS): Proporcionó acceso a literatura en español y portugués, incluyendo estudios regionales, es útil para encontrar evidencia contextualizada a América Latina, la combinación de DeCS y operadores booleanos permite filtrar de forma precisa.
- PubMed: Amplió la búsqueda a literatura internacional de alto impacto, el uso de MeSH y filtros avanzados garantiza mayor precisión y calidad metodológica de los estudios identificados, permite guardar estrategias y recibir alertas para actualizaciones.
- SciELO: Aportó artículos originales y revisiones en español y portugués, relevantes para entornos con recursos limitados o características socioeconómicas similares, facilita el acceso a texto completo gratuito, sin barreras de pago.
- Cochrane Library: Ofreció revisiones sistemáticas de alta calidad metodológica, permite obtener, en minutos, una síntesis de la mejor evidencia disponible.

9. RECOMENDACIONES:

- Ninguna base de datos por sí sola es suficiente.
- La búsqueda combinada maximiza las posibilidades de encontrar información completa y actualizada.
- Usar términos controlados (DeCS/MeSH) y operadores booleanos aumenta la precisión y la sensibilidad de la búsqueda.
- Adaptar la estrategia de búsqueda al idioma, tipo de estudio y contexto del paciente es esencial para la aplicabilidad clínica.
- Documentar y guardar la estrategia permite repetir y actualizar la búsqueda en el futuro.

Para agilizar la búsqueda en diversas plataformas, esta tabla que permite contener los términos MeSH/DeCS, pregunta de investigación y estrategia de búsqueda con operadores booleanos:

https://doi.org/10.62486/978-9915-9928-3-9.Ch08

Tabla 1. Guía práctica para la búsqueda de artículos científicos en salud

ESTRATEGIA DE BÚSQUEDA CON PICO			
Pregunta de investigación:			
	Términos	MeSH	DeCS
P(Pacientes)			
I (Intervención)			
C (Comparación)			
O (Outcome)			
Estrategia en Ingles (MeSH)			
Estrategia en Español (DeCS)			

Fuente: Elaboración propia

EJEMPLO:

ESTRATEGIA DE BÚSQUEDA CON PICO			
Pregunta de investigación: En pacientes con EPOC, ¿la rehabilitación pulmonar domiciliaria o por telemedicina es tan efectiva como la rehabilitación presencial en mejorar la capacidad funcional, la calidad de vida y reducir exacerbaciones?			
	Términos	MeSH	DeCS
P(Pacientes)	EPOC	Pulmonary Disease, Chronic Obstructive	Enfermedad Pulmonar Obstructiva Crónica
I (Intervención)	Rehabilitación pulmonar domiciliaria, tele-rehabilitación	Pulmonary Rehabilitation; Telemedicine; Home Care Services	Rehabilitación Pulmonar; Telemedicina; Atención Domiciliaria de Salud
C (Comparación)	Rehabilitación pulmonar presencial, centro de salud	Pulmonary Rehabilitation (sin calificadores de modalidad)	Rehabilitación Pulmonar (sin calificadores de modalidad)
O (Outcome)	Capacidad funcional, calidad de vida, exacerbaciones, hospitalizaciones	Exercise Tolerance; Quality of Life; Hospitalization; Disease Exacerbation	Tolerancia al Ejercicio; Calidad de Vida; Hospitalización; Exacerbación de la Enfermedad
Estrategia en Ingles (MeSH)	"Pulmonary Disease, Chronic Obstructive"[MeSH] AND "Pulmonary Rehabilitation"[MeSH] AND ("Telemedicine"[MeSH] OR "Home Care Services"[MeSH])		
Estrategia en Español (DeCS)	"Enfermedad Pulmonar Obstructiva Crónica" [DeCS] AND "Rehabilitación Pulmonar" [DeCS] AND ("Telemedicina" [DeCS] OR "Atención Domiciliaria de Salud" [DeCS])		

10. PREGUNTAS DE AUTOEVALUACIÓN

PREGUNTAS DE OPCIÓN MÚLTIPLE

1. En la Biblioteca Virtual en Salud (BVS), ¿qué herramienta se utiliza para identificar términos controlados en español y portugués que estandarizan la búsqueda bibliográfica?

a) Medical Subject Headings (MeSH)
b) Descriptores en Ciencias de la Salud (DeCS)
c) Digital Object Identifier (DOI)
d) International Standard Serial Number (ISSN)

2. ¿Qué operador booleano es el más adecuado para combinar sinónimos de un mismo concepto y así aumentar la sensibilidad de la búsqueda?

a) AND
b) OR
c) NOT
d) NEAR

3. En PubMed, ¿cuál es el término MeSH correcto para el concepto "Enfermedad Pulmonar Obstructiva Crónica"?

a) Chronic Pulmonary Disease

https://doi.org/10.62486/978-9915-9928-3-9.Ch08

b) Pulmonary Disease, Chronic Obstructive
c) Chronic Obstructive Lung Disease
d) Pulmonary Obstructive Syndrome

4. ¿Cuál es la principal ventaja de SciELO en comparación con PubMed al realizar búsquedas clínicas en América Latina?
a) Mayor número de ensayos clínicos multicéntricos
b) Acceso predominante a artículos en inglés
c) Mayor cobertura de literatura regional en español y portugués
d) Indexación exclusiva mediante MeSH

5. En Cochrane Library, ¿qué tipo de documento ofrece la síntesis metodológicamente más robusta para apoyar la toma de decisiones clínicas?
a) Clinical Trials
b) Editorials
c) Cochrane Reviews
d) Case Reports

PREGUNTAS ABIERTAS

6. Explique por qué el uso de descriptores controlados (DeCS/MeSH) mejora la calidad de una búsqueda bibliográfica en comparación con el uso exclusivo de términos libres.

7. Construya una estrategia de búsqueda en la BVS para el caso clínico de don Jorge, utilizando descriptores DeCS y operadores booleanos.

8. Mencione un filtro clave que aplicaría en PubMed para aumentar la relevancia clínica de los resultados y justifique su elección.

9. ¿Qué tipo de estudios priorizaría en Cochrane Library para responder de forma rápida y confiable la pregunta clínica planteada en este capítulo?

10. Compare SciELO y PubMed, señalando una ventaja y una limitación de SciELO en relación con PubMed.

RESPUESTAS:

1: b) Descriptores en Ciencias de la Salud (DeCS. DeCS es el tesauro utilizado por la BVS para estandarizar conceptos en español y portugués. Permite recuperar artículos aunque los autores usen sinónimos diferentes, aumentando la precisión y reproducibilidad de la búsqueda.

2: b) OR. El operador OR se utiliza para unir sinónimos o términos equivalentes. Su función es aumentar la sensibilidad de la búsqueda, es decir, recuperar el mayor número posible de artículos relevantes sobre un mismo concepto.

3: b) Pulmonary Disease, Chronic Obstructive. Este es el término oficial MeSH utilizado por PubMed para indexar artículos sobre EPOC. Usarlo garantiza que se recuperen todos los artículos clasificados bajo ese concepto, independientemente del término usado por los autores.

4: c) Mayor cobertura de literatura regional en español y portugués. SciELO es especialmente útil para acceder a estudios realizados en América Latina, con contextos clínicos, sociales y económicos similares a los de nuestros pacientes, y con acceso gratuito a texto completo.

5. c) Cochrane Reviews. Las revisiones sistemáticas Cochrane siguen metodologías rigurosas, evalúan el riesgo de sesgo y sintetizan la evidencia disponible, por lo que representan uno de los niveles más altos de evidencia científica.

6. El uso de DeCS/MeSH permite una búsqueda estandarizada y reproducible, evita la pérdida de artículos por variaciones terminológicas y mejora tanto la precisión como la sensibilidad de los resultados.

7. “Enfermedad Pulmonar Obstructiva Crónica” [DeCS] AND “Rehabilitación Pulmonar” [DeCS] AND (“Telemedicina” [DeCS] OR “Atención Domiciliaria de Salud” [DeCS])

8. Filtro por tipo de estudio (revisiones sistemáticas, ensayos clínicos) o por población (humanos), ya que aumenta la relevancia clínica y reduce resultados irrelevantes.

9. Revisiones sistemáticas Cochrane, ya que permiten obtener una síntesis crítica y confiable de la evidencia en poco tiempo.

10.Ventaja: acceso a literatura regional y en idioma local. Limitación: menor número de estudios de alto impacto internacional en comparación con PubMed.

https://doi.org/10.62486/978-9915-9928-3-9.Ch08

REFERENCIAS

Rathbone J, Carter M, Hoffmann T, Glasziou P. A comparison of the performance of seven key bibliographic databases in identifying all relevant systematic reviews of interventions for hypertension. Syst Rev. 2016 Feb 9;5(1):27. Available at: https://doi.org/10.1186/s13643-016-0197-5.
Bramer WM, Rethlefsen ML, Kleijnen J, Franco OH. Optimal database combinations for literature searches in systematic reviews: A prospective exploratory study. Syst Rev. 2017 Dec 6;6(1):245. Available at: https://doi.org/10.1186/s13643-017-0644-y.
Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev. 2021;10(1):39. Available at: https://doi.org/10.1186/s13643-020-01542-z.
Carías A, Auza-Santivañez JC, Carias P, Condori Villca N, Vallejos-Rejas DRE, Velez Alejo RW, et al. Knowledge of research and scientific publication in medical students. Salud, Ciencia y Tecnología - Serie de Conferencias [Internet]. 2025 Jan. 1 [cited 2026 Jun. 14];4:689. Available from: https://doi.org/10.56294/sctconf2025689
Fontanive VN, Müller LP, Riva F, Seoane M, Celeste RK. Development and validation of high-sensitivity and high-specificity PubMed search filters for systematic and non-systematic reviews. BMC Med Res Methodol. 2025;25(1):66. Available at: https://doi.org/10.1186/s12874-025-02666-3.
DeMars MM, Perruso C. MeSH and text-word search strategies: Precision, recall, and their implications for library instruction. J Med Libr Assoc. 2022 Jan;110(1):23–33. Available at: https://doi.org/10.5195/jmla.2022.1283.
McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS peer review of electronic search strategies: 2015 guideline statement. J Clin Epidemiol. 2016 Jul;75:40–46. Available at: https://doi.org/10.1016/j.jclinepi.2016.01.021.
Navarro-Ruan T, Haynes RB. Preliminary comparison of the performance of the National Library of Medicine's systematic review publication type and the sensitive clinical queries filter for systematic reviews in PubMed. J Med Libr Assoc. 2022 Jan;110(1):43–46. Available at: https://doi.org/10.5195/jmla.2022.1286.
Salvador-Oliván JA, Marco-Cuenca G, Arquero-Avilés R. Errors in search strategies used in systematic reviews and their effects on information retrieval. J Med Libr Assoc. 2019 Apr;107(2):210–221. Available at: https://doi.org/10.5195/jmla.2019.567.
Leblanc V, Hamroun A, Bentegeac R, Le Guellec B, Lenain R, Chazard E. Added value of Medical Subject Headings terms in search strategies of systematic reviews: Comparative study. J Med Internet Res. 2024 Nov 19;26:e53781. Available at: https://doi.org/10.2196/53781.
Auza-Santiváñez JC, Quispe Cornejo AA, Hayes Dorado JP, Díaz Pérez B. Science education from the approach of innovation, science and technology. Salud, Ciencia y Tecnología [Internet]. 2022 Jul. 23 [cited 2026 Jun. 14];2:64. Availablefrom: https://doi.org/10.56294/saludcyt202264
Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. Cochrane Handbook for Systematic Reviews of Interventions. Version 6.5.1. Chichester (UK): John Wiley & Sons; 2025. Chapter 4, Searching for and selecting studies. Available at: https://training.cochrane.org/handbook/current/chapter-04.
Lefebvre C, Glanville J, Briscoe S, Littlewood A, Marshall C, Metzendorf MI, et al. Searching for and selecting studies. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. Cochrane Handbook for Systematic Reviews of Interventions. Version 6.5.1. Chichester (UK): John Wiley & Sons; 2025. Available at: https://training.cochrane.org/handbook/current/chapter-04.
Auza-Santiváñez JC, Carías Díaz JA, Vedia Cruz OA, Robles-Nina SM, Escalante CS, Apaza Huanca B. Interactive formats: considerations for scientific publications. Seminars in Medical Writing and Education [Internet]. 2023 Nov. 7 [cited 2026 Jun. 14];2:27. Available from: https://doi.org/10.56294/mw202327

Chapter 9 / Capítulo 9

Critical reading of scientific articles on health: A practical guide (Spanish Edition)
ISBN: 978-9915-9928-3-9
DOI: 10.62486/978-9915-9928-3-9.Ch09
Pages: 98-115

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Case Report and Case Series

Reporte de Casos y Serie de Casos

Jorge Marquez-Molina
Alberto Martin Diaz-Seminario
Rolando Antonio Quiroz Quispe

1 Hospital del Norte. Cochabamba, Bolivia.
2 Instituto Nacional Materno Perinatal - INMP. Perú.
3 Caja Bancaria Estatal de Salud. La Paz, Bolivia.

ABSTRACT

Introduction: Case reports and case series are fundamental descriptive designs in Evidence-Based Medicine due to their ability to identify early clinical signs, describe phenotypes, recognize adverse events, and generate research hypotheses. Although they do not establish causality or demonstrate therapeutic efficacy, they represent a valuable source of information for clinical practice, pharmacovigilance, and the detection of emerging health problems. Development: This chapter analyzes the methodological foundations, validity criteria, and principles of interpretation of both designs, using as an example a case of rhabdomyolysis secondary to the interaction between a statin and a macrolide. The importance of clinical chronology, the exclusion of differential diagnoses, biological plausibility, and dechallenge is emphasized as elements that strengthen the credibility of the findings. Descriptive analysis tools, such as medians, interquartile ranges, box plots, and Kaplan-Meier curves, are described, along with the main biases and limitations inherent in these studies. Conclusions: Critical appraisal of case reports and case series allows for the transformation of clinical observations into useful knowledge for healthcare decision-making. Their true value lies in generating early warnings and guiding future research, provided they are interpreted with methodological rigor and within their limitations. It should be emphasized that these designs serve as a reminder that systematic clinical observation remains an essential pillar for the advancement of medical knowledge and the protection of patient safety.

Keywords: Case Report; Case Series; Evidence-Based Medicine; Pharmacovigilance; Selection Bias

RESUMEN

Introducción: Los reportes de caso y las series de casos constituyen diseños descriptivos fundamentales en la Medicina Basada en la Evidencia, debido a su capacidad para identificar señales clínicas tempranas, describir fenotipos, reconocer eventos adversos y generar hipótesis de investigación. Aunque no permiten establecer causalidad ni demostrar eficacia terapéutica, representan una fuente valiosa de información para la práctica clínica, la farmacovigilancia y la detección de problemas emergentes en salud. Desarrollo: El capítulo analiza los fundamentos metodológicos, criterios de validez y principios de interpretación de ambos diseños utilizando como ejemplo un caso de rabdomiólisis secundaria a la interacción entre una estatina y un macrólido. Se enfatiza la importancia de la cronología clínica, el descarte de diagnósticos diferenciales, la plausibilidad biológica y el dechallenge como elementos que fortalecen la credibilidad de los hallazgos. De esta manera, se describen herramientas de análisis descriptivo, como medianas, rangos intercuartílicos, diagramas de caja y curvas de Kaplan-Meier, además de los principales sesgos y limitaciones inherentes a estos estudios. Conclusiones: La lectura crítica de reportes y series de casos permite transformar observaciones clínicas en conocimiento útil para la toma de decisiones sanitarias. Su verdadero valor radica en generar alertas tempranas y orientar investigaciones futuras, siempre que se interpreten con rigor metodológico y dentro de sus límites. Se debe destacar que estos diseños recuerdan que la observación clínica sistemática continúa siendo un pilar esencial para el avance del conocimiento médico y la protección de la seguridad del paciente.

Palabras clave: Reporte de Caso; Serie de Casos; Medicina Basada en la Evidencia; Farmacovigilancia; Sesgo de Selección

1. CASO CLÍNICO

Paciente de 68 años de edad, con antecedentes de hipertensión arterial y dislipidemia adecuadamente controladas, que consulta por mialgias intensas y debilidad muscular progresiva de 48 horas de evolución, asociadas a coluria y disminución del volumen urinario. Tres días antes del inicio de los síntomas había comenzado tratamiento con claritromicina por un cuadro de bronquitis bacteriana. Su medicación habitual incluía atorvastatina 40 mg/día, losartán y ácido acetilsalicílico a dosis bajas. En la evaluación clínica se evidenció dolor muscular difuso y sensibilidad aumentada a la palpación de grupos musculares proximales, particularmente en muslos y cinturas escapular y pélvica, sin fiebre ni déficit neurológico focal. Los estudios complementarios revelaron una elevación marcada de creatinfosfoquinasa (CK), incremento de la creatinina sérica respecto a sus valores basales y una tira reactiva positiva para sangre con escasos eritrocitos en el examen microscópico del sedimento urinario, hallazgos compatibles con mioglobinuria. Ante la sospecha de rabdomiólisis, se suspendieron de forma inmediata la atorvastatina y la claritromicina, instaurándose hidratación intravenosa intensiva y monitorización estrecha de la función renal y del equilibrio hidroelectrolítico. Durante la evolución clínica se observó mejoría progresiva de la sintomatología, descenso sostenido de los niveles de CK, recuperación de la diuresis y normalización gradual de la función renal.

La secuencia temporal de los acontecimientos resulta clínicamente coherente y biológicamente plausible. Se trata de un paciente previamente estable que recibía una estatina metabolizada predominantemente por la isoenzima CYP3A4 y que, tras la introducción de un macrólido con potente efecto inhibidor sobre dicha vía metabólica, desarrolló en un intervalo de 48 a 72 horas un cuadro compatible con rabdomiólisis con compromiso renal agudo. El razonamiento diagnóstico contempló diversas alternativas etiológicas, entre ellas las mialgias asociadas al uso de estatinas, las miopatías inflamatorias y otras causas metabólicas o tóxicas de lesión muscular. No obstante, la magnitud de la elevación de CK, la presencia de mioglobinuria, el deterioro renal concomitante y la estrecha relación temporal con la introducción de claritromicina orientaron hacia una rabdomiólisis secundaria a una interacción farmacológica. Adicionalmente, la mejoría clínica y bioquímica observada tras la suspensión de ambos fármacos constituye un dechallenge positivo, elemento que fortalece la plausibilidad causal de la asociación descrita. Las evidencias que sustentan el diagnóstico corresponden a hallazgos clínicos y de laboratorio de alta relevancia diagnóstica en este contexto: elevación significativa de CK, mioglobinuria documentada, deterioro renal agudo incipiente y una exposición reciente a una combinación farmacológica ampliamente reconocida por incrementar el riesgo de toxicidad muscular inducida por estatinas. En consecuencia, la fortaleza del caso no radica en la complejidad de las pruebas complementarias empleadas, sino en la consistencia de la cronología clínica, la claridad del razonamiento diagnóstico y la exclusión razonada de explicaciones alternativas.

2. OBJETIVOS DE UN REPORTE DE CASO Y SERIE DE CASOS

2.1 REPORTE DE CASO:

Este tipo de artículos tienen los siguientes objetivos.

- Identificar una señal clínica puntual (reacción adversa, complicación rara, presentación atípica, enfermedad emergente).
- Evaluar la coherencia temporal entre exposición y desenlace.
- Confirmar el retiro de la exposición (dechallenge), la reexposición (rechallenge), cuando existan, y la evolución clínica posterior.
- Reconocer el razonamiento diagnóstico, incluyendo el descarte de alternativas.
- Valorar la plausibilidad biológica de la explicación propuesta.
- Extraer un mensaje clínico breve y aplicable (alerta, enseñanza, hipótesis).

2.2 SERIE DE CASOS:

Este tipo de artículos tienen los siguientes objetivos

- Describir el fenotipo clínico de un grupo de pacientes con la misma condición.
- Delinear la variabilidad en síntomas, hallazgos y desenlaces.
- Reconocer patrones temporales (tiempo hasta recuperación, complicación o muerte).

https://doi.org/10.62486/978-9915-9928-3-9.Ch09

- Evaluar la consistencia metodológica (captación consecutiva, periodo/ámbito definidos, definiciones diagnósticas estables).
- Interpretar el análisis descriptivo (medianas, proporciones, tiempo a evento) sin inferir causalidad.
- Considerar la serie como una plataforma de hipótesis para futuros estudios analíticos.
- El objetivo compartido al leer críticamente en ambos tipos de estudio es:
- Detectar señales plausibles y relevantes para la práctica clínica.
- Reconocer los sesgos y limitaciones que condicionan la confianza en el reporte o serie.
- Formular, cuando la señal lo justifique, una pregunta PICO que oriente la búsqueda en estudios de mayor nivel (cohortes, casos y controles, ensayos).

3. INTRODUCCIÓN

Los reportes de caso y las series de casos pertenecen a la familia de los estudios descriptivos: documentan, con detalle clínico, uno o varios pacientes que comparten un fenómeno relevante. Se conceptualiza también como el informe detallado de un individuo o individuos con características comunes que se utilizan para describir algún aspecto clínico, fisiopatológico u operativo de una enfermedad, tratamiento, exposición o procedimiento diagnóstico. Su utilidad no radica en probar causalidad ni eficacia, sino en hacer visibles patrones, generar hipótesis y alertar sobre señales clínicas o de seguridad que merecen atención. Bien leídos, permiten reconocer presentaciones atípicas, delinear fenotipos y anticipar riesgos; mal leídos, pueden inducir a conclusiones que estos diseños no están en condiciones de sostener.Un reporte de caso de calidad ofrece una narrativa clínica completa y coherente: motivo de consulta, cronología verificable, razonamiento diagnóstico con alternativas consideradas y pruebas que realmente inclinan la balanza, además de la evolución tras las decisiones terapéuticas. Una serie de casos añade orden y consistencia: define un periodo y un ámbito, explicita criterios de inclusión y exclusión, utiliza definiciones operacionales estables y resume la información con estadística descriptiva apropiada. Se ha sugerido que un informe con más de cuatro casos se denomine una serie de casos y aquellos con menos de cuatro, un informe de caso. En ambos, la transparencia metodológica es más importante que la rareza del tema: el lector debe poder distinguir qué es dato y qué es interpretación del autor.

En la jerarquía de la evidencia, estos diseños ocupan un nivel bajo para preguntas de causalidad o eficacia, pero su valor pragmático en la práctica clínica es alto. En farmacovigilancia, por ejemplo, las primeras señales de una reacción adversa emergen a menudo de casos con temporalidad coherente y mejoría tras retirar el fármaco. En enfermedades raras o emergentes, los reportes y series trazan el mapa inicial de manifestaciones, pruebas útiles y desenlaces, y orientan algoritmos diagnósticos que luego se pondrán a prueba en estudios analíticos. Esa es la clave para el lector: aprovechar la señal sin exigirle a este tipo de evidencia lo que no puede dar.

La lectura crítica, en reportes, se valora la claridad de la línea del tiempo, la explicitación del diagnóstico diferencial y la solidez de las pruebas clave; en eventos adversos, además, se espera una estimación razonada de causalidad clínica (temporalidad, dechallenge/rechallenge, descarte de explicaciones alternativas). En series, la credibilidad depende de cómo se seleccionaron los pacientes (idealmente de forma consecutiva), de la definición del periodo y el ámbito, de la consistencia de las variables y de la completitud del seguimiento. En ambos, el lector debe vigilar sesgos previsibles y preguntarse qué impacto tendrían sobre el mensaje clínico.

A lo largo del capítulo se utilizará el caso reportado en un inicio—rabdomiólisis tras la combinación estatina–macrólido— para entrenar estas competencias. Con esto el lector logrará interpretar reportes y series con criterio, extraerá un mensaje clínico seguro y los convertirá en punto de partida para una búsqueda dirigida que guíe decisiones clínicas o futuros estudios.

4. METODOLOGÍA

En la lectura crítica, los objetivos son diferentes para un reporte de caso y para una serie de casos, aunque comparten el carácter descriptivo y generador de hipótesis.

4.1. REPORTE DE CASO: ¿PARA QUÉ SIRVE CUANDO LO LEO?

El objetivo clínico de un reporte de caso es hacer visible una señal: una presentación

https://doi.org/10.62486/978-9915-9928-3-9.Ch09

atípica, una combinación poco usual de hallazgos, una reacción adversa o una complicación rara pero relevante. Como lector, lo primero es verificar si la cronología está bien documentada (temporalidad coherente, "dechallenge/rechallenge" cuando aplique) y si la explicación propuesta tiene plausibilidad biológica sin saltos lógicos.

El objetivo metodológico es transparentar el razonamiento diagnóstico. Un buen reporte muestra alternativas consideradas y descartadas, y señala las pruebas que cambiaron la probabilidad (qué dato "inclinó la balanza"). La lectura crítica exige distinguir datos (examen, imagen, evolución) de interpretación (hipótesis del autor) y estimar si el desenlace tiene seguimiento suficiente para ser creíble.El objetivo educativo es enseñar decisiones clínicas en incertidumbre: qué se hizo, por qué, con qué riesgos y cómo se vigiló la seguridad. Si el caso sugiere causalidad, el lector debería encontrar una estimación explícita de causalidad clínica y un descarte razonable de explicaciones alternativas.

Aplicado a nuestro caso, el objetivo válido para el lector es identificar una señal de seguridad accionable (interacción farmacológica grave), reconocer la cadena temporal y extraer un mensaje clínico breve y prudente (vigilar síntomas, medir CK, suspender a tiempo).

4.2. SERIE DE CASOS: ¿QUÉ AÑADE CUANDO LA INTERPRETO?

La serie de casos persigue ordenar varias observaciones comparables para delinear un fenotipo y su variabilidad. Su objetivo es describir, no comparar. Como lector, el hilo conductor es: ¿los pacientes fueron captados de forma consecutiva en un periodo y ámbito definidos?, ¿hay criterios operacionales claros y consistentes (diagnóstico, exposición, desenlaces)?, ¿el seguimiento permite confiar en lo reportado?.En términos de resultados, el objetivo legítimo es ofrecer resúmenes descriptivos robustos (medianas y rangos intercuartílicos, proporciones con intervalos, eventualmente tiempo a evento dentro de la serie). Se puede explorar patrones ("los casos con X característica parecen evolucionar peor"), pero el lector debe recordar que no hay control de confusión ni base para inferir eficacia. Si la serie sugiere seguridad o daño, lo correcto es tratarlo como señal temprana que justifica estudios analíticos posteriores, no como veredicto definitivo.

Una buena serie también cumple un objetivo de factibilidad: ayuda al lector a juzgar si existe volumen de pacientes y definiciones estables como para diseñar una cohorte, un estudio de casos y controles o, más adelante, un ensayo clínico.

Regla de oro para el lector crítico:

- Si su pregunta es "¿funciona mejor X que Y?" o "¿cuál es el riesgo relativo de Z?", ni un reporte ni una serie podrán responder.
- Si su pregunta es "¿esto puede pasar?", "¿cómo se manifiesta?", "¿qué debo vigilar y cuándo?" o "¿hay una señal de daño que requiera precaución inmediata?", entonces está frente al diseño correcto y debe exprimir su valor descriptivo y de alerta.

4.3. METODOLOGÍA DE REPORTES Y SERIES DE CASOS: CÓMO INTERPRETARLAS Y QUÉ EXIGIR

REPORTE DE CASO: ELEMENTOS METODOLÓGICOS QUE SUSTENTAN LA CONFIANZA

La lectura de la metodología de un reporte de caso debe responder, al menos, dos preguntas: si es posible reconstruir la ruta clínica sin saltos y si la explicación propuesta es plausible frente a alternativas razonables.

En primer término, debe precisarse el contexto asistencial (servicio, nivel de complejidad) y la fuente de datos (historia clínica, laboratorio, imágenes, entrevista). La cronología constituye el armazón metodológico: las fechas o intervalos han de permitir seguir el orden de síntomas, pruebas e intervenciones. Cuando la línea de tiempo es confusa, la inferencia también lo es.

La metodología sólida permite la definición operativa del diagnóstico y el diagnóstico diferencial. No basta con "revelar" un diagnóstico; se requiere mostrar cómo se descartaron alternativas (criterios clínicos, pruebas negativas relevantes, umbrales aplicados). En reacciones adversas a medicamentos, debe existir una estimación explícita de causalidad clínica: coherencia temporal, presencia de dechallenge (mejoría tras retirar el fármaco) y, cuando exista, rechallenge. El uso de escalas estandarizadas no es universalmente obligatorio, pero incrementa la credibilidad cuando se aplica y se justifica.

La exposición/intervención debe estar descrita con precisión (dosis, vía, duración, cambios

https://doi.org/10.62486/978-9915-9928-3-9.Ch09

y motivos). Los desenlaces han de definirse con claridad (clínicos, de laboratorio, imagen), con periodicidad y seguimiento suficientes para creer el resultado (recuperación, recaída, complicación). Dos componentes metodológicos a menudo omitidos y que conviene exigir son: una mini-búsqueda de literatura que contextualice el caso (bases, términos y fecha de corte) y la constancia de consentimiento informado y anonimización (especialmente si se incluyen imágenes).

Por último, la metodología debe permitir distinguir datos de interpretaciones. Cuando las conclusiones descansan en razonamientos no sustentados por los datos presentados, o se omiten explicaciones alternativas plausibles, la confianza disminuye. Un buen reporte no pretende probar causalidad ni eficacia; aporta claridad, orden y plausibilidad.

SERIE DE CASOS: LA METODOLOGÍA QUE LEGITIMA LA DESCRIPCIÓN

En una serie de casos, la metodología es determinante. Lo primero que debe verificarse es cómo se formó la serie: periodo calendario acotado, ámbito asistencial y, de manera central, si la captación fue consecutiva (todos los casos que cumplían criterios) o de conveniencia. La consecutividad reduce el sesgo de selección y suele ser condición para confiar en la descripción del fenotipo. Los informes de casos (o series de casos) suelen ser retrospectivos, aunque ocasionalmente pueden ser prospectivos.

La definición operativa del caso ha de ser replicable (criterios diagnósticos explícitos, pruebas mínimas requeridas, umbrales definidos). Deben especificarse criterios de inclusión/exclusión y presentarse un diagrama de flujo con elegibles, incluidos y motivos de exclusión. Seguidamente, las variables deben quedar operacionalizadas (qué significa “mejoría”, qué punto de corte define “elevado”, qué prueba de imagen se considera “positiva”) y el seguimiento debe contemplar ventanas temporales apropiadas para capturar desenlaces (p. ej., a 30/90 días, o según la condición).El análisis en una serie legítimamente es descriptivo: medidas de tendencia central y dispersión adecuadas (medianas y rangos intercuartílicos cuando la distribución es asimétrica; medias y desviaciones estándar sólo con normalidad defendible), proporciones con denominador claro (n/N), manejo transparente de datos faltantes y, cuando procede, curvas de Kaplan–Meier internas a la serie. Pueden mostrarse exploraciones (p. ej., “los casos con X parecen recuperarse más lento”), pero no corresponde inferir diferencias causales. La presencia de valores p o modelos multivariables para comparar tratamientos dentro de una serie sugiere exceso inferencial.

La metodología debe reconocer sesgos previsibles (selección, duplicación de casos, clasificación de desenlaces, pérdidas al seguimiento) y describir estrategias para mitigarlos. En el plano ético, muchas series requieren aprobación de comité de ética o constancia de exención; en reportes individuales, al menos consentimiento. La existencia explícita de estos elementos incrementa la confianza en lo comunicado.

4.4. QUÉ MEDIDAS Y ANÁLISIS ESPERAR (Y CUÁLES NO)

De un reporte se esperan datos puntuales (valores de laboratorio, hallazgos de imagen) y un seguimiento que cierre el desenlace; no se esperan inferencias estadísticas. De una serie, se esperan resúmenes descriptivos sólidos: conteos, proporciones y, cuando es pertinente, intervalos de confianza para comunicar incertidumbre (con tamaños muestrales razonables). Las tasas por persona-tiempo sólo son pertinentes si el seguimiento es uniforme y bien medido. No corresponde estimar riesgos comparativos ni concluir eficacia.

4.5. SEÑALES DE CALIDAD Y BANDERAS ROJAS EN LA METODOLOGÍA

Constituyen señales de calidad: cronología clara; definiciones operacionales explícitas para caso/serie; captación consecutiva con periodo y ámbito definidos; seguimiento suficiente; manejo transparente de datos faltantes; y cumplimiento ético documentado. En eventos adversos, suma credibilidad una estimación razonada de causalidad y, cuando procede, la aplicación de una escala estandarizada.

Son banderas rojas: cronología imposible de reconstruir; diagnóstico afirmado sin mostrar cómo se descartaron alternativas; en series, muestra de conveniencia no declarada, denominadores imprecisos o cambios de definiciones a mitad del periodo; presentación de p-valor o modelos causales sin diseño para comparación; ausencia de consentimiento/anonimización (reporte) o de

https://doi.org/10.62486/978-9915-9928-3-9.Ch09

revisión ética/exención (serie, cuando aplica).

4.6. DE LA METODOLOGÍA A DECISIONES PRÁCTICAS

Aplicado al caso, si un paciente en estatina inicia un macrólido, la CK se eleva masivamente a las 48–72 horas, aparece mioglobinuria y, tras suspender ambos fármacos, se observa mejoría (dechallenge), la metodología presentada es suficiente para extraer una alerta de seguridad accionable: vigilar síntomas musculares ante combinaciones de estatinas metabolizadas por CYP3A4 con inhibidores potentes, medir CK y suspender a tiempo. En una serie sobre rabdomiólisis por interacciones, la interpretación será confiable cuando existan consecutividad, criterios operacionales uniformes y un análisis descriptivo prudente que permita trasladar a la práctica qué vigilar, cuándo y en quién.

5. EVALUACIÓN DE VALIDEZ Y RELEVANCIA CLÍNICA

La lectura crítica de reportes y series de casos exige separar dos juicios complementarios: validez (¿es creíble lo descrito?) y relevancia/aplicabilidad (¿debería cambiar la práctica?). En diseños descriptivos, la validez no se apoya en comparaciones ni estimaciones de efecto, sino en la calidad del registro clínico, la coherencia lógica de la explicación y la transparencia sobre sesgos y limitaciones.

5.1. MARCO GENERAL DE VALIDEZ (APLICABLE A AMBOS DISEÑOS)

Un lector riguroso suele basarse en seis pilares:

- **Cronología verificable:** los eventos clínicos y terapéuticos deben poder reconstruirse sin lagunas.
- **Plausibilidad biológica y clínica:** la explicación propuesta debe ser consistente con la fisiopatología y la literatura disponible.
- **Descartar alternativas razonables:** la ruta diagnóstica debe mostrar cómo se excluyeron hipótesis competidoras.
- **Mediciones válidas y consistentes:** pruebas e imágenes con criterios explícitos, mismos métodos a lo largo del tiempo.
- **Seguimiento suficiente:** tiempo adecuado para observar el desenlace de interés (recuperación, recaída, complicación).
- **Transparencia ética y de fuentes:** consentimiento/anonimización (reporte), aprobación o exención ética (serie), fuentes de datos claras.

Tabla 1. Pilares para la evaluación crítica de la validez en reportes y series de casos

Pilar	Qué debe buscar el lector	Pregunta clave en lectura crítica
Cronología verificable	Los eventos clínicos, diagnósticos y terapéuticos están ordenados en el tiempo y pueden reconstruirse sin lagunas.	¿Puedo seguir claramente qué ocurrió primero, qué después y en qué intervalo?
Plausibilidad biológica y clínica	La explicación propuesta es coherente con la fisiopatología conocida y con la evidencia previa publicada.	¿Lo que describen tiene sentido desde el punto de vista biológico y clínico?
Descarte de alternativas razonables	Se presentan diagnósticos diferenciales relevantes y se explica cómo fueron excluidos.	¿Se descartaron de forma explícita otras causas plausibles del desenlace?
Mediciones válidas y consistentes	Pruebas diagnósticas e imágenes con criterios explícitos, métodos reproducibles y consistencia a lo largo del seguimiento.	¿Las mediciones son confiables y comparables en el tiempo?
Seguimiento suficiente	El tiempo de observación permite evaluar adecuadamente el desenlace (recuperación, recaída o complicación).	¿El seguimiento es suficiente para creer el desenlace reportado?
Transparencia ética y de fuentes	Consentimiento informado y anonimización (reporte); aprobación o exención ética (serie); fuentes de datos claras.	¿El estudio es éticamente sólido y transparente en el origen de la información?

Fuente: Elaboración propia

https://doi.org/10.62486/978-9915-9928-3-9.Ch09

5.2. VALIDEZ EN REPORTES DE CASO: CRITERIOS ESPECÍFICOS

En un reporte individual, el núcleo de la validez es el razonamiento causal-clínico. La temporalidad (aparición del fenómeno tras la exposición), el dechallenge y, cuando ocurre, el rechallenge no intencional, incrementan la credibilidad. En reacciones adversas, el uso de una escala estandarizada de causalidad añade estructura al juicio, siempre subordinado al contexto clínico.

También se valora la calidad del diferencial (no basta con listar diagnósticos; debe explicarse por qué se descartaron), la pertinencia de las pruebas (que realmente cambian la probabilidad diagnóstica) y la mini-búsqueda que demuestra conocimiento del estado del arte. La ausencia de estos elementos reduce la validez, incluso si el desenlace fue favorable.

5.3. VALIDEZ EN SERIES DE CASOS: CRITERIOS ESPECÍFICOS

Depende fundamentalmente de cómo se conformó la cohorte descriptiva y de la consistencia con la que se registraron y analizaron las observaciones. A diferencia de los estudios analíticos, en este tipo de diseño no existe un grupo de comparación; por ello, la credibilidad de los hallazgos se apoya en la transparencia del proceso de selección y en la estandarización de las definiciones y mediciones.

Un elemento central es la captación consecutiva de los casos dentro de un periodo y un ámbito claramente definidos, lo que reduce el sesgo de selección y permite que la serie represente, de forma razonable, la experiencia clínica real del escenario estudiado. La definición operativa de "caso" debe ser explícita y replicable, especificando los criterios diagnósticos utilizados, los umbrales clínicos o bioquímicos aplicados y las pruebas mínimas requeridas para la inclusión. La ausencia de estas definiciones impide interpretar adecuadamente la homogeneidad de la serie.

Resulta indispensable la uniformidad en las mediciones a lo largo del periodo de estudio, así como una descripción clara del manejo de los datos faltantes, dado que la disponibilidad desigual de información puede distorsionar la interpretación de los desenlaces. La presentación de un diagrama de flujo que muestre los pacientes elegibles, incluidos y excluidos, junto con las estrategias empleadas para evitar la duplicación de casos, contribuye de manera significativa a la transparencia metodológica.La validez también está condicionada por un seguimiento suficiente y homogéneo para los desenlaces de interés, de modo que los resultados observados reflejen la evolución clínica real y no diferencias artificiales derivadas de pérdidas tempranas o ventanas de observación desiguales. Finalmente, el análisis debe mantenerse estrictamente descriptivo, utilizando medidas apropiadas como medianas y rangos intercuartílicos para variables continuas, proporciones con denominadores claramente especificados y, cuando corresponda, análisis de tiempo a evento dentro de la propia serie, evitando cualquier intento de inferencia causal o de evaluación de eficacia comparativa.

La validez de una serie de casos disminuye de forma significativa cuando se utilizan muestras de conveniencia no declaradas, cuando los denominadores varían sin explicación entre las variables analizadas, cuando las definiciones diagnósticas cambian a mitad del periodo o cuando se intenta extrapolar los resultados hacia conclusiones de eficacia o riesgo, objetivos que exceden las capacidades inherentes de este diseño.

5.4. RELEVANCIA Y APLICABILIDAD CLÍNICA

Incluso un reporte/serie válidos pueden ser poco útiles si la aplicabilidad es baja. Importa valorar:

- **Concordancia de población y escenario:** semejanza con pacientes, nivel de atención y recursos locales.
- **Gravedad y urgencia de la señal:** eventos raros pero graves (p. ej., toxicidades) pueden justificar cambios prudenciales inmediatos.
- **Factibilidad de la recomendación implícita:** disponibilidad de pruebas/terapias y capacidad de vigilancia.
- **Balance de riesgo/beneficio de actuar con señal preliminar:** en seguridad del paciente, la prudencia favorece la prevención; en eficacia, conviene esperar evidencia analítica.

En el caso de rabdomiólisis tras combinación estatina–macrólido, la temporalidad estrecha, la mioglobinuria, la CK marcadamente elevada y el dechallenge positivo sustentan un juicio de validez alta para una señal de seguridad. La aplicabilidad es elevada en múltiples contextos

https://doi.org/10.62486/978-9915-9928-3-9.Ch09

porque la interacción farmacológica es frecuente y prevenible. El semáforo resultaría verde para la acción clínica prudencial (evitar la combinación o intensificar vigilancia y suspender a tiempo), sin extrapolar a comparaciones de eficacia.

6. INTERPRETACIÓN ESTADÍSTICA

La estadística en reportes y series de casos es descriptiva y su propósito es resumir y hacer legibles las observaciones, no probar hipótesis ni estimar efectos causales. La lectura crítica debe concentrarse en la adecuación de los resúmenes, la transparencia de los denominadores, la gestión de la incertidumbre y la coherencia entre las medidas reportadas y la realidad clínica de la serie.

6.1. PRINCIPIOS GENERALES

En este tipo de artículos, la estadística útil:

- Caracteriza quiénes son los pacientes (distribuciones de edad, sexo, comorbilidades);
- Muestra qué ocurrió (valores clínicos y de laboratorio, hallazgos de imagen);
- Ordena cuándo ocurrió (línea de tiempo, evolución, tiempo a eventos clínicos

No corresponde comparar tratamientos ni reportar pruebas de significación entre subgrupos: esos análisis inducen a exceso inferencial y no son fiables sin un diseño comparativo.

6.2. EN REPORTES DE CASO (N=1)

La interpretación se basa en la coherencia interna de los datos y en su capacidad para respaldar la narrativa clínica. El lector debe centrarse en los valores absolutos de las variables clínicas, bioquímicas o imagenológicas por ejemplo, creatinfosfoquinasa (CK), creatinina sérica o marcadores inflamatorios, siempre interpretados en relación con los rangos de referencia y con el contexto clínico del paciente.

Un aspecto clave es la evaluación de los cambios a lo largo del tiempo dentro del mismo individuo. La presencia de tendencias claras, acompañadas de fechas o intervalos bien definidos, permite valorar la evolución del cuadro y su relación con las intervenciones realizadas. Esta lectura temporal es especialmente relevante cuando se plantea una posible asociación causal, ya que la coherencia temporal aparición del fenómeno tras la exposición sospechada y mejoría posterior a su retiradaconstituye uno de los pilares del razonamiento clínico.

En este contexto, las cifras deben respaldar y no sustituir la narrativa clínica. No tiene sentido reportar medias, porcentajes u otras medidas de resumen poblacional, ya que no aportan información válida en un análisis centrado en un solo paciente y pueden inducir a interpretaciones erróneas.

6.3. EN SERIES DE CASOS (N PEQUEÑO A MODERADO)

En las series de casos, la estadística cumple una función descriptiva, orientada a resumir de manera ordenada las características y desenlaces de un grupo de pacientes, sin pretender establecer comparaciones causales. Para variables continuas que suelen presentar distribuciones asimétricas como niveles de CK, proteína C reactiva o duración de la hospitalización, deben preferirse medianas acompañadas de rangos intercuartílicos, ya que describen mejor la tendencia central y la dispersión. El uso de medias y desviaciones estándar sólo es aceptable cuando la distribución es aproximadamente normal y dicha elección se encuentra justificada.

Las variables categóricas o binarias deben presentarse como proporciones con denominadores explícitos (n/N), de modo que el lector pueda identificar con claridad cuántos pacientes aportan información para cada variable. Cuando existen datos faltantes, es fundamental informar el número exacto de observaciones disponibles, evitando la ilusión de completitud que puede distorsionar la interpretación.

La información estadística se organiza en tablas claras que resumen las características basales y los desenlaces principales de la serie. El objetivo de la estadística en una serie de casos es describir con precisión, no demostrar diferencias ni efectos.

Por todo ello la estadística es de resumen. Debe preferirse:

https://doi.org/10.62486/978-9915-9928-3-9.Ch09

- **Medianas y rangos intercuartílicos (IQR)** para variables continuas típicamente asimétricas (CK, PCR, tiempos de hospitalización). Las medias y desviaciones estándar sólo son pertinentes si la distribución es aproximadamente normal y está justificado.
- **Proporciones con denominador explícito (n/N)** para variables binarias (p. ej., "lesión renal aguda en 6/12 pacientes"); informar el N exacto en cada variable si hay datos faltantes.
- **Tablas claras** de características basales y desenlaces, evitando dividir la serie en subgrupos "tratados vs. no tratados" como si fuera un estudio analítico.

6.4. INCERTIDUMBRE CON TAMAÑOS PEQUEÑOS: CÓMO LEER INTERVALOS DE CONFIANZA

Aunque las series de casos no tengan como objetivo realizar inferencias poblacionales, los intervalos de confianza (IC) cumplen una función clave en la lectura crítica: expresar la imprecisión inherente a los resúmenes obtenidos a partir de tamaños muestrales pequeños. Su correcta interpretación permite al lector calibrar expectativas y evitar conclusiones excesivas a partir de proporciones aparentemente contundentes.

Cuando el tamaño de la serie es reducido, los intervalos de confianza exactos (binomiales) o los intervalos de Wilson son los que se utilizan para intervalos normales. Por ejemplo, si en una serie se observa una proporción de 6 eventos en 12 pacientes (50%), el intervalo de confianza exacto al 95% será amplio, aproximadamente entre 21% y 79%, lo que comunica de manera explícita la baja precisión de la estimación y la elevada variabilidad asociada a ese tamaño muestral.

Un escenario particularmente propenso a la sobreinterpretación es la ausencia de eventos. Cuando se observa una proporción de 0/N, se aplica la denominada regla del tres, según la cual, con un 95% de confianza, la proporción verdadera del evento podría ser de hasta aproximadamente 3/N. Así, en una serie de 12 pacientes sin eventos, el límite superior del intervalo sería cercano al 25%, recordando al lector que la ausencia observada no equivale a ausencia real de riesgo.

Es fundamental enfatizar que estos intervalos no transforman una serie de casos en una estimación poblacional válida. Su utilidad radica exclusivamente en advertir cuán inestable y dependiente del tamaño es el resumen estadístico presentado.

6.5 TIEMPO DE EVENTO HASTA LA SERIE

En determinadas series de casos, el interés principal no reside únicamente en si ocurrió o no un desenlace, sino en cuándo ocurrió. En estos contextos por ejemplo, tiempo hasta normalización de creatinfosfoquinasa, tiempo hasta alta hospitalaria o tiempo hasta una complicación resulta apropiado utilizar análisis de tiempo a evento, como las curvas de Kaplan–Meier, siempre que se presenten de forma interna a la serie y con fines descriptivos.

La lectura crítica de estas curvas exige verificar, en primer lugar, que el evento esté claramente definido, es decir, que exista un punto inequívoco que marque su ocurrencia.

En segundo lugar, debe evaluarse la censura, identificando qué pacientes no alcanzaron el evento al final del seguimiento y las razones de ello. Finalmente, cuando el número de eventos lo permite, la mediana del tiempo a evento puede ofrecer una medida sintética útil de la trayectoria clínica.

En este diseño no corresponde calcular razones de riesgo, hazard ratios ni realizar comparaciones formales entre subgrupos. El propósito de estas curvas es exclusivamente describir la evolución temporal de los pacientes incluidos, no establecer diferencias causales.

Figura 1: Curva de Kaplan Meier: tiempo hasta la normalización de CK en una serie de casos. Se muestran el evento definido con CK < 200 U/L, datos censurados por alta médica (▲) y por perdida de seguimiento (Δ) y mediana del tiempo a 10 dias.

Esta curva de Kaplan–Meier muestra cuánto tiempo tardan los pacientes en alcanzar un desenlace específico, en este caso la normalización de la CK. El eje horizontal representa el tiempo de seguimiento, mientras que el eje vertical indica la proporción de pacientes que aún no han alcanzado el evento.

Cada descenso escalonado de la curva corresponde a uno o más pacientes que alcanzan el desenlace en ese momento del seguimiento. Es decir, cada bajada indica que algunos pacientes normalizaron la CK en ese punto del tiempo. Los tramos horizontales reflejan periodos en los que no ocurrieron eventos, aunque los pacientes continuaron en seguimiento.

https://doi.org/10.62486/978-9915-9928-3-9.Ch09

Figura 1: Tiempo hasta la normalización de Creatinfosfoquinasa

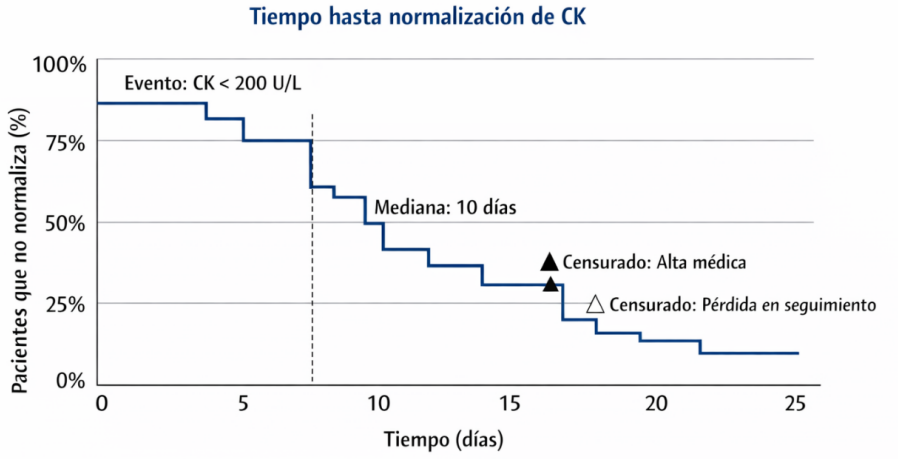


Figura X. Curva de Kaplan-Meier: tiempo hasta la normalización de CK en una serie de casos. Se muestran: evento definido como CK < 200 U/L, datos censurados por alta médica (▲) y por pérdida en seguimiento (△), y mediana del tiempo a evento de 10 días.

Los marcadores de censura (cuando están presentes) señalan pacientes que dejaron de ser observados antes de alcanzar el desenlace, ya sea por alta médica, final del seguimiento o pérdida de contacto. Estos pacientes aportan información hasta el momento en que se censuran, pero no indican fracaso ni éxito clínico.

La mediana del tiempo a evento se identifica cuando la curva cruza el 50% del eje vertical y representa el tiempo en el que la mitad de los pacientes ha alcanzado el desenlace. Este valor resume la trayectoria temporal típica dentro de la serie.

Desde una lectura crítica, esta curva debe interpretarse como una descripción de la evolución temporal de los pacientes incluidos. No compara grupos ni permite establecer relaciones causales; su finalidad es ayudar a comprender cómo y cuándo ocurren los desenlaces en esta serie específica.

6.6. VISUALIZACIÓN DE DATOS

Son útiles los gráficos por paciente (trayectorias individuales “spaghetti plots” de CK, creatinina), dotplots o boxplots para resumir distribuciones y run charts para series temporales. Deben evitarse figuras que induzcan a comparación causal (barras lado a lado “tratamiento A vs. B” con p-values), pues no corresponden a este diseño.

Este boxplot resume el tiempo hasta la normalización de la CK en una serie de casos y permite visualizar, de forma sintética, la distribución y variabilidad de ese desenlace.

La línea central dentro de la caja representa la mediana, que en este ejemplo se sitúa alrededor de 10 días. Esto indica que el 50% de los pacientes normalizó la CK antes de ese tiempo y el otro 50% después. La caja abarca el rango intercuartílico, aproximadamente entre 5 y 15 días, donde se concentra la mitad central de los pacientes; este intervalo refleja la variabilidad clínica típica dentro de la serie.

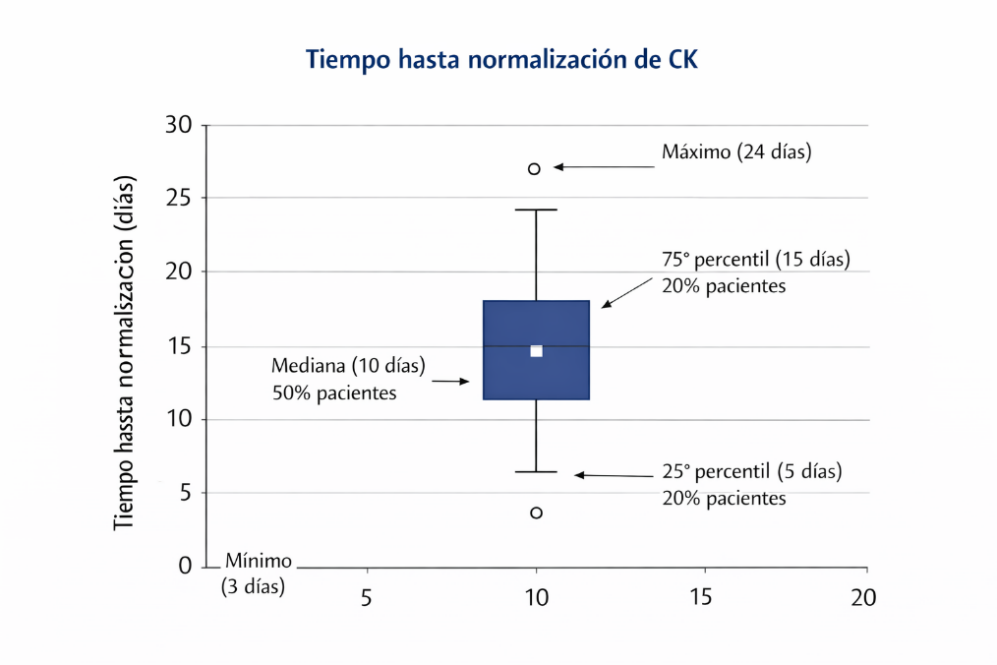
Los bigotes muestran los valores más extremos observados sin considerar valores atípicos, indicando que algunos pacientes normalizaron la CK de forma más temprana y otros requirieron un tiempo mayor. Los puntos aislados (cuando están presentes) corresponden a valores atípicos, es decir, pacientes con evoluciones claramente más rápidas o más prolongadas que la mayoría, lo cual es frecuente en series pequeñas y clínicamente heterogéneas.

Desde una lectura crítica, este gráfico permite comprender qué tan disperso es el tiempo de recuperación, identificar la asimetría de la distribución y reconocer la presencia de casos extremos, pero no permite comparar tratamientos ni inferir eficacia. Su función es descriptiva y complementa la narrativa clínica y otros análisis temporales, como las curvas de Kaplan–Meier,

https://doi.org/10.62486/978-9915-9928-3-9.Ch09

ofreciendo una visión rápida y honesta de la variabilidad dentro de la serie.

Figura 2: Tiempo hasta la normalizacion de la CK en boxplot



6.5. ERRORES FRECUENTES DE INTERPRETACIÓN

- Leer porcentajes sin denominador o con denominadores cambiantes.
- Elegir medias en distribuciones asimétricas (sesga la impresión de "promedio").
- Confundir ausencia de eventos con ausencia de riesgo (olvidar la regla del tres).
- Comparar subgrupos como si existiera aleatorización o control de confusión.
- Ignorar la censura en descripciones de tiempo a evento.
- Overclaim: traducir un patrón descriptivo en efectividad o seguridad comparativa.

En el caso clínico, si tuviéramos 12 pacientes similares, la lectura estadística esperable sería: edad mediana 70 años (IQR 66–74); CK mediana 12 000 U/L (IQR 8 500–18 000); lesión renal aguda 6/12 = 50% (IC exacto 95%: ≈21%–79%); requerimiento de diálisis 0/12, lo cual sugiere que el riesgo verdadero podría ser hasta ≈25% (regla del tres). El tiempo hasta normalización de CK podría representarse con Kaplan–Meier interna, con mediana de 6 días si al menos la mitad alcanzó el evento. Nada de esto prueba diferencias entre estatinas o antibióticos; simplemente delinea el fenotipo clínico y la evolución, con su incertidumbre explícita.

7. SESGOS Y LIMITACIONES

La interpretación de reportes y series de casos exige reconocer que se trata de diseños descriptivos: muestran qué ocurrió en uno o varios pacientes, pero no comparan grupos ni controlan confusión de manera formal. Su mayor fortaleza la cercanía al caso clínico reales, al mismo tiempo, su mayor vulnerabilidad: la evidencia depende de cómo se seleccionaron y documentaron las observaciones. Este acápite describe los sesgos más frecuentes, las limitaciones inherentes y cómo mitigarlos al leer.

7.1. POR QUÉ IMPORTAN LOS SESGOS EN DISEÑOS DESCRIPTIVOS

En ausencia de un contrafactual explícito, pequeñas deficiencias en la selección de pacientes, en la definición de caso o en la cronología pueden transformar una señal válida en una impresión engañosa. El lector competente no busca "perfección metodológica", sino coherencia

https://doi.org/10.62486/978-9915-9928-3-9.Ch09

y transparencia suficientes para confiar en el mensaje clínico prudente que estos diseños sí pueden aportar.

7.2. SESGOS TÍPICOS EN REPORTES DE CASO

Sesgo de selección del caso (notoriedad/rareza): se eligen casos llamativos o exitosos y se omiten presentaciones atípicas sin desenlace "interesante". Limita la inferencia sobre frecuencia o gravedad.

Sesgo de publicación y de reporte: tienden a publicarse hallazgos positivos/raros; dentro del manuscrito, pueden omitirse datos que debilitan el reporte de caso (p. ej., valores intermedios, pruebas negativas).

Sesgo de confirmación e "hindsight": la narrativa puede reorganizar hechos a la luz del diagnóstico final, exagerando la inevitabilidad del mismo y minimizando alternativas descartadas.

Clasificación/medición (misclasificación): definiciones imprecisas del diagnóstico o del desenlace; mediciones no estandarizadas (p. ej., distintos ensayos de laboratorio sin referencia).

Sospecha/ascertainment: se busca con mayor intensidad el fenómeno esperado y se pasan por alto hallazgos que lo contradicen; frecuente cuando el autor sospecha una reacción adversa.

Cronología incompleta: secuencias temporales borrosas impiden juzgar plausibilidad causal (especialmente en farmacovigilancia).

Confusión clínica no controlada: comorbilidades, polifarmacia, co-intervenciones o factores fisiológicos (edad, función renal/tiroidea) pueden explicar la observación tanto o más que la exposición postulada.

Riesgo ético y de privacidad: en enfermedades raras, la reidentificación es posible aun con anonimización deficiente; su ausencia reduce la confianza en todo el reporte.

7.3. SESGOS TÍPICOS EN SERIES DE CASOS

No consecutividad / muestra de conveniencia: incluir sólo pacientes "a mano" o remitidos sesga el fenotipo hacia cuadros más graves o particulares.

Definición de caso inestable (drift): cambiar criterios diagnósticos o umbrales a la mitad del periodo que altera la composición de la serie.

Duplicación: el mismo paciente puede aparecer en más de una publicación; "aumenta" artificialmente eventos raros.

Sesgo temporal/secular: cambios en prácticas clínicas o disponibilidad de pruebas durante el periodo producen diferencias que se atribuyen erróneamente al fenómeno descrito.

Heterogeneidad de co-intervenciones: múltiples tratamientos simultáneos impiden atribuir trayectorias a una intervención específica (aunque la serie no deba hacerlo, el lector podría sobreinterpretar).

Seguimiento insuficiente o diferencial: pérdidas tempranas o seguimiento más corto en quienes evolucionan mal distorsionan desenlaces.

Exceso inferencial: uso de p-valor o modelos comparativos como si existiera un diseño analítico.

7.4. LIMITACIONES INHERENTES QUE NO PUEDEN "ARREGLARSE" CON ESTADÍSTICA

Ausencia de contrafactual y control de confusión: no es posible aislar el efecto de una exposición de manera robusta.

Imposibilidad de estimar riesgos poblacionales: la serie describe lo observado, no la incidencia o el riesgo relativo en la población.

Generalización limitada: lo que vale en un entorno/periodo puede no aplicar en otros (casuística, recursos, pericia).

Tamaño muestral pequeño y variabilidad alta: la incertidumbre es intrínsecamente amplia, aun con resúmenes e intervalos de confianza.

7.5. CÓMO MITIGAR SESGOS AL LEER (ESTRATEGIAS DE INTERPRETACIÓN)

La lectura crítica de reportes y series de casos requiere verificar, en primer lugar, la existencia

https://doi.org/10.62486/978-9915-9928-3-9.Ch09

de una cronología clara, que permita establecer coherencia temporal entre exposición y desenlace; en ausencia de ella, la confianza en la asociación propuesta se reduce de forma importante. Es igualmente necesario examinar si se presenta un diagnóstico diferencial adecuado, con una explicación explícita de cómo se descartaron alternativas razonables, ya que su omisión suele reflejar sesgo de confirmación.

El lector debe valorar la claridad de las definiciones operacionales del diagnóstico y de los desenlaces, así como considerar el escenario clínico en el que se captaron los casos, evaluando si la consecutividad, el periodo y el ámbito asistencial limitan la generalización. La revisión de los datos faltantes y los denominadores por variable permite anticipar cómo la ausencia de información podría modificar la interpretación, evitando conclusiones optimistas injustificadas.

Finalmente, resulta fundamental distinguir los datos de las interpretaciones del autor y contrastar la observación con la literatura existente, mediante una búsqueda breve, para ubicar la señal clínica dentro del conjunto de la evidencia disponible y calibrar adecuadamente su relevancia.

CHECKLIST DE LECTURA CRÍTICA

Tabla 2. Checklist de lectura crítica para reportes de caso y series de casos

Dominio	Ítem de verificación	Pregunta guía para el lector crítico	Guía de respaldo
A. Verificación rápida	Identificación del diseño	¿Está claramente definido si es un reporte de caso único o una serie de casos?	CARE, JBI
	Mensaje clínico prudente	¿El mensaje es descriptivo y evita inferencias causales o de eficacia?	CARE, JBI
	Cronología comprensible	¿La secuencia temporal se entiende sin lagunas relevantes?	CARE
	Cumplimiento ético	¿Se declara consentimiento y anonimización (reporte) o aprobación/exención ética (serie)?	CARE, JBI
B. Reporte de caso	Línea de tiempo explícita	¿Se describe claramente la secuencia síntomas → pruebas → intervención → evolución?	CARE
	Diagnóstico diferencial	¿Se presentan alternativas razonables y se explica cómo fueron descartadas?	CARE, JBI
	Pruebas clave	¿Las pruebas diagnósticas realmente cambian la probabilidad clínica?	CARE
	Temporalidad y dechallenge	En eventos adversos, ¿hay coherencia temporal y mejoría tras retirar la exposición (dechallenge ± rechallenge)?	CARE, Naranjo, RUCAM
	Descripción de la intervención	¿Se detallan dosis, duración, cambios y motivos de la intervención/exposición?	CARE
	Seguimiento suficiente	¿El seguimiento permite creer el desenlace reportado?	CARE
	Contextualización	¿Incluye una mini-búsqueda bibliográfica que sitúe el caso en la literatura?	CARE
C. Serie de casos	Captación consecutiva	¿Los casos fueron incluidos de forma consecutiva en un periodo y ámbito definidos?	JBI, NHLBI
	Definición operativa de caso	¿El "caso" está definido con criterios replicables (diagnóstico, umbrales, pruebas mínimas)?	JBI
	Inclusión/exclusión y flujo	¿Se reportan criterios claros y un diagrama de flujo (elegibles, incluidos, excluidos)?	JBI, PROCESS
	Variables operacionalizadas	¿Los desenlaces y variables están claramente definidos?	JBI
	Seguimiento homogéneo	¿El tiempo de seguimiento es suficiente y comparable entre pacientes?	JBI
	Análisis descriptivo	¿El análisis es estrictamente descriptivo (mediana/IQR, n/N, IC cuando procede)?	JBI
	Datos faltantes	¿Se declaran datos faltantes y denominadores por variable?	JBI, NHLBI
	Duplicación y cambios seculares	¿Se evitó duplicar casos y se discutieron cambios temporales relevantes?	JBI
D. Interpretación estadística mínima	Resumen de continuas	¿Variables asimétricas se presentan con mediana/IQR (no medias engañosas)?	JBI

https://doi.org/10.62486/978-9915-9928-3-9.Ch09

Dominio	Ítem de verificación	Pregunta guía para el lector crítico	Guía de respaldo
	Proporciones claras	¿Las proporciones incluyen n/N y se considera la incertidumbre (IC o regla del tres, 3/N)?	JAMA, BMJ EBM
	Tiempo a evento	¿Se define evento, censura y mediana sin comparar subgrupos?	JBI
	Evitar inferencia indebida	¿Se evitan p-valores, modelos o comparaciones causales?	JBI
E. Sesgos y limitaciones	Sesgo de selección	¿Se reconoce posible sesgo por caso memorable, centro terciario o no consecutividad?	JBI
	Coherencia temporal	¿La cronología respalda la hipótesis sin sesgo retrospectivo ("hindsight")?	CARE
	Estabilidad de definiciones	¿Las definiciones se mantuvieron constantes durante la serie?	JBI
	Publicación y duplicación	¿Se discute el riesgo de publicación selectiva o duplicación?	BMJ
	Dato vs. interpretación	¿Se distingue claramente la observación de la interpretación del autor?	CARE
F. Aplicabilidad clínica	Similaridad del contexto	¿La población y el escenario se parecen a tu práctica clínica?	MBE
	Acción de bajo riesgo	¿La señal permite una acción prudente y de bajo riesgo?	MBE
	Necesidad de triangulación	¿Requiere contrastarse con otras capas de evidencia antes de actuar?	MBE
G. Juicio final (semáforo)	Verde	Señal coherente, método transparente, acción de bajo riesgo/alto beneficio	Síntesis MBE
	Amarillo	Vacíos relevantes; vigilar, buscar evidencia analítica y formular PICO	Síntesis MBE
	Rojo	Inferencias indebidas o metodología débil; no cambiar conducta clínica	Síntesis MBE

Fuente: Elaboración propia

9. UTILIDAD

En conjunto, los reportes y series de casos cumplen un rol relevante en distintos niveles del sistema de salud, siempre que se interpreten dentro de los límites propios de los diseños descriptivos.

En la práctica clínica, funcionan como alertas tempranas que permiten identificar reacciones adversas, interacciones farmacológicas y patrones clínicos poco frecuentes antes de que exista evidencia analítica. Además, contribuyen al razonamiento diagnóstico al mostrar secuencias temporales y trayectorias clínicas que afinan la sospecha y orientan el seguimiento, con la cautela que exige la incertidumbre de tamaños pequeños.

Desde la perspectiva de la **seguridad institucional y la gestión**, estas observaciones pueden activar comités de seguridad, auditorías clínicas y mejoras de procesos, traduciéndose en cambios de protocolos, alertas en sistemas de prescripción y estrategias de prevención de eventos adversos.

En el ámbito **educativo**, constituyen herramientas eficaces para entrenar el pensamiento clínico, hacer explícito el razonamiento diagnóstico y enseñar lectura crítica, ayudando a distinguir entre dato e interpretación y a reconocer sesgos frecuentes en estudios descriptivos.

Como insumo para la **investigación**, los reportes y series actúan principalmente como **generadores de hipótesis**. Facilitan la formulación de preguntas PICO realistas, aportan definiciones operacionales iniciales y permiten valorar la factibilidad de estudios analíticos posteriores más robustos.

Cuando las señales son plausibles y consistentes, y el riesgo de no actuar es elevado mientras que la intervención preventiva es de bajo costo o bajo riesgo, esta evidencia puede influir en guías clínicas, consensos o políticas institucionales, al menos de forma transitoria, hasta contar con estudios confirmatorios.

Su utilidad debe leerse junto a sus **límites**: no permiten estimar riesgos poblacionales ni comparar eficacia entre intervenciones, y no sustituyen la triangulación con otras capas de evidencia. Su uso prudente combina la valoración crítica de la calidad del reporte con el principio

https://doi.org/10.62486/978-9915-9928-3-9.Ch09

de precaución, actuando cuando el potencial de daño es alto y la acción propuesta es segura.

10. PREGUNTAS DE AUTOEVALUACIÓN

PREGUNTAS DE OPCIÓN MÚLTIPLE:

1. ¿Cuál es el uso correcto de reportes y series de casos? a) Estimar el riesgo relativo de una intervención b) Demostrar eficacia comparativa entre tratamientos c) Describir patrones clínicos y generar hipótesis sin probar causalidad d) Ajustar confusores con modelos multivariables

2. En un reporte que sugiere una reacción adversa a fármacos, lo que más aumenta la credibilidad es:

- Un valor $p < 0,05$
- Dechallenge positivo y temporalidad coherente (± rechallenge)
- Un gráfico de barras comparando dos tratamientos
- Un largo listado de diagnósticos alternativos sin explicación

3. En una serie de casos, ¿qué práctica reduce mejor el sesgo de selección?

a) Incluir únicamente los casos más graves b) Captación consecutiva en un periodo y ámbito definidos c) Cambiar definiciones diagnósticas según disponibilidad de pruebas d) Excluir a quienes presentan datos faltantes

4. Para resumir creatinfosfoquinasa (CK), con distribución asimétrica, en una serie de casos, la medida más apropiada es:

- Media y desviación estándar
- Mediana y rango intercuartílico (IQR)
- Moda
- Promedio truncado al 5%

5. En una serie de 12 pacientes no se registró necesidad de diálisis (0/12). Aproximadamente, ¿cuál es el límite superior del 95% para esa proporción según la regla del tres?

- 3%
- 8%
- 15%
- 25%

PREGUNTAS ABIERTAS:

6. ¿Por qué los reportes y series de casos no permiten afirmar que un tratamiento es eficaz?

7. Mencione dos situaciones clínicas en las que un reporte de caso puede ser especialmente útil para el médico asistencial.

8. ¿Qué significa que un estudio sea “generador de hipótesis”?

9. ¿Por qué es importante que un reporte de caso tenga una cronología clara de los eventos clínicos?

10. En una serie de casos, ¿por qué es preferible reportar medianas en lugar de medias?

RESPUESTAS

1. c). Los reportes y series de casos son diseños descriptivos. Su finalidad es identificar patrones clínicos y generar hipótesis, no estimar efectos ni establecer causalidad.

2. b) La coherencia temporal y la mejoría tras retirar la exposición sospechada (dechallenge) son pilares de la causalidad clínica en reacciones adversas.

3. b) La captación consecutiva, dentro de un marco temporal y asistencial definido, minimiza el sesgo de selección y mejora la validez descriptiva.

4. b). En variables continuas asimétricas, la mediana y el IQR representan mejor la tendencia central y la dispersión que la media y la desviación estándar.

5. d). La regla del tres estima el límite superior del 95% como 3/N; en este caso, 3/12 ≈ 25%, evitando interpretar erróneamente la “ausencia de riesgo”.

6. Porque no tienen grupo de comparación, no controlan factores de confusión y su objetivo es descriptivo, no causal.

7. Detección de reacciones adversas a medicamentos; presentación inusual de una enfermedad; complicaciones poco frecuentes; nuevas interacciones farmacológicas.

https://doi.org/10.62486/978-9915-9928-3-9.Ch09

8. Que no responde una pregunta definitiva, pero sugiere una posible asociación que debe estudiarse luego con diseños analíticos más robustos.

9. Porque permite evaluar la relación temporal entre exposición e desenlace, aumentando la credibilidad de la observación clínica.

10. Porque muchas variables clínicas tienen distribución asimétrica y la mediana representa mejor el valor típico sin distorsión por valores extremos.

https://doi.org/10.62486/978-9915-9928-3-9.Ch09

REFERENCIAS

Gagnier JJ, Kienle G, Altman DG, Moher D, Sox H, Riley DS, et al.; CARE Group. The CARE guidelines: Consensus-based clinical case report guideline development. J Clin Epidemiol. 2014;67(1):46–51. Available at: https://doi.org/10.1016/j.jclinepi.2013.08.003.
Riley DS, Barber MS, Kienle GS, Aronson JK, von Schoen-Angerer T, Tugwell P, et al. CARE guidelines for case reports: Explanation and elaboration document. J Clin Epidemiol. 2017;89:218–235. Available at: https://doi.org/10.1016/j.jclinepi.2017.04.026.
Murad MH, Sultan S, Haffar S, Bazerbachi F. Methodological quality and synthesis of case series and case reports. BMJ Evid Based Med. 2018;23(2):60–63. Available at: https://doi.org/10.1136/bmjebm-2017-110853.
Dekkers OM, Egger M, Altman DG, Vandenbroucke JP. Distinguishing case series from cohort studies. Ann Intern Med. 2012;156(1 Pt 1):37–40. Availableat: https://doi.org/10.7326/0003-4819-156-1-201201030-00006.
Vandenbroucke JP. In defense of case reports and case series. Ann Intern Med. 2001;134(4):330–334. Available at: https://doi.org/10.7326/0003-4819-134-4-200102200-00017.
Grimes DA, Schulz KF. Descriptive studies: What they can and cannot do. Lancet. 2002;359(9301):145–149. Availableat: https://doi.org/10.1016/S0140-6736(02)07373-7.
Hanley JA, Lippman-Hand A. If nothing goes wrong, is everything all right? Interpreting zero numerators. JAMA. 1983;249(13):1743–1745. Available at: https://doi.org/10.1001/jama.1983.03330370053031.
von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Explanation and elaboration. PLoS Med. 2007;4(10):e297. Availableat: https://doi.org/10.1371/journal.pmed.0040297.
Vandenbroucke JP. Observational research, randomised trials, and two views of medical science. PLoS Med. 2008;5(3):e67. Availableat: https://doi.org/10.1371/journal.pmed.0050067.
Loke YK, Price D, Derry S, Aronson JK. Case reports of suspected adverse drug reactions—systematic literature survey of follow-up. BMJ. 2006;332(7537):335–339. Availableat: https://doi.org/10.1136/bmj.38701.399942.63.
Albrecht J, Meves A, Bigby M. Case reports and case series from The Lancet had significant impact on medical literature. J Clin Epidemiol. 2005;58(12):1227–1232. Available at: https://doi.org/10.1016/j.jclinepi.2005.04.003.
Hill AB. The environment and disease: Association or causation? Proc R Soc Med. 1965;58(5):295–300. Available at: https://doi.org/10.1177/003591576505800503.

https://doi.org/10.62486/978-9915-9928-3-9.Ch09

Chapter 10 / Capítulo 10

Critical reading of scientific articles on health: A practical guide (Spanish Edition)
ISBN: 978-9915-9928-3-9
DOI: 10.62486/978-9915-9928-3-9.Ch10
Pages: 116-135

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Descriptive studies (cross-sectional, ecological, agreement)

Estudios descriptivos (transversal, ecológico, concordancia)

Mauricio Alejandro Paz Del Rio
Alejandro Carías
María Lourdes del Rosario Escalera Rivero
Adrian Avila-Hilari

1 Clínica Los Andes. La Paz, Bolivia.
2 Departamento de Cirugía de la Universidad Nacional Autónoma de Honduras. Tegucigalpa, Honduras.
3 Hospital de Clínicas. La Paz, Bolivia.
4 Clínica CIES "Critical Care Altitude" La Paz-Bolivia.

ABSTRACT

Introduction: Descriptive studies form the basis of epidemiological and clinical research, allowing for the quantification of the frequency and distribution of health problems in specific populations. Their main utility lies in estimating prevalence and incidence, identifying epiedemiological patterns, evaluating the agreement of measurements, and generating hypotheses for subsequent research, without establishing causal relationships. Development: This chapter analyzes the main descriptive designs: cross-sectional, ecological, concordance, case series, and longitudinal descriptive studies. It emphasizes the essential methodological elements for their critical appraisal, including the definition of the target population, sample representativeness, measurement quality, and the interpretation of prevalence, incidence, confidence intervals, and measures of agreement. It also examines the main biases associated with each design, such as selection bias, information bias, the ecological fallacy, and loss to follow-up. Conclusions: Descriptive studies are indispensable tools for understanding the magnitude and distribution of health problems, guiding clinical decisions, and supporting health planning. Their proper interpretation requires a rigorous evaluation of the validity, accuracy, and applicability of the results. Reflectively, these designs remind us that systematic observation is the first step in building scientific knowledge and generating evidence useful for clinical practice and public health.

Keywords: Descriptive Studies; Prevalence; Incidence; Ecological Fallacy; Concordance

RESUMEN

Introducción: Los estudios descriptivos constituyen la base de la investigación epidemiológica y clínica, al permitir cuantificar la frecuencia y distribución de los problemas de salud en poblaciones específicas. Su principal utilidad radica en estimar prevalencias e incidencias, identificar patrones epidemiológicos, evaluar la concordancia de mediciones y generar hipótesis para investigaciones posteriores, sin establecer relaciones causales. Desarrollo: El capítulo analiza los principales diseños descriptivos: estudios transversales, ecológicos, de concordancia, series de casos y estudios longitudinales descriptivos. Se enfatizan los elementos metodológicos esenciales para su lectura crítica, incluyendo la definición de la población objetivo, la representatividad muestral, la calidad de las mediciones, la interpretación de prevalencias, incidencias, intervalos de confianza y medidas de concordancia. Asimismo, se examinan los principales sesgos asociados a cada diseño, como el sesgo de selección, el sesgo de información, la falacia ecológica y las pérdidas de seguimiento. Conclusiones: Los estudios descriptivos son herramientas indispensables para comprender la magnitud y distribución de los problemas de salud, orientar decisiones clínicas y apoyar la planificación sanitaria. Su adecuada interpretación exige una evaluación rigurosa de la validez, precisión y aplicabilidad de los resultados. Reflexivamente, estos diseños recuerdan que la observación sistemática constituye el primer paso en la construcción del conocimiento científico y en la generación de evidencia útil para la práctica clínica y la salud pública.

Palabras clave: Estudios Descriptivos; Prevalencia; Incidencia; Falacia Ecológica; Concordancia

1. CASO CLÍNICO

En el Hospital Municipal de segundo nivel, la doctora de consulta externa nota un aumento en pacientes jóvenes con obesidad. María es uno de ellos, con 29 años, con IMC 32,1 kg/m² y acantosis nigricans. En la práctica clínica diaria, estos hallazgos individuales suelen percibirse como eventos aislados, con frecuencia reflejan cambios epidemiológicos más amplios en la población. Al buscar información para comprender el contexto, la doctora encuentra cuatro artículos publicados:

- Estudio transversal sobre prevalencia de obesidad en adultos de 18–39 años en La Paz.
- Estudio ecológico que compara tasas de obesidad por distrito y su relación con la densidad de tiendas de ultraprocesados.
- Estudio de concordancia que evalúa el acuerdo entre dos métodos de medición del perímetro de cintura.
- Serie de casos con descripción clínica y metabólica de 15 pacientes jóvenes con obesidad severa y complicaciones endocrinas.

Su reto como lectora crítica es determinar:

- Validez interna: ¿los estudios miden lo que dicen medir?, ¿hay sesgos de selección o de información?, ¿se aplicaron protocolos de medición confiables?
- Precisión: ¿cómo se calcularon intervalos de confianza y medidas de concordancia?, ¿los análisis son apropiados para el diseño?
- Aplicabilidad: ¿los resultados se pueden extrapolar a la población de su hospital o de toda la ciudad?
- Coherencia entre estudios: ¿los hallazgos de los diferentes diseños son consistentes o contradictorios?

A lo largo del capítulo, el lector debe usar este caso como hilo conductor para aplicar herramientas de evaluación crítica a cada uno de los subtipos de estudios descriptivos, aprendiendo a identificar fortalezas, limitaciones y utilidad práctica de la evidencia presentada.

2. OBJETIVOS

Un estudio descriptivo busca:

- Medir la frecuencia de una condición (prevalencia, incidencia, tasas).
- Describir su distribución en la población (por edad, sexo, ubicación, tiempo).
- Identificar patrones o tendencias que ayuden a comprender el problema.
- Evaluar la concordancia o fiabilidad de las mediciones, cuando corresponde.
- Generar hipótesis para estudios analíticos posteriores, sin establecer causalidad.

3. INTRODUCCIÓN

Los estudios descriptivos constituyen la base de la investigación epidemiológica y clínica. Su propósito principal es mostrar cómo se distribuyen las enfermedades o condiciones de salud en una población, sin pretender establecer relaciones causales. Son, en esencia, una herramienta para contar y caracterizar: cuántos enferman, en qué grupos, en qué lugares y cómo cambia esa frecuencia con el tiempo.

En la práctica médica, estos estudios permiten al clínico reconocer la magnitud de los problemas que atiende en su consulta diaria. En la salud pública, ayudan a identificar prioridades, asignar recursos y diseñar políticas de prevención. Y en la investigación, son el primer peldaño para generar hipótesis que luego podrán probarse con diseños analíticos o experimentales. Sin estudios descriptivos, los problemas no se dimensionan ni se intervienen.El objetivo de este capítulo es que el lector adquiera la capacidad de leer críticamente un estudio descriptivo, reconociendo qué aporta, cuáles son sus limitaciones y cómo interpretar sus resultados con precisión y prudencia. Así, al enfrentarse a un artículo de este tipo, no se limitará a aceptar sus cifras, sino que sabrá analizar su validez, confiabilidad y aplicabilidad en el contexto clínico y poblacional en el que trabaja.

3.1 ESTUDIOS DE PREVALENCIA (O ESTUDIOS TRANSVERSALES)

El estudio de prevalencia o transversal es un diseño que mide la proporción de personas con una condición en un momento determinado. Se caracteriza por recolectar datos en un único punto

https://doi.org/10.62486/978-9915-9928-3-9.Ch10

temporal, evaluar simultáneamente variables de exposición y resultado, y ofrecer una fotografía de la magnitud de un problema de salud sin establecer causalidad. Se utiliza principalmente para estimar la frecuencia de enfermedades, describir su distribución por subgrupos y planificar recursos. Ejemplo: En el caso de la obesidad en La Paz, corresponde al estudio que mide la prevalencia en adultos de 18 a 39 años. El lector crítico debe comprobar que la población y el periodo de medición estén claramente definidos, que el muestreo sea representativo, que las mediciones de peso y talla se hayan hecho con protocolos estandarizados y que los intervalos de confianza se calculen con métodos apropiados. El tamaño muestral debe calcularse considerando prevalencia esperada, nivel de confianza y error aceptable. El intervalo de puntuación de Wilson (Wilson Score Interval) es uno de los métodos estadísticos más recomendados y apropiados para calcular intervalos de confianza (IC) en estudios de prevalencia.

3.2 ESTUDIOS ECOLÓGICOS

El estudio ecológico analiza datos agregados por grupos o regiones, en lugar de datos individuales, y se caracteriza por el uso frecuente de fuentes secundarias como censos o registros. Sus unidades de análisis son colectivas, lo que permite detectar patrones geográficos o temporales, pero conlleva el riesgo de falacia ecológica y efecto MAUP (Modifiable Areal UnitProblem, o Problema de la Unidad Areal Modificable) es un sesgo que ocurre cuando las conclusiones de un análisis varían significativamente según cómo se agrupen o delimiten geográficamente los datos. Y donde una asociación encontrada a nivel de grupo no necesariamente se aplica a nivel individual. La elección de la unidad geográfica es, por tanto, crucial para evitar conclusiones erróneas si se interpretan resultados grupales como aplicables a individuos. Se utiliza para observar tendencias, generar hipótesis y evaluar impactos de políticas a gran escala.

Ejemplo: En el caso, sería el estudio que compara tasas de obesidad por distrito y su relación con la densidad de tiendas de ultraprocesados. El lector crítico debe verificar que las unidades de análisis estén bien definidas, que las tasas se hayan ajustado por edad y sexo, y que se presenten medidas de incertidumbre, además de reconocer las limitaciones inherentes al diseño.

3.3 ESTUDIOS DE CONCORDANCIA

El estudio de concordancia evalúa el grado de acuerdo entre observadores o métodos de medición. Sus características incluyen la posibilidad de trabajar con variables nominales, ordinales o continuas, y el uso de índices específicos como kappa, kappa ponderado o coeficiente de correlación intraclase (ICC), según corresponda. Se utiliza para validar instrumentos, asegurar la consistencia de mediciones y comparar métodos diagnósticos. Son esenciales en investigación clínica para determinar la fiabilidad y la intercambiabilidad de métodos. No debe confundirse correlación con concordancia

Ejemplo: En el caso, sería el estudio que analiza la concordancia entre dos observadores que miden el perímetro de cintura en adultos jóvenes. En la lectura crítica es fundamental comprobar que el índice elegido sea el adecuado, que se reporten intervalos de confianza, que exista una descripción clara del protocolo de medición y de la capacitación

3.4 ESTUDIOS DE SERIES DE CASOS

La serie de casos describe las características clínicas, evolución y desenlaces de un grupo de pacientes con la misma condición. Se distingue por su naturaleza descriptiva, sin grupo de comparación, y por el énfasis en documentar detalles clínicos. Se utiliza para reconocer patrones, describir complicaciones o comportamientos inusuales y generar hipótesis para futuros estudios. En el caso de La Paz, sería el informe sobre 15 pacientes jóvenes con obesidad severa y complicaciones endocrinas, no permiten estimar prevalencias ni riesgos. El lector crítico debe comprobar que los criterios de inclusión estén bien definidos, que la selección de casos no esté sesgada y que las conclusiones no excedan lo que un diseño descriptivo permite.

Aunque son descriptivos, las series de casos comparten características únicas al documentar múltiples pacientes con la misma condición. Estos estudios proporcionan información detallada sobre los síntomas, tratamientos y desenlaces. Este tema ya se vio en el capítulo 8.

https://doi.org/10.62486/978-9915-9928-3-9.Ch10

3.5 ESTUDIOS LONGITUDINALES DESCRIPTIVOS

Por último, el estudio descriptivo longitudinal sigue a una población a lo largo del tiempo para describir cambios en la frecuencia o en las características de una condición. Recoge datos en múltiples momentos y, aunque no establece causalidad, permite observar tendencias. Se utiliza para monitorizar el impacto de políticas, observar la evolución de enfermedades y detectar cambios en patrones de salud.

Ejemplo: En el caso, sería un registro municipal que documenta la prevalencia de obesidad en adultos jóvenes de 2015 a 2024. El lector crítico debe evaluar si el seguimiento es consistente, si los métodos de medición se mantuvieron uniformes, si las pérdidas se informaron y analizaron, y si las variaciones observadas reflejan cambios reales y no artefactos metodológicos.Aunque son menos comunes, los estudios descriptivos longitudinales registran datos a lo largo del tiempo, describiendo cambios en la salud de una población.

4. METODOLOGÍA DE LOS ESTUDIOS DESCRIPTIVOS

La metodología es el corazón de todo estudio descriptivo. En la lectura crítica, el objetivo no es solo comprender cómo se diseñó la investigación, sino identificar si esos métodos son lo suficientemente sólidos como para confiar en los resultados. Al evaluar un artículo descriptivo, el lector debe fijarse en cinco aspectos esenciales: definición de la población, selección de la muestra, medición de variables, procedimientos de análisis y control de sesgos.

4.1 POBLACIÓN OBJETIVO.

El estudio debe dejar en claro a quién se quiere describir: adultos jóvenes de 18 a 39 años residentes en La Paz en 2024, por ejemplo. La lectura crítica debe comprobar si esa definición es precisa y si el marco de referencia que se utilizó (censo, registros municipales, padrones de salud) representa adecuadamente a toda la población. Cuando un estudio no especifica claramente su población, se abre la puerta a conclusiones ambiguas o inexactas.

4.2 MUESTRA.

Un artículo de calidad debe describir cómo se seleccionaron los participantes y justificar el tamaño muestral. En un estudio transversal, por ejemplo, el lector debe preguntarse si el muestreo fue aleatorio y representativo o si se limitó a conveniencia (como pacientes de un hospital), lo cual puede generar sesgo de selección. También debe evaluar si el tamaño calculado fue suficiente para estimar la prevalencia con la precisión requerida y si se consideraron factores como el efecto de diseño en muestreos por conglomerados.

4.3 MEDICIÓN DE VARIABLES.

El artículo debe explicar cómo se definieron y registraron las variables principales. En el caso de la obesidad, ¿el peso y la talla se midieron con balanzas y tallímetros calibrados o se obtuvieron de autorreporte? ¿Se capacitó al personal para aplicar protocolos estandarizados? En estudios de concordancia, ¿se detalla el entrenamiento de los observadores y se utilizan índices adecuados como el kappa o el ICC? El lector crítico debe estar atento a posibles fuentes de error de medición y valorar si se implementaron estrategias para reducirlas.

4.4 PROCEDIMIENTOS DE ANÁLISIS.

En estudios de prevalencia, la expectativa es que se calculen proporciones con intervalos de confianza apropiados, que se estratifique por subgrupos relevantes y, cuando procede, que se ajusten las tasas por edad o sexo para facilitar comparaciones. En estudios ecológicos, el lector debe verificar si las tasas fueron estandarizadas y si se discuten las limitaciones de trabajar con datos agregados. En estudios de concordancia, es esencial revisar si se eligió el estadístico correcto y si se reportan intervalos de confianza, no solo valores puntuales. En todos los casos, el análisis debe ser coherente con los objetivos planteados y transparente en la presentación de resultados.

https://doi.org/10.62486/978-9915-9928-3-9.Ch10

4.5 CONTROL DE SESGOS Y LIMITACIONES.

Un buen artículo descriptivo no los oculta, sino que los reconoce y discute. El lector crítico debe evaluar si los autores identificaron el riesgo de sesgo de selección, la posibilidad de error de información, las pérdidas en estudios longitudinales o la interpretación indebida de resultados en estudios ecológicos. Cuando un artículo ignora estas debilidades, su credibilidad se reduce significativamente. Aplicado al caso base, un lector crítico no se conformaría solo con conocer la prevalencia de obesidad reportada, sino que preguntaría: ¿La muestra de adultos jóvenes proviene de toda la ciudad o solo de ciertos distritos? ¿Los datos se obtuvieron con mediciones directas o de cuestionarios? ¿Los intervalos de confianza están calculados de manera adecuada? ¿Se reconocen los sesgos inherentes al diseño? Este ejercicio es lo que diferencia la lectura pasiva de la lectura crítica, y lo que permite que la información se convierta en evidencia confiable para la práctica clínica y la salud pública.

5. VENTAJAS Y LIMITACIONES

5.1 VENTAJAS:

Los estudios descriptivos son la base de la epidemiología y suelen ser la primera aproximación para dimensionar un problema de salud. Su principal ventaja es la simplicidad: permiten obtener información útil con relativa rapidez y menor costo en comparación con estudios analíticos o experimentales. Gracias a ellos se puede estimar la prevalencia de una enfermedad, describir cómo se distribuye en la población y generar hipótesis para investigaciones posteriores.

Un lector crítico debe reconocer que esta sencillez no los hace triviales. En el caso de la obesidad en La Paz, un estudio transversal puede dar cifras claras sobre cuántos jóvenes de 18 a 39 años viven con exceso de peso, mientras que un estudio ecológico puede mostrar diferencias entre distritos y señalar áreas con mayor densidad de ultraprocesados. La fortaleza de estos diseños está en documentar la magnitud y el contexto del problema, lo cual resulta esencial para la planificación de recursos y políticas. Asimismo, los estudios de concordancia permiten validar herramientas y asegurar la calidad de las mediciones, y las series de casos aportan información clínica detallada que puede alertar sobre complicaciones emergentes.

5.2 LIMITACIONES:

Sus limitaciones deben analizarse con la misma atención. La más evidente es que no establecen causalidad: un estudio transversal puede mostrar que la obesidad es más frecuente en ciertos distritos, pero no puede concluir que la causa sean las tiendas de ultraprocesados; del mismo modo, una serie de casos no permite generalizar sus hallazgos al conjunto de la población. Otra limitación importante es la susceptibilidad a sesgos. El sesgo de selección aparece cuando la muestra no representa a la población general, como ocurriría si solo se incluyeran pacientes hospitalizados. El sesgo de información surge cuando las mediciones son inexactas, por ejemplo, si el peso se obtiene por autorreporte en lugar de medición directa.

Los estudios ecológicos, además, enfrentan el riesgo de falacia ecológica, es decir, interpretar indebidamente resultados grupales como si fueran aplicables a nivel individual. Los longitudinales descriptivos, aunque útiles para mostrar tendencias, pueden verse afectados por cambios en la metodología de recolección de datos a lo largo del tiempo, lo que puede generar artefactos en la interpretación de la evolución de una enfermedad.El lector crítico debe habituarse a preguntarse, ante cualquier artículo descriptivo: ¿Las ventajas de este diseño están bien aprovechadas? ¿Las limitaciones se reconocen y se discuten de forma transparente? En el caso de La Paz, un estudio que reporte la prevalencia de obesidad en jóvenes solo será útil si aclara cómo se seleccionó la muestra, cómo se midió la variable y qué sesgos pueden afectar los resultados. Un artículo que omita estas aclaraciones no debería considerarse confiable para orientar decisiones clínicas o de salud pública.

En síntesis, los estudios descriptivos son herramientas valiosas para dimensionar y describir, pero el lector crítico debe recordar siempre que su utilidad radica en informar y orientar, no en explicar causas. Su fortaleza es ser el primer paso; su debilidad, que por sí solos nunca bastan para responder preguntas de causalidad.

Ejemplo. Un estudio transversal podría indicar que la prevalencia de obesidad en jóvenes es del 25%, pero no explicará por qué. Un estudio ecológico mostraría que los distritos con más

https://doi.org/10.62486/978-9915-9928-3-9.Ch10

ultraprocesados tienen mayores tasas, aunque sin probar que esa sea la causa. Un estudio de concordancia serviría para comprobar si dos observadores miden de manera consistente el perímetro de cintura. Una serie de casos podría describir complicaciones endocrinas en jóvenes con obesidad severa. Finalmente, un estudio longitudinal permitiría ver si la prevalencia aumentó de 2015 a 2024.

Tabla 1: Análisis crítico: ventajas y limitaciones de los principales diseños descriptivos

Tipo de estudio descriptivo	Ventajas principales	Limitaciones principales
Transversal (prevalencia)	- Estima la magnitud de un problema en un momento específico.	- No establece causalidad.
	- Rápido y relativamente económico.	- Vulnerable al sesgo de selección (muestra no representativa).
	- Útil para identificar subgrupos de mayor riesgo.	- Puede haber sesgo de información (mediciones autorreportadas).
	- Genera hipótesis iniciales.	- No distingue causa y efecto (temporalidad).
Ecológico	- Analiza patrones geográficos y temporales.	- Riesgo de falacia ecológica.
	- Útil para comparar regiones o evaluar impacto de políticas.	- Dependencia de calidad de datos secundarios.
	- Genera hipótesis a gran escala.	- Dificultad para controlar variables de confusión.
Concordancia	- Evalúa la fiabilidad de métodos y observadores.	- Resultados dependen del estadístico elegido (kappa, ICC, Bland–Altman).
	- Permite estandarizar protocolos y mejorar calidad de datos.	- Tamaños muestrales pequeños limitan precisión.
	- Fundamental en estudios diagnósticos.	- Requiere protocolos de medición bien definidos.
Serie de casos	- Describe en detalle características clínicas de pacientes.	- No permite estimar frecuencia poblacional.
	- Útil para condiciones poco frecuentes o emergentes.	- Alto riesgo de sesgo de selección.
	- Genera hipótesis clínicas.	- No incluye grupo comparativo.
		- Limitada generalización.
Longitudinal descriptivo	- Permite observar tendencias y evolución de enfermedades.	- Puede verse afectado por cambios en la calidad de los datos o métodos de medición.
	- Útil para monitorear impacto de políticas o intervenciones.	- Riesgo de pérdidas de seguimiento.
	- Describe cambios en salud pública a lo largo del tiempo.	- No establece causalidad.

Fuente: Elaboración propia

6. INTERPRETACIÓN ESTADÍSTICA

En los estudios descriptivos, la interpretación estadística se centra en medidas de frecuencia (como prevalencia e incidencia), en resúmenes numéricos (como medias y proporciones) y en la precisión de estas estimaciones (intervalos de confianza). El papel del lector crítico no es calcular, sino saber leer e interpretar lo que ya está publicado, distinguiendo resultados sólidos de estimaciones poco confiables.

Tabla 2. Ejemplos de medidas estadísticas en estudios descriptivos aplicadas al caso de obesidad en jóvenes de La Paz

Medida	Ejemplo aplicado (La Paz)
Prevalencia	162 de 540 jóvenes con obesidad → 30%
Incidencia acumulada	200 casos nuevos en 5.000 jóvenes en 2 años → 4% en 2 años
Proporción	300 de 1.000 jóvenes con cintura elevada → 30%
Intervalo de confianza (IC95%)	Prevalencia = 30% (IC95%: 26%–34%)

Nota: Los datos presentados son ejemplos ficticios utilizados únicamente con fines didácticos.
Fuente: Elaboración propia

https://doi.org/10.62486/978-9915-9928-3-9.Ch10

6.1 PREVALENCIA.

La prevalencia es la proporción de personas con una condición en un momento específico. Dicho de forma simple: es el "fotograma" de cuántas personas tienen el problema de salud en el instante de la medición.

- En el caso (Tabla 2), si 162 de 540 jóvenes evaluados tienen obesidad, la prevalencia es 30%.

Todo artículo debe reportarla con un intervalo de confianza (IC), que es el rango donde probablemente se encuentra el valor real en la población (IC 95%: 26%–34%).

Lectura crítica: el lector debe preguntarse si el artículo explica cómo seleccionó la muestra, si la medición fue directa y estandarizada (balanza calibrada, protocolo definido) y si el IC se calculó con un método apropiado. Si solo se da el porcentaje "30%" sin IC, la estimación es incompleta y poco útil.

6.2 INCIDENCIA.

La incidencia mide cuántos casos nuevos aparecen en un periodo entre personas que al inicio no tenían la enfermedad. Mientras que la prevalencia es una foto, la incidencia es como un "video" que muestra la aparición de casos en el tiempo.

- En nuestro caso (tabla 2): Si en 5.000 jóvenes sin obesidad al inicio se detectan 200 casos nuevos en dos años, la incidencia acumulada es del 4**%** en 2 años.

Cuando se considera el tiempo de seguimiento (persona-años), se calcula una tasa de incidencia.

Lectura crítica: el lector debe verificar si el artículo definió claramente la población en riesgo y si explica cómo manejó pérdidas de seguimiento. Un estudio que solo diga "hubo 200 casos nuevos" sin aclarar el denominador o el tiempo observado no permite interpretar correctamente la magnitud del problema.

6.3 PROPORCIONES Y MEDIDAS DE RESUMEN

Las proporciones expresan cuántas personas de un grupo tienen cierta característica.

- Ejemplo (tabla 2): si 300 de 1.000 jóvenes presentan cintura elevada, la proporción es 30%.

Las medidas de resumen como la media, mediana y moda describen cómo se distribuyen los valores de una variable continua. En el caso de la obesidad, la media de IMC puede dar una idea general, pero la mediana es más robusta si hay valores extremos.

Lectura crítica: el lector debe revisar si el artículo presenta los datos con medidas de dispersión (desviación estándar, percentiles, rango intercuartílico) y no solo con una media aislada, que puede ocultar una distribución muy desigual.

6.4 INTERVALOS DE CONFIANZA

Un intervalo de confianza es un rango de valores que refleja la precisión de la estimación. Cuanto más estrecho el IC, mayor precisión (generalmente por mayor tamaño de muestra). Cuanto más amplio, menor certeza y más cautela al interpretar.

- Si la prevalencia es 30% con IC95% de 26%–34% (tabla 2), significa que el valor real de la población probablemente esté dentro de ese rango.

Lectura crítica: el lector debe desconfiar de artículos que presentan porcentajes sin IC, o con IC imposibles (por ejemplo, −2% a 35% de prevalencia). También debe verificar si se menciona cómo se calculó el IC. El método de Wilson o el exacto de Clopper–Pearson suelen ser más confiables que el método clásico de Wald.

6.5 MEDICIÓN IMPERFECTA Y AJUSTES

Cuando la condición se identifica con una prueba imperfecta (por ejemplo, un cuestionario en lugar de medición directa), la prevalencia observada puede estar sesgada. Existen fórmulas que ajustan el resultado considerando la sensibilidad (probabilidad de detectar un verdadero positivo) y la especificidad (probabilidad de detectar un verdadero negativo).

Lectura crítica: si el artículo usó cuestionarios de autorreporte para el peso y talla, debería

https://doi.org/10.62486/978-9915-9928-3-9.Ch10

al menos discutir que esto puede subestimar la prevalencia real de obesidad. Un buen artículo incluso presentará la prevalencia ajustada teniendo en cuenta la imperfección de la medición.

6.6 CONCORDANCIA

La concordancia mide cuánto coinciden dos observadores o métodos al evaluar lo mismo.

- Para variables categóricas (ej. "cintura elevada: sí/no"), se usa el kappa.
- Para variables continuas (ej. perímetro de cintura en cm), se usa el coeficiente de correlación intraclase (ICC) o el método de Bland–Altman.

Tabla 3. Ejemplo de interpretación crítica de un estudio de concordancia en mediciones de perímetro de cintura

Medida de concordancia	Resultado	Interpretación crítica
Coeficiente de correlación intraclase (ICC, IC95%)	ICC = 0,94 (IC95%: 0,90–0,96)	El acuerdo entre observadores es excelente. El artículo debe especificar el modelo de ICC utilizado y el protocolo de medición.
Bland–Altman (sesgo y límites de acuerdo)	Sesgo = −0,4 cm; LOA = ±3,0 cm	Aunque el sesgo promedio es bajo, los observadores pueden diferir hasta 3 cm. El lector crítico debe juzgar si esta variación es clínicamente aceptable para la clasificación de riesgo.

Nota: Datos ficticios presentados con fines ilustrativos. LOA = limitsofagreement (límites de acuerdo).

Lectura crítica: un artículo de calidad siempre reporta el índice de concordancia con su IC95% y explica el protocolo de medición. Por ejemplo, si dos observadores tienen un ICC de 0,94, el acuerdo es excelente. Pero si los límites de acuerdo muestran que sus mediciones pueden variar hasta ±3 cm, el lector debe juzgar si esa diferencia es clínicamente aceptable para decidir si un paciente está en riesgo.

Ejemplo para el caso:

- Un buen estudio transversal de obesidad debe reportar la prevalencia con IC, describir el muestreo y detallar cómo midió peso y talla.
- Un estudio longitudinal debe dejar claro el denominador y el tiempo de observación.
- Una serie de proporciones debe acompañarse de medidas de dispersión.
- En estudios de concordancia, lo clave es que el índice esté bien elegido y explicado.

En cada situación, el lector crítico debe preguntarse: ¿La estimación es válida, precisa y aplicable? Si la respuesta es no, la conclusión del artículo pierde fuerza.Los artículos descriptivos suelen apoyarse en tablas y gráficos para resumir resultados. El lector crítico debe aprender a leerlos más allá de los números, detectando si están bien construidos y si los datos están completos.

6.7 TABLAS Y GRÁFICAS

TABLAS DE PREVALENCIA Y PROPORCIONES:

La tabla clásica de un estudio transversal suele mostrar la prevalencia de la condición por grupos de edad, sexo o región. Al leerla, el primer paso es identificar el denominador (número de personas evaluadas en cada subgrupo). Un error común es que los autores reporten porcentajes sin especificar el tamaño de la muestra, lo que impide juzgar la solidez del resultado. El lector debe revisar también si se presentan intervalos de confianza y si se usó el método adecuado para calcularlos.

Tabla 4. Prevalencia de obesidad en adultos jóvenes de La Paz, según sexo

Sexo	n/N	Prevalencia %	IC95%
Hombres	70/270	25,9	20,9 – 31,5
Mujeres	92/270	34,1	28,6 – 40,0
Total	162/540	30	26,3 – 34,0

Nota: Datos ficticios presentados con fines didácticos. n = número de casos observados; N = número total evaluado. IC95% = intervalo de confianza al 95%, calculado con método de Wilson.
Fuente: Elaboración propia.

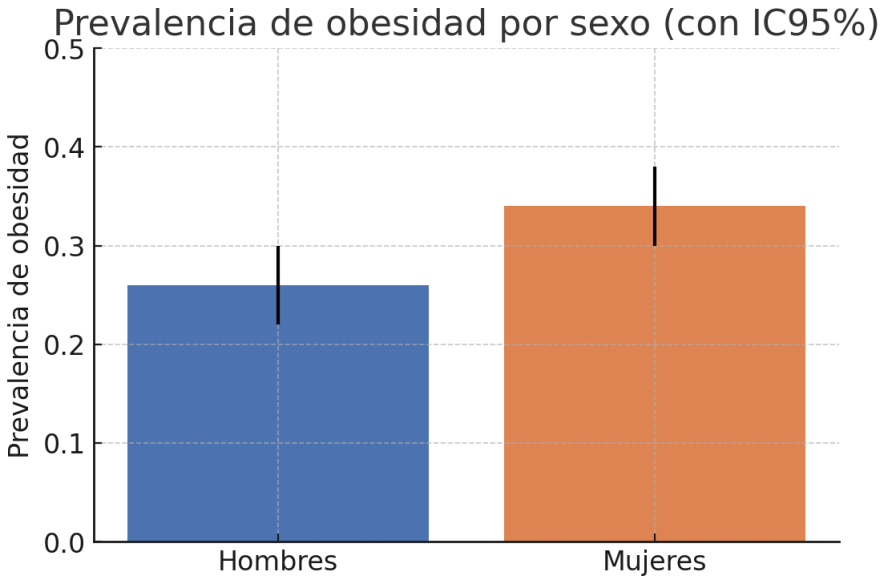
https://doi.org/10.62486/978-9915-9928-3-9.Ch10

Ejemplo (tabla 4): si un artículo sobre obesidad en La Paz reporta "prevalencia en mujeres = 34%", el lector debe buscar inmediatamente cuántas mujeres fueron evaluadas (¿fueron 50, 500 o 5.000?). Una prevalencia del 34% en 50 personas es mucho menos confiable que en 500.

GRÁFICOS DE BARRAS O COLUMNAS:

Se usan para comparar proporciones entre categorías (hombres vs. mujeres, distritos, grupos etarios). El lector crítico debe fijarse en que los ejes no estén manipulados (p. ej., truncar el eje Y en 20% exagera diferencias). Idealmente, las barras deben acompañarse de barras de error que representen intervalos de confianza. Ejemplo aplicado: un gráfico que muestre la obesidad en distritos de La Paz, con barras que van de 20% a 40%. El lector debe comprobar si las diferencias entre distritos realmente son estadística y clínicamente relevantes, o si los IC se superponen y por tanto no son tan distintas como parecen.

Figura 1 Prevalencia de obesidad en adultos jóvenes de La Paz, según sexo (IC95%)



Fuente: Elaboración propia

Este gráfico muestra la prevalencia de obesidad en hombres (26%) y mujeres (34%), con sus respectivos intervalos de confianza al 95% (rayas negras).

- La diferencia entre sexos parece clara (8 puntos porcentuales más en mujeres).
- Pero el lector debe fijarse en las barras de error: si se superponen parcialmente, la diferencia podría no ser tan sólida estadísticamente.
- El denominador es crucial: si la muestra de mujeres fue pequeña, el IC será más amplio y la precisión menor.
- El artículo que muestre este gráfico debería aclarar cuántas mujeres y hombres fueron incluidos y si el muestreo fue representativo de la ciudad. Si no lo hace, la comparación puede ser engañosa.

INCIDENCIA ACUMULADA:

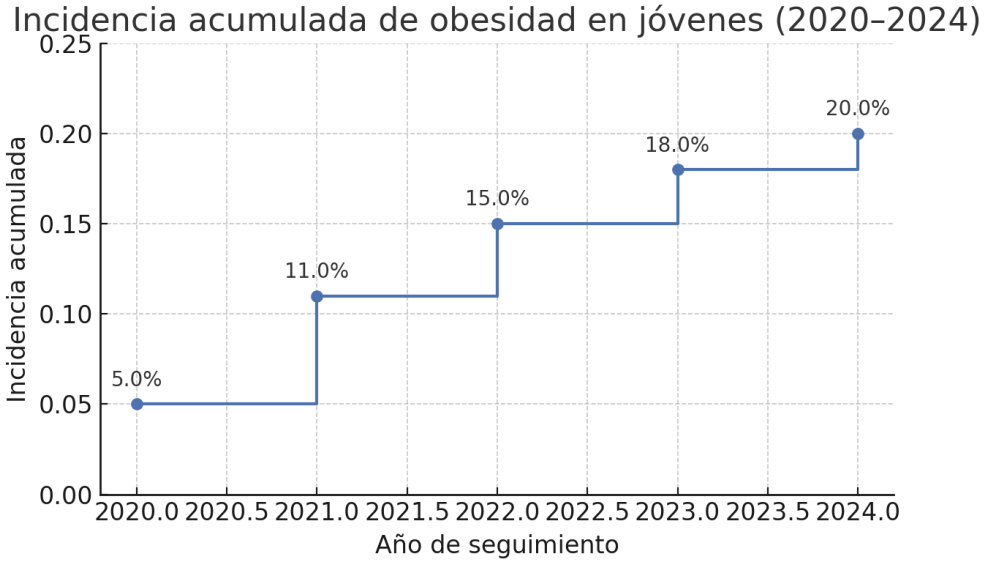
La incidencia acumulada representa la proporción de personas que desarrollan una condición nueva en un periodo de tiempo, dentro de una población inicialmente sana. Es, en esencia, la "probabilidad" de enfermar durante el seguimiento.

Ejemplo, partimos de 1.000 jóvenes sin obesidad en 2020. A lo largo de 5 años se registraron casos nuevos: 50 en 2020, 60 en 2021, 40 en 2022, 30 en 2023 y 20 en 2024.

- Al final de 2020, la incidencia acumulada es 5%.
- Al final de 2021, se eleva a 11% (50 + 60 casos / 1.000).
- Para 2022, llega a 15%, y así sucesivamente hasta alcanzar 20% en 2024.

https://doi.org/10.62486/978-9915-9928-3-9.Ch10

Figura 2 Incidencia acumulada de obesidad en adultos jóvenes de La Paz (2020–2024)



Fuente: Elaboración propia

Para interpretar la gráfica:

- La curva escalonada muestra cómo se va acumulando el riesgo a lo largo de los años.
- Cada punto indica el porcentaje de jóvenes que desarrollaron obesidad hasta ese momento.

El lector crítico debe verificar: Si el estudio define con claridad la población en riesgo al inicio. Si se aclara cómo se manejaron las pérdidas de seguimiento (quienes dejaron de participar). Si se especifica el periodo exacto de observación (año calendario, meses, etc.). Si la medición de obesidad fue consistente en todos los años (misma definición y método). Aplicado al caso: Un artículo que presente este gráfico para jóvenes en La Paz debe informar no solo la curva, sino también el número de participantes seguidos cada año y si hubo pérdidas significativas. Sin estos detalles, la incidencia acumulada puede sobreestimar o subestimar el riesgo real.

MAPAS (ESTUDIOS ECOLÓGICOS):

Los mapas de calor o coropléticos representan tasas o prevalencias por áreas geográficas. El lector debe preguntar: ¿se ajustaron las tasas por edad?, ¿los colores reflejan diferencias reales o solo artefactos de escala?, ¿se muestran datos de incertidumbre? Un mapa muy vistoso puede ocultar una base de datos débil.

Ejemplo aplicado: un mapa que pinte en rojo los distritos con obesidad “alta” en La Paz puede llamar la atención, pero si no hubo ajuste por edad, es posible que un distrito con más adultos mayores aparezca como “más enfermo” solo por su estructura poblacional.

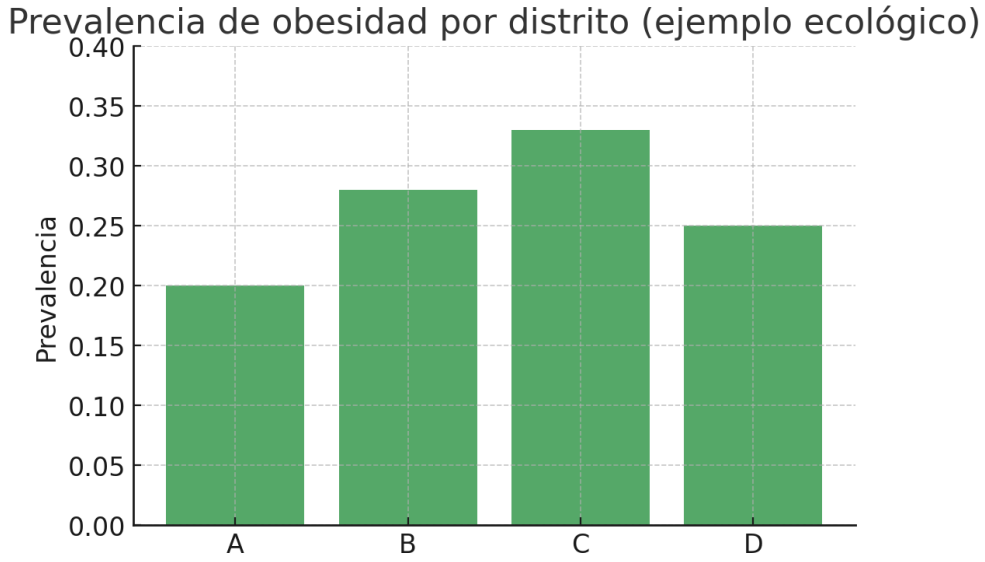
Aquí se comparan distritos ficticios (A–D), con prevalencias que van del 20% al 33%, cómo interpretarlo críticamente:

- El lector podría pensar que el distrito C tiene un problema mucho mayor (33% vs. 20%).
- Sin embargo, debe preguntarse: ¿se ajustaron las tasas por edad? Si un distrito tiene más población de adultos cercanos a los 40 años, la prevalencia puede ser mayor simplemente por estructura demográfica.
- Otro aspecto: ¿los IC de cada distrito están reportados? Si no aparecen, el gráfico puede exagerar diferencias que en realidad no son estadísticamente significativas.

Ejemplo aplicado: en un estudio de La Paz, un mapa o gráfico como este debe ir acompañado de aclaraciones sobre estandarización y márgenes de error; de lo contrario, la visualización es atractiva pero poco confiable.

https://doi.org/10.62486/978-9915-9928-3-9.Ch10

Figura 3 Prevalencia de obesidad en adultos jóvenes de La Paz, según distrito (ejemplo ecológico)



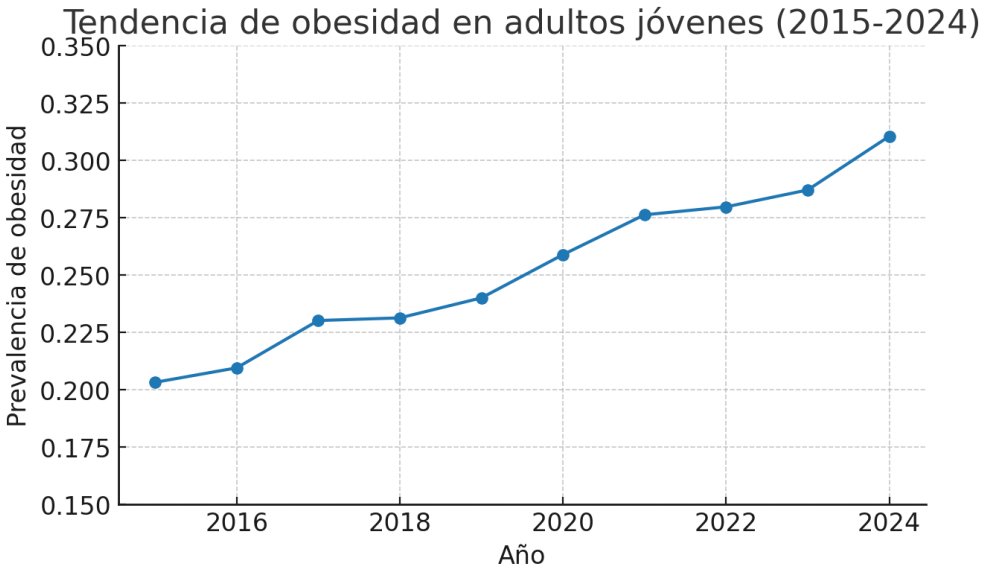
Fuente: Elaboración propia

SERIES TEMPORALES Y TENDENCIAS:

En estudios longitudinales descriptivos se suelen mostrar curvas que representan cómo cambia una condición con el tiempo. El lector debe revisar que el eje X indique claramente los años o periodos, que el eje Y esté correctamente escalado y que se especifique si los datos fueron recolectados con métodos consistentes en todos los años. Cambiar la forma de medir en medio de la serie puede generar un "salto" artificial en la curva.

Ejemplo aplicado: si se muestra que la prevalencia de obesidad en La Paz subió de 20% en 2015 a 30% en 2024, el lector crítico debe verificar si en todo el periodo se usaron los mismos criterios diagnósticos y las mismas técnicas de medición. De lo contrario, la tendencia podría deberse a un cambio metodológico y no a un aumento real.

Figura 4 Tendencia de la prevalencia de obesidad en adultos jóvenes de La Paz (2015–2024)



Fuente: Elaboración propia

https://doi.org/10.62486/978-9915-9928-3-9.Ch10

La curva muestra un aumento gradual de obesidad en jóvenes, de alrededor del 20% en 2015 hasta el 30% en 2024.Para interpretarlo críticamente:

- A primera vista, parece un incremento real y preocupante.
- El lector debe verificar si el método de medición fue consistente en todos los años. Si en 2020 cambiaron la definición de obesidad o el equipo de medición, la curva podría mostrar un "salto artificial".
- También es clave mirar si se indican intervalos de confianza o bandas de incertidumbre alrededor de la línea. Sin ellas, la tendencia puede ser sobreinterpretada.

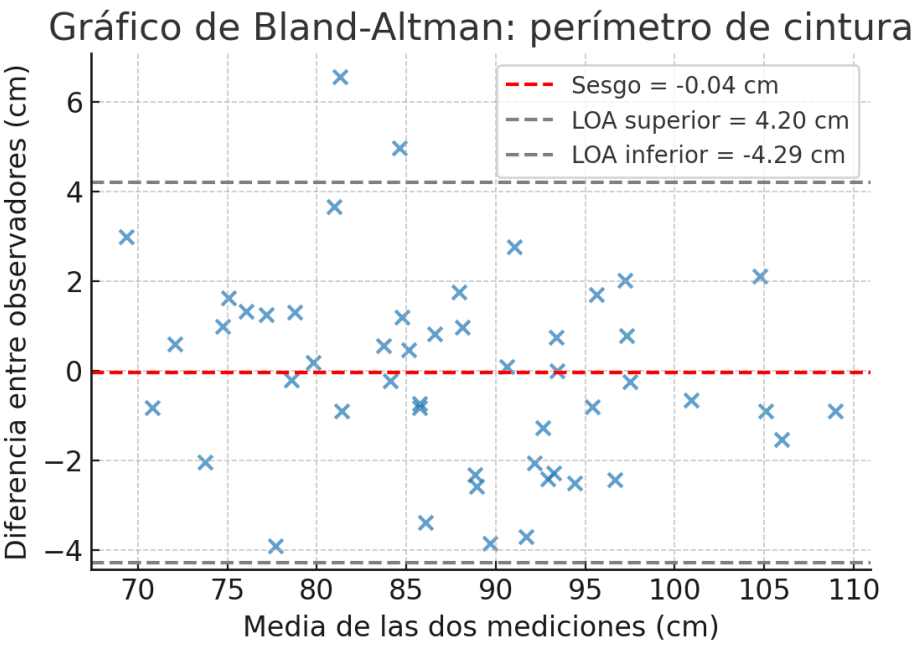
Ejemplo aplicado: un artículo que muestre este gráfico debería informar si cada año se midió con la misma balanza, con personal entrenado y con idéntico punto de corte de IMC. Si no, el aparente aumento puede deberse a un cambio metodológico y no a un fenómeno real.

GRÁFICOS DE CONCORDANCIA (BLAND–ALTMAN):

Cuando se comparan métodos o evaluadores, el gráfico de Bland–Altman es clásico: coloca en el eje X la media de las dos mediciones y en el eje Y la diferencia entre ellas. El lector debe fijarse en el sesgo promedio (qué tan sistemáticamente se diferencian los observadores) y en los límites de acuerdo (hasta cuánto puede variar la medición entre ellos). Si esos límites son muy amplios respecto al punto de corte clínico, la concordancia no es aceptable aunque el ICC sea "alto".

Ejemplo aplicado: en el caso del perímetro de cintura, un gráfico que muestre un sesgo de −0,4 cm y límites de acuerdo de ±2,8 cm puede considerarse aceptable si la decisión clínica depende de un umbral de 5 cm, pero problemático si el umbral es estrecho.

Figura 5 Gráfico de Bland–Altman para la medición del perímetro de cintura por dos observadores



Fuente: Elaboración propia

Cada punto representa la diferencia entre dos observadores midiendo el mismo perímetro de cintura. La línea roja marca el sesgo promedio (−0,4 cm) y las grises los límites de acuerdo (LOA), que en este caso son aproximadamente ±2,8 cm.

Para interpretarlo críticamente:

- El sesgo cercano a cero indica que los observadores, en promedio, miden casi lo mismo.
- Pero lo importante son los LOA: significan que una medición puede diferir hasta 3 cm entre observadores.
- El lector crítico debe preguntarse: ¿esta variabilidad es clínicamente aceptable? Si el punto de corte de riesgo es 90 cm, una diferencia de ±3 cm podría hacer que un mismo paciente sea

https://doi.org/10.62486/978-9915-9928-3-9.Ch10

clasificado como “riesgoso” o “no riesgoso” según quién lo mida.

Tabla 5: Medidas estadísticas en estudios descriptivos: definición, interpretación y lectura crítica

Medida	Qué es	Cómo se interpreta	Qué revisar críticamente	Ejemplo aplicado (obesidad en La Paz)
Prevalencia	Proporción de personas con una condición en un momento determinado.	Representa un “fotograma” de cuántos tienen la condición en ese instante.	Población y periodo definidos; muestreo representativo; mediciones directas; IC95% reportado (idealmente Wilson o exacto).	162 de 540 jóvenes con obesidad → prevalencia 30% (IC95%: 26–34%).
Incidencia	Proporción o tasa de casos nuevos en un periodo.	Mide el riesgo de desarrollar la enfermedad en el tiempo (“video”).	Denominador claro; definición de población en riesgo; control de pérdidas; cálculo de persona-tiempo si es tasa.	200 casos nuevos en 5.000 jóvenes en 2 años → incidencia acumulada 4%.
Tasas	Frecuencia de eventos en relación con persona-tiempo observado.	Expresan velocidad de aparición de eventos.	Correcto cálculo de persona-tiempo; consistencia en seguimiento; pérdidas informadas.	200 casos nuevos / 9.500 persona-año → tasa = 21 por 1.000 persona-año.
Proporciones	Relación entre parte y todo (en %).	Expresan fracciones simples de la población.	Denominadores correctos; IC reportado; claridad sobre quién entra en el numerador.	300 de 1.000 jóvenes con cintura elevada → proporción = 30%.
Medidas de resumen (media, mediana, moda)	Valores que describen la distribución de variables continuas.	Media refleja promedio, mediana el valor central, moda el más frecuente.	Acompañar con dispersión (DE, IQR); evaluar asimetría; evitar reportar solo media.	Media de IMC = 28,4 kg/m²; mediana = 27,5 kg/m²; moda = 26,8 kg/m².
Intervalo de confianza (IC)	Rango que muestra la precisión de una estimación.	Indica hasta dónde puede variar el valor real en la población.	Método de cálculo explicado; IC coherente con el tamaño de muestra; no usar solo valores puntuales.	Prevalencia 30% (IC95%: 26–34%). IC estrecho = más preciso.
Concordancia (kappa, ICC, Bland–Altman)	Grado de acuerdo entre observadores o métodos.	Mide fiabilidad: kappa (categóricas), ICC (continuas), Bland–Altman (métodos).	Uso del índice adecuado; IC95% reportado; protocolos claros; muestra suficiente.	ICC=0,94 para medición de cintura (IC95% 0,90–0,96); Bland–Altman: sesgo −0,4 cm, LOA ±2,8 cm.

Fuente: Elaboración propia

7. SESGOS Y LIMITACIONES

Los estudios descriptivos, aunque fundamentales, son especialmente vulnerables a errores metodológicos que distorsionan la realidad. El lector crítico debe entrenarse para identificar estos sesgos y juzgar qué tan serios son, ya que de ellos depende la validez y aplicabilidad de los hallazgos.

7.1 EN ESTUDIOS TRANSVERSALES (PREVALENCIA)

El sesgo más común es el sesgo de selección: ocurre cuando la muestra no representa a toda la población objetivo. En el caso de la obesidad en La Paz, un estudio que incluya solo pacientes atendidos en hospitales probablemente sobrestime la prevalencia, pues estas personas tienden a tener más problemas de salud que la población general. Otro problema es el sesgo de información, que aparece cuando las mediciones no son precisas. Si el peso y la talla se obtienen por autorreporte, muchos participantes tenderán a subestimar su peso y sobrestimar su altura, reduciendo artificialmente la prevalencia calculada. También debe vigilarse la falta de temporalidad: la prevalencia muestra un “fotograma” y no permite saber si la exposición precedió al desenlace.

7.2 EN ESTUDIOS ECOLÓGICOS

El principal riesgo es la falacia ecológica, que consiste en interpretar indebidamente relaciones grupales como si fueran válidas a nivel individual. Si en distritos con más tiendas de

https://doi.org/10.62486/978-9915-9928-3-9.Ch10

ultraprocesados hay mayor obesidad, esto no significa que cada persona con obesidad compre en esas tiendas. Además, este diseño depende de datos secundarios cuya calidad varía; censos incompletos o registros mal actualizados pueden generar artefactos. También existe el problema del MAUP (Modifiable Areal UnitProblem): los resultados pueden cambiar según cómo se definan las fronteras geográficas o la escala de análisis (distritos vs. barrios).

7.3 EN ESTUDIOS DE CONCORDANCIA

El sesgo más frecuente surge de protocolos de medición poco estandarizados. Si los observadores no reciben la misma capacitación, las diferencias entre ellos reflejarán variabilidad técnica más que verdadera falta de acuerdo. Otra limitación es usar un estadístico inadecuado: reportar solo la correlación de Pearson entre dos mediciones de cintura no es correcto, porque mide asociación, no acuerdo. Además, cuando la prevalencia de una categoría es muy baja o muy alta, el índice kappa puede dar valores bajos incluso con alta concordancia aparente, lo que se conoce como la paradoja de la prevalencia. El lector crítico debe estar atento a si el artículo reconoce estas limitaciones.

7.4 EN SERIES DE CASOS

El mayor riesgo es el sesgo de selección: los casos reportados suelen provenir de un hospital específico o de pacientes más graves, lo que limita la generalización. En la serie de 15 pacientes jóvenes con obesidad severa, es probable que se hayan incluido solo los casos más complejos que llegaron al hospital, dejando fuera a muchos otros con formas más leves. También existe la tentación de extrapolar resultados descriptivos hacia conclusiones causales, lo cual no corresponde. Un artículo serio limitará sus afirmaciones al terreno de lo descriptivo.

7.5 EN ESTUDIOS LONGITUDINALES DESCRIPTIVOS

El principal problema es la pérdida de seguimiento. Si muchos participantes dejan de ser evaluados a lo largo de los años, y no se aclara cuántos ni por qué, las tasas observadas pueden distorsionarse. También es crítico el cambio de definiciones o métodos en el tiempo: si en 2015 se usó un punto de corte distinto de IMC para definir obesidad, la comparación con 2024 será inválida. Incluso mejoras tecnológicas, como cambiar de balanzas manuales a digitales, pueden introducir variaciones que aparenten tendencias inexistentes.

8. COMPARACIÓN CON OTROS ESTUDIOS

Una parte clave de la lectura crítica es entender que los resultados de un estudio descriptivo no deben interpretarse en aislamiento. El verdadero valor emerge cuando se comparan con otros trabajos que han evaluado el mismo problema, aunque usen metodologías diferentes. El lector debe preguntarse siempre: ¿son consistentes los hallazgos? ¿Si no lo son, se deben a diferencias reales en la población o a variaciones metodológicas?

8.1 EJEMPLO APLICADO AL CASO DE OBESIDAD EN LA PAZ

Un estudio transversal local puede reportar una prevalencia de obesidad del 30% en jóvenes de 18 a 39 años en 2024. Al compararlo con un estudio nacional publicado el año anterior, que estimó una prevalencia del 25% en ese mismo grupo etario, el lector debe analizar las diferencias. ¿Se deben a que la situación realmente empeoró en La Paz? ¿O a que el estudio nacional usó un muestreo más amplio pero incluyó también áreas rurales con menor obesidad? Aquí, la representatividad de la muestra y los métodos de estandarización se vuelven determinantes para interpretar la discrepancia.

Los estudios ecológicos también aportan a la comparación. Imaginemos que en un análisis de varios países andinos se reporta que Bolivia tiene una de las tasas más altas de obesidad urbana en mujeres jóvenes. Si nuestro estudio local encuentra el mismo patrón en La Paz, se refuerza la coherencia del hallazgo. Pero si no coincide, debemos revisar si las definiciones de obesidad fueron

https://doi.org/10.62486/978-9915-9928-3-9.Ch10

equivalentes (IMC ≥ 30 en ambos casos), si las fuentes de datos fueron comparables (encuestas poblacionales vs. registros hospitalarios) y si las fechas de recolección son similares.

En estudios de concordancia, la comparación también es esencial. Supongamos que un trabajo boliviano encuentra un coeficiente de correlación intraclase (ICC) de 0,94 para la medición de cintura, mientras que un estudio en Perú reporta 0,80 con el mismo protocolo. El lector debe analizar si la diferencia refleja variaciones reales en la capacitación del personal o si se debe a un tamaño muestral insuficiente en uno de los estudios.

Las series de casos son aún más contextuales. Una serie paceña puede describir complicaciones endocrinas frecuentes en jóvenes con obesidad severa, mientras que otra serie en México muestra predominio de complicaciones cardiovasculares. La discrepancia puede deberse a diferencias genéticas, ambientales o en el acceso a la atención, pero también puede ser producto de cómo se seleccionaron los pacientes y qué hospital reportó los casos.

Finalmente, los estudios longitudinales descriptivos permiten comparar tendencias. Si en La Paz la obesidad en jóvenes pasó de 20% en 2015 a 30% en 2024, pero un estudio similar en Santa Cruz reporta una tendencia estable en el mismo periodo, el lector debe cuestionar: ¿la diferencia refleja realidades epidemiológicas distintas o simplemente cambios metodológicos en la recolección de datos?

Clave de lectura critica: Cuando los resultados de distintos estudios son coherentes, fortalecen la confianza en los hallazgos. Cuando son discrepantes, el lector crítico debe identificar si las diferencias provienen de la población estudiada, la definición de la condición, el método de muestreo, la medición de las variables o el análisis estadístico. Solo así podrá decidir si las variaciones reflejan la realidad o si son artefactos metodológicos.

9. CHECKLIST DE EVALUACIÓN CRÍTICA

Este checklist ofrece preguntas clave que el lector crítico debe responder al revisar un artículo descriptivo. Esta lista de verificación estructurada y que no reemplaza el juicio crítico. Cada ítem proviene de recomendaciones de STROBE(StrengtheningtheReportingofObservationalStudies in Epidemiology o “Fortalecimiento de la comunicación de estudios observacionales en epidemiología”), JBI(Joanna Briggs Institute, enfocado en la validez del método) y RECORD(específico para estudios basados en registros médicos), adaptado a un nivel práctico y didáctico.

Tabla 6. Checklist de lectura crítica para estudios descriptivos

Categoría	Preguntas de lectura crítica
A. Validez del diseño	1. ¿Se especifica claramente la población objetivo (edad, sexo, lugar, periodo)?
	2. ¿Se describe de dónde se seleccionó la muestra y si representa a toda la población?
	3. ¿El muestreo fue probabilístico y transparente (aleatorio, estratificado, por conglomerados)?
	4. ¿El artículo justifica el tamaño muestral y considera pérdidas o efecto de diseño?
B. Medición de variables	5. ¿Se usaron criterios diagnósticos claros y reconocidos?
	6. ¿Las variables se midieron de forma estandarizada y con instrumentos validados?
	7. Si había más de un observador o método, ¿se reporta el grado de acuerdo (kappa, ICC, Bland–Altman) con intervalos de confianza?
C. Análisis estadístico	8. ¿Se presentan prevalencias, incidencias o tasas con intervalos de confianza apropiados (no solo porcentajes)?
	9. ¿Se aplicaron estandarizaciones (edad, sexo) o pesos de encuesta si el diseño lo requería?
	10. ¿El estudio explica cómo trató los valores perdidos?
D. Sesgos y limitaciones	11. ¿El estudio discute si la muestra puede diferir de la población general?
	12. ¿Reconoce posibles errores de medición o autorreporte?
	13. En estudios grupales, ¿los autores advierten la posibilidad de falacia ecológica?
	14. En estudios longitudinales, ¿se reporta cuántos abandonaron y si podrían sesgar los resultados?
E. Transparencia del reporte	15. ¿El artículo sigue las guías STROBE o JBI para estudios de prevalencia?
	16. ¿Los autores reconocen explícitamente las principales limitaciones?
	17. ¿El estudio discute si sus hallazgos son generalizables a la práctica clínica o a otras poblaciones?

Nota: Adaptado de guías STROBE, JBI y RECORD. Preguntas diseñadas para facilitar la lectura crítica de estudios descriptivos.

https://doi.org/10.62486/978-9915-9928-3-9.Ch10

Cómo usar el checklist:

- Cada respuesta puede ser Sí, No o Incierto.
- Un estudio con muchos "No" o "Incierto" tiene menor calidad y sus resultados deben interpretarse con cautela.
- No se trata de descartar automáticamente el artículo, sino de ponderar su valor como evidencia.

10. UTILIDAD

Los estudios descriptivos son útiles porque permiten dimensionar un problema de salud, mostrar cómo se distribuye en la población y servir de base para decisiones clínicas, de salud pública y generar hipótesis para investigaciones posteriores. Aunque no establecen relaciones causales, aportan una información inicial que orienta tanto la práctica médica como la planificación sanitaria.

En la práctica clínica, su utilidad radica en ofrecer al médico una idea clara de la frecuencia de las enfermedades que atiende. Conocer, por ejemplo, que la obesidad afecta a un 30% de los jóvenes en La Paz ayuda a entender por qué cada vez es más común encontrar complicaciones metabólicas en pacientes de consulta externa. Para el lector crítico, esto significa valorar si los datos de un estudio son aplicables a su contexto clínico y a su población de pacientes.

En salud pública, los estudios descriptivos permiten comparar regiones, monitorear tendencias y priorizar intervenciones. Un estudio ecológico que muestra mayor prevalencia de obesidad en distritos con más expendios de ultraprocesados no demuestra causalidad, pero sí alerta sobre un patrón que merece atención de políticas locales. La utilidad aquí es abrir la discusión y orientar recursos hacia los lugares de mayor necesidad.En el campo de la investigación, los descriptivos son el primer peldaño del método científico. Identifican problemas, describen características clínicas o poblacionales y generan hipótesis que luego podrán probarse con estudios analíticos o experimentales. Una serie de casos que documenta complicaciones endocrinas en jóvenes con obesidad severa puede motivar estudios de cohortes o ensayos clínicos enfocados en intervenciones tempranas.

La utilidad también está en la docencia y el entrenamiento del pensamiento crítico. Para internos y residentes, los estudios descriptivos son ideales como ejercicio inicial de lectura crítica, ya que exigen comprender conceptos básicos de muestreo, medición y estimación estadística sin la complejidad de modelos multivariados. Evaluar si un estudio de prevalencia está bien hecho es un excelente primer paso para desarrollar la mirada crítica frente a la literatura científica.

En síntesis, los estudios descriptivos son útiles porque ponen el problema sobre la mesa, lo cuantifican y lo hacen visible para clínicos, gestores e investigadores. El lector crítico debe aprovecharlos como herramienta de diagnóstico poblacional y como base para el razonamiento epidemiológico, siempre recordando que su papel es describir y generar preguntas, no responder causalidad.

https://doi.org/10.62486/978-9915-9928-3-9.Ch10

11. PREGUNTAS DE AUTOEVALUACIÓN

PREGUNTAS DE SELECCIÓN MÚLTIPLE

1. ¿Cuál es el objetivo principal de un estudio descriptivo?
a) Establecer causalidad entre exposición y desenlace
b) Medir la frecuencia y distribución de una condición en una población
c) Comparar la eficacia de dos tratamientos
d) Controlar variables de confusión
2. En un estudio transversal sobre obesidad en La Paz, ¿qué error metodológico limitaría más la validez de los resultados?
a) Usar balanzas calibradas para medir peso
b) Reportar la prevalencia con IC95%
c) Seleccionar la muestra solo en hospitales
d) Describir la edad y el sexo de los participantes
3. ¿Cuál es la principal limitación de un estudio ecológico?
a) No puede calcular prevalencia
b) Depende de registros secundarios
c) No permite inferir resultados individuales a partir de datos grupales
d) Tiene pérdidas de seguimiento elevadas
4. En un estudio de concordancia que evalúa medición de cintura, ¿qué indicador sería más apropiado para variables continuas?
a) Razón de momios (OR)
b) Kappa simple
c) ICC (coeficiente de correlación intraclase)
d) Valor p
5. Un estudio longitudinal descriptivo sobre obesidad en jóvenes muestra un aumento de 20% a 30% en 10 años. ¿Qué debería revisar el lector crítico antes de concluir que la tendencia es real?
a) Si el cálculo de odds ratio fue correcto
b) Si las definiciones y métodos de medición se mantuvieron constantes en el tiempo
c) Si los autores aplicaron un ensayo clínico aleatorizado
d) Si los distritos urbanos fueron excluidos del análisis

PREGUNTAS ABIERTAS

6. Explique la diferencia entre prevalencia e incidencia y dé un ejemplo aplicado al caso de obesidad en La Paz.
7. ¿Por qué los estudios ecológicos son útiles y al mismo tiempo peligrosos de interpretar?
8. ¿Qué aspectos debería revisar un lector crítico en la tabla de prevalencias por sexo de un estudio transversal?
9. Mencione dos ejemplos de sesgo en estudios descriptivos y explique cómo podrían afectar un estudio sobre obesidad.
10. ¿Cómo interpretaría un gráfico de Bland–Altman que muestre un sesgo promedio de −0,4 cm y límites de acuerdo de ±3 cm entre dos observadores que miden cintura?

RESPUESTAS

1. b) Estudios descriptivos: Se limitan a detallar la presencia, características, magnitud (prevalencia o incidencia) y distribución de un fenómeno (enfermedad, evento, factor de riesgo) según variables de tiempo, lugar y persona. No buscan causas, sino describir qué sucede.
2. c): Este error metodológico limitaría más la validez de los resultados por las siguientes razones: -Sesgo de Selección (o Sesgo de Berkson): Seleccionar la muestra únicamente en hospitales provoca que la prevalencia de obesidad no sea representativa de la población general de La Paz. Las personas en hospitales tienen mayor probabilidad de sufrir enfermedades crónicas asociadas a la obesidad, lo que resultaría en una sobreestimación de la prevalencia real de obesidad. Falta de Validez Externa: Los resultados no se pueden generalizar a la población sana o comunitaria, limitando la utilidad del estudio para la planificación de salud pública. Comparación: Las opciones A, B y D son acciones correctas o estándar en la realización de estudios transversales

https://doi.org/10.62486/978-9915-9928-3-9.Ch10

de prevalencia.

3. c): La principal desventaja de los estudios ecológicos (o de correlación) es que utilizan datos agregados o promedios de poblaciones (países, ciudades, escuelas) en lugar de datos de individuos

4. c): La medición de la cintura (en cm) es una variable numérica continua.

5. b): En los estudios longitudinales (que analizan tendencias a lo largo del tiempo), el mayor riesgo de sesgo es el "sesgo de medición" o cambio en las definiciones. Si las herramientas (básculas, tallímetros, instrumentos de encuesta) o las definiciones (ej. el punto de corte del IMC para sobrepeso) cambiaron entre el inicio y el final de los 10 años, el aumento observado podría ser artificial, debido a una mayor precisión en la medición o a criterios más amplios, en lugar de un aumento real en la obesidad

6. La prevalencia mide cuántas personas tienen obesidad en un momento específico (ejemplo: 30% de los jóvenes en 2024). La incidencia mide cuántos casos nuevos aparecen en un periodo (ejemplo: 200 nuevos casos en 5.000 jóvenes seguidos durante 2 años → incidencia acumulada 4%).

7. Son útiles para mostrar patrones geográficos o de políticas públicas, pero peligrosos porque se puede cometer falacia ecológica al interpretar los resultados grupales como si fueran individuales.

8. El denominador en cada grupo, la forma en que se midió la obesidad, si la muestra es representativa y si se reportan intervalos de confianza que indiquen la precisión de la estimación.

9. Sesgo de selección: si solo se encuestan pacientes hospitalizados, se sobreestimaría la obesidad. Sesgo de información: si se usan pesos autorreportados, se subestimaría la prevalencia.

10. Indica que en promedio los observadores coinciden casi perfectamente (sesgo cercano a 0), pero pueden diferir hasta 3 cm en sus mediciones. El lector crítico debe juzgar si esta variabilidad es clínicamente aceptable para clasificar a los pacientes en riesgo.

https://doi.org/10.62486/978-9915-9928-3-9.Ch10

REFERENCIAS

Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow C, Pocock SJ, et al.; STROBE Initiative. Strengthening the reporting of observational studies in epidemiology (STROBE): Explanation and elaboration. PLoS Med. 2007 Oct 16;4(10):e297. Available at: https://doi.org/10.1371/journal.pmed.0040297.
Munn Z, Moola S, Lisy K, Riitano D, Tufanaru C. Methodological guidance for systematic reviews of observational epidemiological studies reporting prevalence and cumulative incidence data. JBI Evid Synth. 2020 Oct;18(10):2117–26. Available at: https://doi.org/10.11124/JBISRIR-D-19-00007.
Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al.; RECORD Working Committee. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med. 2015 Oct;12(10):e1001885. Available at: https://doi.org/10.1371/journal.pmed.1001885.
Porta M, editor. A dictionary of epidemiology. 6th ed. Oxford: Oxford University Press; 2014.
Rothman KJ, Greenland S, Lash TL. Modern epidemiology. 4th ed. Philadelphia: Wolters Kluwer; 2021.
Friis RH, Sellers TA. Epidemiology for public health practice. 6th ed. Burlington: Jones & Bartlett Learning; 2020.
Naing L, Winn T, Rusli BN. Practical issues in calculating the sample size for prevalence studies. Arch Orofac Sci. 2006;1:9–14. Available at: https://www.cabdirect.org/cabdirect/abstract/20073130210.
Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016 Jun;15(2):155–63. Available at: https://doi.org/10.1016/j.jcm.2016.02.012.
Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986 Feb 8;1(8476):307–10. Available at: https://doi.org/10.1016/S0140-6736(86)90837-8.
World Health Organization. WHO STEPS surveillance manual: The WHO STEPwise approach to chronic disease risk factor surveillance. 3rd ed. Geneva: World Health Organization; 2017. Available at: https://www.who.int/ncds/surveillance/steps.

https://doi.org/10.62486/978-9915-9928-3-9.Ch10

Chapter 11 / Capítulo 11

Critical reading of scientific articles on health: A practical guide (Spanish Edition)
ISBN: 978-9915-9928-3-9
DOI: 10.62486/978-9915-9928-3-9.Ch11
Pages: 136-155

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Diagnostic Test Accuracy Studies

Estudios de prueba diagnóstica

Edgar Juan José Chávez Navarro
Nayra Condori-Villca
Marco Antonio Gumucio-Villarroel

1 Hospital de la Mujer. La Paz, Bolivia.
2 Hospital del Gran Chaco Fray Quebracho de III nivel. Tarija, Bolivia.
3 Hospital de Tercer Nivel Dr. Hernán Messuti Ribera. Pando, Bolivia.

ABSTRACT

Introduction: Diagnostic test studies are a fundamental tool in Evidence-Based Medicine, as they allow for the evaluation of a test's ability to correctly identify the presence or absence of a disease. Their purpose is to determine the validity, accuracy, and clinical utility of a diagnostic test by comparing it to a reference standard. Development: This chapter addresses the essential methodological principles for evaluating diagnostic tests, including appropriate patient selection, the use of reference standards, blinding, and bias control. It analyzes performance indicators such as sensitivity, specificity, predictive values, likelihood ratios, ROC curves, and measures of agreement. The importance of interpreting these parameters within the clinical, epidemiological, and healthcare context in which they are applied is emphasized. Conclusions: The critical evaluation of diagnostic tests goes beyond the simple interpretation of statistical indicators, requiring an assessment of their applicability, precision, and impact on clinical decision-making. A diagnostic test acquires true value when it contributes to improving patient care, optimizing resources, and supporting decisions based on high-quality scientific evidence.

Keywords: Sensitivity and Specificity; Diagnostic Tests; Likelihood Ratios; ROC Curve; Reproducibility of Results

RESUMEN

Introducción: Los estudios de prueba diagnóstica constituyen una herramienta fundamental en la Medicina Basada en la Evidencia, ya que permiten evaluar la capacidad de una prueba para identificar correctamente la presencia o ausencia de una enfermedad. Su propósito es determinar la validez, exactitud y utilidad clínica de una prueba diagnóstica mediante la comparación con un estándar de referencia. Desarrollo: El capítulo aborda los principios metodológicos esenciales para evaluar pruebas diagnósticas, incluyendo la selección adecuada de pacientes, el uso de estándares de referencia, el cegamiento y el control de sesgos. De esta manera, analiza indicadores de desempeño como sensibilidad, especificidad, valores predictivos, razones de verosimilitud, curvas ROC y medidas de concordancia. Se enfatiza la importancia de interpretar estos parámetros dentro del contexto clínico, epidemiológico y asistencial donde se aplican. Conclusiones: La evaluación crítica de las pruebas diagnósticas trasciende la simple interpretación de indicadores estadísticos, requiriendo valorar su aplicabilidad, precisión y repercusión en la toma de decisiones clínicas. Una prueba diagnóstica adquiere verdadero valor cuando contribuye a mejorar la atención del paciente, optimizar recursos y respaldar decisiones fundamentadas en evidencia científica de calidad.

Palabras clave: Sensibilidad y Especificidad; Pruebas Diagnósticas; Razones de Verosimilitud; Curva ROC; Reproducibilidad de Resultados

1. CASO CLÍNICO

En un hospital municipal de II nivel de La Paz, acude Juan**,** paciente masculino de 34 años, sin enfermedades previas, que consulta por cuadro clínico de 3 semanas con presencia de accesos de tos con expectoración, pérdida de peso de 4 kg y sudoración nocturna. En la radiografía de tórax se observa un infiltrado en el lóbulo superior derecho con posible cavitación. Su primo tuvo tuberculosis pulmonar (TB) hace menos de un año. El médico de guardia debe tomar una decisión inmediata: ¿aislar al paciente Juan e iniciar tratamiento antibacilar empírico o esperar confirmación diagnóstica?El hospital cuenta con disponibilidad de las siguientes pruebas:

- GenXpert MTB/RIF Ultra: resultado en pocas horas (prueba índice).
- Cultivo de micobacterias: considerado el estándar de referencia, pero demora semanas.
- Baciloscopia: disponible, pero menos sensible.

Entonces el médico se plantea la pregunta PICO: En adultos con sospecha clínica de TB pulmonar (Paciente), ¿la GenXpert MTB/RIF Ultra en esputo (Prueba índice), comparada con el cultivo micobacteriano (Estándar de referencia), permite confirmar o descartar tuberculosis (Outcome) con la validez y utilidad clínica necesarias para decidir aislamiento y tratamiento inmediato en nuestro contexto?

Encuentra un estudio (simulado) con 300 pacientes consecutivos con sospecha de TB:

TABLA 1: Matriz 2×2 de resultados simulados de la prueba Xpert MTB/RIF Ultra frente al cultivo de Mycobacterium tuberculosis en pacientes con sospecha clínica de tuberculosis pulmonar

	Cultivo + (enfermo)	Cultivo – (no enfermo)	Total
Xpert positivo	90 (VP)	20 (FP)	110
Xpert negativo	10 (FN)	180 (VN)	190
Total	100	200	300

Fuente: Elaboración propia

Los resultados mencionaron: Sensibilidad (Se): 90 %, Especificidad (Sp): 90 %, VPP: 81,8 %, VPN: 94,7 %, LR+: 9,0, LR–: 0,11; Además, se registraron 10 resultados indeterminados, reportados aparte.

2. OBJETIVOS

Un estudio de prueba diagnóstica tiene como objetivo central determinar si una prueba índice (la que queremos evaluar) es válida y útil para detectar, confirmar o descartar una enfermedad, en comparación con un estándar de referencia (Gold standard).

Es importante recalcar que las pruebas diagnósticas son el complemento a la clínica y no son el fundamento del diagnóstico. Tenemos que tomarlas como un aditamento tecnológico que nos ayuda a incrementar nuestra capacidad de percepción sobre lo que sucede con cada paciente y nos aproxima a la realidad.En otras palabras: saber qué tan bien la prueba discrimina entre enfermos y no enfermos, y si esa información es aplicable para tomar decisiones clínicas en un contexto real. De manera más concreta, sus objetivos son:

- Determinar la validez de la prueba: Estimar sensibilidad y especificidad. Estos indicadores muestran la capacidad discriminativa básica de la prueba.
- Calcular el valor predictivo en la práctica clínica: Estimar el valor predictivo positivo (VPP) y el valor predictivo negativo (VPN), que dependen no solo de la validez intrínseca de la prueba sino también de la prevalencia de la enfermedad.
- Cuantificar la utilidad de la prueba para modificar probabilidades clínicas: Medir razones de verosimilitud (LR+ y LR–) y, a partir de ellas, calcular probabilidades post-prueba con la ayuda de herramientas como el nomograma de Fagan. Permite saber si la prueba cambia realmente la toma de decisiones (iniciar tratamiento, aislar, descartar enfermedad).
- Definir la capacidad global de discriminación: Evaluar la curva ROC y el área bajo la curva (AUC) para valorar el desempeño de la prueba en distintos puntos de corte. Seleccionar el umbral diagnóstico óptimo considerando el equilibrio entre falsos positivos y falsos negativos.
- Evaluar la aplicabilidad y utilidad clínica: Valorar si los resultados obtenidos en el estudio son extrapolables al contexto local (nivel de atención, prevalencia, disponibilidad de recursos). Considerar además aspectos de rapidez, costo, accesibilidad y factibilidad de uso rutinario.

https://doi.org/10.62486/978-9915-9928-3-9.Ch11

3. INTRODUCCIÓN

En medicina, gran parte de las decisiones clínicas comienzan con una sospecha diagnóstica. Ante un paciente con síntomas y signos, el médico necesita apoyarse en pruebas complementarias (laboratorio, imágenes, pruebas rápidas, biomarcadores) para confirmar o descartar enfermedades. Sin embargo, no todas las pruebas que utilizamos tienen la misma capacidad de diferenciar entre enfermos y sanos.

Un estudio de prueba diagnóstica es el diseño de investigación que permite medir y validar esa capacidad discriminativa. Su esencia consiste en comparar la prueba que queremos evaluar (llamada prueba índice) con la mejor referencia disponible (llamada estándar de referencia o *gold standard*). De esta forma, podemos conocer con qué grado de certeza la prueba índice identifica correctamente a quienes tienen la enfermedad (verdaderos positivos) y a quienes no la tienen (verdaderos negativos).

Importancia en la práctica clínica:

- Para el médico general: le ayuda a comprender por qué una prueba "negativa" no siempre descarta la enfermedad, y a interpretar correctamente lo que significan sensibilidad y especificidad.
- Para el residente: le enseña a evaluar críticamente si el estudio que valida una nueva prueba fue bien diseñado, evitando sesgos y errores metodológicos.
- Para el especialista: le permite decidir si una prueba es realmente útil en su práctica diaria, considerando el impacto en costos, rapidez, toma de decisiones y salud pública.

En países como Bolivia y en Latinoamérica en general, interpretar críticamente estudios de pruebas diagnósticas es esencial porque muchas veces se usan pruebas de menor costo o acceso rápido, aunque con exactitud limitada, la prevalencia de enfermedades como tuberculosis, dengue o Chagas es distinta a la de países desarrollados, lo que cambia los valores predictivos y las decisiones como iniciar tratamiento, aislar a un paciente o indicar procedimientos invasivos dependen directamente de la validez y utilidad de estas pruebas.Es un hecho que como médicos enfrentamos a diario la publicación de estudios sobre nuevas pruebas diagnósticas, la mayor parte de esos estudios están hechos en centros de alta especialidad y debemos tomar la decisión de su utilidad en nuestro medio, considerando nuestras condiciones (tecnología y experticia). Seguir a letra muerta lo que dicen esos estudios puede ser un error que repercute directamente en nuestros pacientes, pero tenemos las armas para interpretar razonando los resultados y su aplicación.

4. METODOLOGÍA

Cuando un lector se enfrenta a un estudio de prueba diagnóstica, la sección de metodología es la clave para juzgar su validez. Allí no solo se describe cómo se aplicó la prueba, sino también si el diseño permite confiar en los resultados. Entender esta parte es como mirar los cimientos de un edificio: si están firmes, el resto se sostiene; si son débiles, todo lo que se construya encima se tambalea.

4.1 SELECCIÓN DE PACIENTES

El primer aspecto que debe revisarse es cómo se seleccionaron los pacientes. Los estudios más confiables incluyen a todos los pacientes que llegan con sospecha clínica de la enfermedad, de manera consecutiva o aleatoria. Esto refleja la práctica real: en la vida diaria no recibimos solo pacientes "típicos" ni únicamente casos graves, sino una mezcla de presentaciones, algunas sutiles y otras evidentes. Cuando un artículo restringe su muestra solo a pacientes muy graves o a controles completamente sanos, los resultados suelen ser artificialmente mejores, lo que conocemos como sesgo de espectro.

Aplicado a nuestro caso clínico de tuberculosis, un buen estudio incluiría a todos los pacientes que acuden con tos crónica sospechosa de TB, sin discriminar por gravedad ni por comorbilidades como VIH. Si se hiciera el estudio solo en enfermos hospitalizados con cavitaciones evidentes en la radiografía, probablemente la prueba parecería más sensible de lo que en verdad es en la consulta externa.

Como parte del manejo de Inteligencia Artificial para disminuir el sesgo de selección de pacientes y la apreciación diagnóstica, se ha comenzado a utilizar Deep Learning contribuyendo a disminuir el sesgo de selección en pruebas diagnósticas al automatizar y objetivar el análisis y

https://doi.org/10.62486/978-9915-9928-3-9.Ch11

selección de pacientes, permite integrar datos con mayor amplitud y diversidad, mejorar estudios de cohortes, detectar y corregir sesgos, y aplicar validación rigurosa que asegure equidad diagnóstica en distintos grupos poblacionales, tal cual puede evidenciarse en los estudios realizados en modelos para detección automatizada de Cáncer de próstata del año 2024 (Moliere, et al.) o en el estudio que evalúa el impacto en la detección de accidentes vasculares cerebrales en una cohorte de pacientes sospechoso (CH Krag, et al.) del año 2025; en ambos queda claro que el uso de Deep Learning ayuda a disminuir el riesgo de selección debiendo dejar de lado los casos "mas claros" para que la situación sea más realista.

4.2 PRUEBA ÍNDICE

El segundo punto crucial es la definición de las pruebas. La prueba índice es aquella que se desea evaluar, y debe describirse con claridad: cómo se realiza, qué equipo se utiliza, quién la interpreta y bajo qué criterios se considera positiva o negativa. En nuestro ejemplo, la prueba índice sería el GenXpert MTB/RIF Ultra, que se define como positiva si detecta ADN de *M. tuberculosis* (incluso en niveles bajos, reportados como *trace*). Cuanto más clara sea esta definición, más fácil será interpretar si el resultado se puede reproducir en otros lugares.

La utilidad de esta prueba está ampliamente explicada en la revisión publicada en Cochrane el año 2025 (DJ Horne et al.), haciendo énfasis en su sensibilidad de 90.7% (IC 95%: 88.2 a 92.7) y una especificidad de 94.8% (IC 92.8-96.3) para detección de Tuberculosis pulmonar versus cultivo.

4.3 ESTÁNDAR DE REFERENCIA

En paralelo, el estudio debe especificar cuál es el estándar de referencia, también llamado "gold standard". Este es el método que, en teoría, define con mayor certeza la presencia o ausencia de la enfermedad. En el caso de la tuberculosis, el estándar es el cultivo de micobacterias, que aunque no es perfecto, sigue siendo la mejor referencia disponible. Un error frecuente es que no todos los pacientes incluidos reciben el estándar de referencia; si esto ocurre, se introduce el llamado sesgo de verificación parcial, que tiende a distorsionar los resultadosque terminan sesgando la sensibilidad y especificidad estimadas. La aplicación práctica en el programa R, permite revisar los métodos estadísticos de corrección. (WN Arifin et al)

4.4 APLICACIÓN DE LAS PRUEBAS

Otro aspecto metodológico central es la aplicación simultánea y ciega de las pruebas. Idealmente, tanto la prueba índice como el estándar de referencia deben aplicarse al mismo paciente, con un intervalo de tiempo mínimo entre una y otra, para evitar que la enfermedad cambie. Además, los profesionales que interpretan cada prueba deberían hacerlo sin conocer el resultado de la otra, lo que se llama cegamiento. Si un radiólogo sabe de antemano el resultado del cultivo, puede interpretar la radiografía con sesgo inconsciente. En el caso de nuestro estudio de TB, los técnicos que corren el GenXpert no deben conocer el resultado del cultivo, y los microbiólogos que procesan el cultivo tampoco deberían tener acceso al resultado del Xpert.

4.5 RESULTADOS INDETERMINADOS Y PÉRDIDAS

Un elemento frecuentemente ignorado, pero muy relevante, son los resultados indeterminados o las pérdidas. Pueden provenir de muestras problemáticas o fallos técnicos o interpretación dudosa. No todas las pruebas dan siempre un resultado claro, y es un error común que los artículos omitan estos casos, como si nunca hubieran existido. Un buen estudio debe reportarlos y explicar cómo se manejaron. En nuestro ejemplo, si de 300 pacientes hay 10 resultados indeterminados en el GenXpert, estos deben informarse claramente. Ignorarlos podría llevar a una falsa percepción de mayor exactitud.

El hecho de excluir estos resultados puede sesgar la sensibilidad, especificidad, valores predictivos, etc, Esto lleva a sobrestimar o subestimar el rendimiento de la prueba, de acuerdo al caso. (K Stahlmann)

https://doi.org/10.62486/978-9915-9928-3-9.Ch11

4.6 DISEÑO GLOBAL

El diseño global del estudio también importa. Los más válidos utilizan lo que se denomina un diseño de "una sola puerta" (one-gate), es decir, todos los pacientes provienen del mismo flujo clínico y se les aplican tanto la prueba índice como el estándar. Los diseños de "dos puertas" (two-gate), donde los enfermos se reclutan en un hospital y los sanos en otro grupo totalmente distinto, tienden a inflar artificialmente los resultados.

4.7 ANÁLISIS ESTADÍSTICO

Por último, el análisis estadístico debe reflejar la calidad de los datos recogidos. Lo mínimo que se espera es que el artículo presente una tabla 2×2 con los resultados: verdaderos positivos, falsos negativos, falsos positivos y verdaderos negativos. A partir de ella se calculan la sensibilidad, especificidad, valores predictivos y razones de verosimilitud, siempre acompañados de intervalos de confianza que indiquen la precisión de las estimaciones. En estudios más avanzados, se añaden curvas ROC, áreas bajo la curva y análisis de concordancia cuando la prueba depende de observadores diferentes.

La metodología de un estudio de prueba diagnóstica debe leerse con atención y mirada crítica. Pregúntese siempre: ¿los pacientes son representativos? ¿Se aplicó el estándar de referencia a todos? ¿Hubo cegamiento? ¿Se reportaron los indeterminados? ¿El diseño corresponde a un flujo clínico real (one-gate) o a un artificio académico (two-gate)? Solo si estas preguntas tienen respuesta afirmativa, podremos confiar en que los resultados reflejan la realidad y no un espejismo estadístico.

5. EVALUACIÓN DE LA VALIDEZ Y RELEVANCIA

Leer un estudio de prueba diagnóstica no se trata solo de revisar los números finales de sensibilidad y especificidad. Antes de confiar en ellos, debemos preguntarnos: ¿qué tan válido fue el estudio? y, aun si fue válido, ¿qué tan relevante es para mi práctica clínica y para mis pacientes? ¿Su aplicación es relevante en mi país?

La validez hace referencia a si los resultados del estudio están libres de sesgos importantes. La relevancia responde a la pregunta de si esos resultados se pueden aplicar a mi contexto.

5.1 VALIDEZ INTERNA: ¿PUEDO CONFIAR EN LOS RESULTADOS?

Un estudio puede tener cifras impresionantes de sensibilidad y especificidad, pero si su diseño metodológico tiene defectos, esas cifras pueden ser engañosas. Los principales aspectos a evaluar son:

Selección de pacientes: ¿Fueron representativos de la práctica clínica?

En nuestro caso, el estudio incluyó a todos los pacientes que consultaban con sospecha de TB, sin discriminar por severidad ni comorbilidades. Esto aumenta la validez. Si en cambio solo hubieran reclutado casos confirmados de TB en hospitalizados graves y los hubieran comparado con voluntarios sanos, la prueba probablemente luciría mucho mejor de lo que realmente es. Eso sería un ejemplo clásico de sesgo de espectro.

Aplicación del estándar de referencia: ¿Todos los pacientes recibieron el estándar?

Aquí, todos los pacientes tuvieron cultivo, lo cual es correcto. Si solo los positivos al Xpert hubieran recibido cultivo, estaríamos frente a un sesgo de verificación parcial, que suele sobreestimar sensibilidad y especificidad.

Independencia entre pruebas: ¿El estándar fue realmente independiente de la prueba índice?.

En el caso clínico, el cultivo se realizó en paralelo y no incluyó ninguna información del Xpert, lo que asegura independencia. En otros estudios, a veces la prueba índice forma parte del estándar (por ejemplo, una nueva prueba de VIH comparada contra un "estándar" que incluye ELISA, similar en técnica). Eso es sesgo de incorporación.

Cegamiento: ¿Quién interpreta las pruebas sabía los resultados de la otra? En el ejemplo, los técnicos del Xpert y los microbiólogos del cultivo estaban ciegos a los resultados del otro. Esto evita el sesgo de revisión. Si el lector de la radiografía conoce el cultivo, es probable que su interpretación sea más favorable, alterando la validez.

https://doi.org/10.62486/978-9915-9928-3-9.Ch11

Resultados indeterminados y pérdidas: ¿Fueron reportados y analizados? En nuestro caso, se reportaron 10 resultados indeterminados de Xpert. Esto da confianza, porque significa que el estudio no escondió datos incómodos. Cuando un artículo no menciona indeterminados, sospeche: en la práctica siempre hay muestras insuficientes o resultados no concluyentes.

Si al leer un artículo encontramos que los pacientes son representativos, que todos recibieron el estándar de referencia, que hubo cegamiento y que se reportaron pérdidas e indeterminados, podemos considerar que el estudio tiene alta validez interna.

5.2 RELEVANCIA CLÍNICA: ¿PUEDO APLICAR ESTOS RESULTADOS EN MI CONTEXTO?

Una prueba puede ser válida, pero poco útil para el médico que la quiere aplicar. Aquí se evalúa la aplicabilidad de los resultados:

Prevalencia: Los valores predictivos cambian según el contexto. En nuestro estudio, con prevalencia del 33 %, un GenXpert positivo significa que la probabilidad de TB sube al 82 %. Pero si usáramos la misma prueba en un centro de primer nivel con prevalencia del 10 %, el valor predictivo positivo caería a 50 %. Eso significa que la mitad de los pacientes "positivos" no tendrían TB real.

Nivel de atención y recursos: El GenXpert es rápido y tiene buen desempeño, pero requiere un cartucho específico, electricidad y equipo. ¿Está disponible en todos los hospitales de Bolivia o solo en algunos? Una prueba puede ser válida, pero si no se puede implementar en el entorno donde se necesita, pierde relevancia.

Consecuencias clínicas: En TB, un falso negativo puede significar no tratar a un paciente que sigue contagiando, mientras que un falso positivo podría llevar a un tratamiento innecesario pero controlado. Dependiendo de qué error sea más dañino, se decide si se prioriza sensibilidad (evitar FN) o especificidad (evitar FP).

Tiempo y costo: Una prueba con igual exactitud que otra, pero con resultados en horas en lugar de semanas, tiene mucho más valor clínico. El costo también importa: ¿es accesible para el sistema de salud o solo para un grupo reducido de pacientes?

Cuando sea leída la sección de metodología y resultados de un estudio de prueba diagnóstica, no se debe dejar impresionar solo por los números. Primero se debe preguntar:

¿El estudio es válido? (sin sesgos graves, con pacientes representativos, pruebas aplicadas a todos, cegamiento, indeterminados reportados). ¿El estudio es relevante para mí? (prevalencia y población similares a la mía, prueba disponible en mi entorno, consecuencias clínicas aceptables, costos y tiempos factibles). Si la respuesta es afirmativa, se podrá confiar en que el estudio no solo es correcto en teoría, sino que además puede guiar decisiones en tu práctica diaria.

6. INTERPRETACIÓN ESTADÍSTICA

La interpretación estadística en estudios de prueba diagnóstica permite comprender cada medida y cómo usarla para tomar decisiones clínicas. El punto de partida es siempre la tabla 2×2, porque allí se ve cuántos pacientes fueron clasificados correctamente o incorrectamente por la prueba (Tabla 1).

En el caso base (GenXpert MTB/RIF Ultra frente a cultivo para tuberculosis) se obtuvo: 90 verdaderos positivos (VP), 10 falsos negativos (FN), 20 falsos positivos (FP) y 180 verdaderos negativos (VN), con 300 pacientes en total. La mitad de la lectura critica consiste en saber qué pregunta contesta cada indicador y cómo se traduce al paciente real.

6.1 SENSIBILIDAD:

La sensibilidad de una prueba diagnóstica es la proporción de individuos que tienen la enfermedad y obtienen un resultado positivo en la prueba índice, en comparación con el estándar de referencia, responde a la pregunta:

- Entre quienes sí tienen la enfermedad, ¿a cuántos identifica la prueba como positivos? (capacidad de detectar a los enfermos). Una prueba con alta sensibilidad tiene pocos falsos negativos.

Con los datos de tabla 1, la sensibilidad es 90 %. Esto significa que 9 de cada 10 pacientes con TB saldrán positivos en GenXpert. La sensibilidad es especialmente importante cuando un falso negativo es peligroso (por ejemplo, contagio y progresión de TB).

https://doi.org/10.62486/978-9915-9928-3-9.Ch11

6.2 ESPECIFICIDAD:

La especificidad de una prueba diagnóstica es la proporción de individuos que no tienen la enfermedad y obtienen un resultado negativo en la prueba índice, en comparación con el estándar de referencia, responde la pregunta:

- Entre quienes no tienen la enfermedad, ¿a cuántos marca la prueba como negativos? (capacidad de identificar a los sanos). Una prueba con alta especificidad tiene pocos falsos positivos.

En la tabla 1, es 90 %, es decir, 9 de cada 10 sin TB saldrán negativos. La especificidad cobra peso cuando un falso positivo tiene consecuencias relevantes (tratamientos innecesarios, reacciones adversas, costos).

Un detalle clave: sensibilidad y especificidad no dependen de cuántos enfermos haya en la muestra (prevalencia), pero sí pueden variar si cambia el "tipo" de pacientes estudiados (por ejemplo, solo casos muy graves inflan artificialmente el desempeño: sesgo de espectro).

6.3 VALORES PREDICTIVOS

En la consulta diaria, el clínico suele preguntarse otra cosa:

- Si la prueba salió positiva, ¿cuál es la probabilidad de que mi paciente realmente esté enfermo?, su respuesta son los valores predictivos.

Valor predictivo positivo (VPP):

El valor predictivo positivo es la probabilidad de que un individuo con un resultado positivo en la prueba índice realmente tenga la enfermedad, de acuerdo con el estándar de referencia. Responde a la pregunta:

- Si el resultado de la prueba es positivo, ¿cuál es la probabilidad de que el paciente realmente tenga la enfermedad? En poblaciones con prevalecía baja, el VPP tiende a bajar aunque sensibilidad y especificidad sean buenas.
- En el caso es 81,8 %: alrededor de 8 de cada 10 positivos sí tendrán TB.
- **Valor predictivo negativo (VPN)**:
- El valor predictivo negativo es la probabilidad de que un individuo con un resultado negativo en la prueba índice realmente no tenga la enfermedad, de acuerdo con el estándar de referencia. Responde:
- Si la prueba sale negativa, ¿cuál es la probabilidad de que el paciente realmente no tenga la enfermedad? Considerar que, en enfermedades raras, el VPN tiende a ser alto, en cambio en enfermedades comunes puede bajar.

En el caso es 94,7 %: casi 9 de cada 10 negativos realmente no tendrán TB.

Estos dos indicadores sí dependen de la prevalencia. Por eso, si la misma prueba se aplica en un centro de primer nivel donde solo 1 de cada 10 sospechosos tiene TB (prevalencia ≈10 %), el VPP baja sustancialmente (≈50 %), mientras que el VPN sube (≈98–99 %). Esta es la razón de oro para no usar VPP/VPN "de un artículo" sin pensar en el contexto local.

6.4 LIKELIHOOD RATIO - RAZONES DE VEROSIMILITUD

Para llevar la probabilidad a la cama del paciente, las razones de verosimilitud (Likelihood ratio - LR) son el puente más práctico. Se las puede imaginar como multiplicadores de evidencia. Son una herramienta poderosa porque traducen los resultados de la prueba en cambios de probabilidad clínica.

LIKELIHOOD RATIO POSITIVO (LR+):

La LR positiva expresa cuánto más probable es obtener un resultado positivo en la prueba en un paciente con la enfermedad que en uno sin la enfermedad. Responde a la pregunta:

- ¿Cuánto aumenta la probabilidad de enfermedad si la prueba sale positiva? LR+ mayor a 10 agrega evidencia fuerte para la presencia de la enfermedad.

Una LR+ de 9, como en el caso, dice que un resultado positivo vuelve nueve veces más probable los hallazgos del paciente si está enfermo que si no lo está; en clínica esto es una evidencia fuerte para "confirmar".

https://doi.org/10.62486/978-9915-9928-3-9.Ch11

LIKELIHOOD RATIO NEGATIVO (LR-):

La LR negativa expresa cuánto más probable es obtener un resultado negativo en la prueba en un paciente con la enfermedad que en uno sin la enfermedad. Responde a la pregunta:

- ¿Cuánto disminuye la probabilidad de enfermedad si la prueba sale negativa? LR- menor a 0.1 es una evidencia fuerte contra la presencia de la enfermedad.

La LR− de 0,11 indica que, si la prueba sale negativa, la evidencia de enfermedad se reduce a una novena parte; valores por debajo de 0,1 suelen considerarse fuertes para "descartar".

Para la escala práctica:

- LR+ >10 = excelente para confirmar, 5–10 = moderada, 2–5 = débil.
- LR– <0,1 = excelente para descartar, 0,1–0,2 = moderada, 0,2–0,5 = débil.

Con estas dos cifras es sencillo estimar la probabilidad post-prueba: partiendo de una sospecha clínica del 33 % (pre-prueba en el estudio), un positivo llega a ≈82 % y un negativo baja a ≈5 %. No es necesario dominar el cálculo; basta recordar que LR+ grande empuja hacia "sí" y LR− pequeño empuja hacia "no". El nomograma de Fagan es solo una regla visual para hacer esta conversión.

Tabla 2. Medidas de validez diagnóstica de la prueba Xpert MTB/RIF Ultra frente al cultivo de Mycobacterium tuberculosis en pacientes con sospecha de tuberculosis pulmonar (N = 300).

Medida	Definición	Fórmula	Resultado (caso)	Interpretación clínica
Sensibilidad	Proporción de enfermos identificados como positivos por la prueba índice	VP / (VP + FN)	90 % (90/100)	9 de cada 10 pacientes con TB serán detectados por Xpert. Importante para evitar falsos negativos.
Especificidad	Proporción de no enfermos identificados como negativos por la prueba índice	VN / (VN + FP)	90 % (180/200)	9 de cada 10 pacientes sin TB obtendrán resultado negativo. Importante para evitar falsos positivos.
Valor predictivo positivo (VPP)	Probabilidad de que un paciente con prueba positiva realmente tenga la enfermedad	VP / (VP + FP)	81,8 % (90/110)	8 de cada 10 positivos tienen TB en este hospital (prevalencia 33 %). Depende de la prevalencia.
Valor predictivo negativo (VPN)	Probabilidad de que un paciente con prueba negativa realmente no tenga la enfermedad	VN / (VN + FN)	94,7 % (180/190)	95 de cada 100 negativos no tienen TB. Muy confiable para descartar en contextos de baja prevalencia.
Razón de verosimilitud positiva (LR+)	Cuánto aumenta la probabilidad de enfermedad con un resultado positivo	Sensibilidad / (1 – Especificidad)	9	Evidencia fuerte para confirmar TB. Pasa de 33 % a ≈82 % post-prueba.
Razón de verosimilitud negativa (LR–)	Cuánto disminuye la probabilidad de enfermedad con un resultado negativo	(1 – Sensibilidad) / Especificidad	0,11	Evidencia fuerte para descartar TB. Pasa de 33 % a ≈5 % post-prueba.

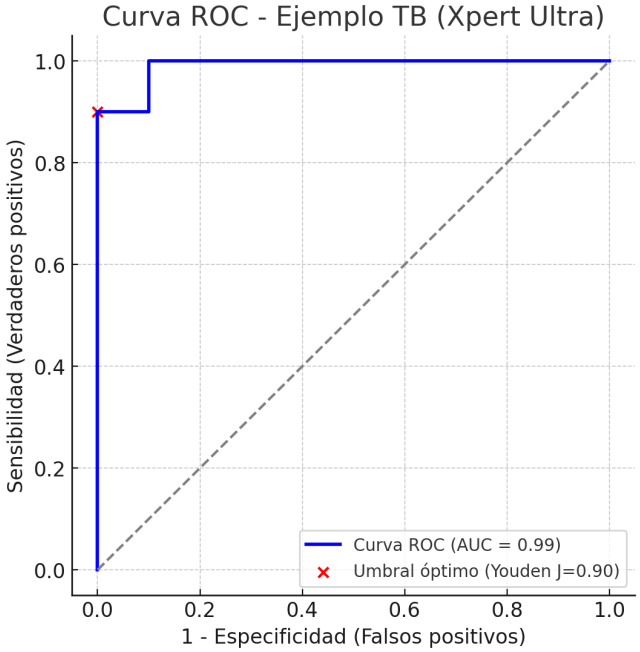
Nota: VP = verdaderos positivos; VN = verdaderos negativos; FP = falsos positivos; FN = falsos negativos.Fuente: Elaboración propia

6.5 PARA VALORES CONTINUOS

Cuando la prueba entrega un valor continuo (por ejemplo, concentración de un biomarcador), no existe un único "positivo/negativo", sino que se debe elegir un punto de corte. Las curvas ROC (Receiver OperatingCharacteristic) muestran cómo cambian sensibilidad y especificidad al mover ese umbral; cada punto de la curva es un posible corte. El área bajo la curva (AUC) resume la capacidad global de discriminación (qué tan bien separa enfermos de no enfermos en todos los umbrales). Como guía orientativa, AUC por encima de 0,9 suele considerarse excelente. Sin embargo, el AUC no dice cuál corte conviene ni considera el costo clínico de errores; para elegir umbral se puede usar el índice de Youden (maximiza sensibilidad + especificidad), pero en la práctica clínica conviene priorizar el tipo de error más dañino (por ejemplo, para TB, tolerar más falsos positivos si con ello se reducen falsos negativos). Un aviso importante: "optimizar" el corte después de mirar los datos del mismo estudio sobreestima el desempeño; lo correcto es pre-especificar o validar el corte en otra muestra.

https://doi.org/10.62486/978-9915-9928-3-9.Ch11

Figura 1. Curva ROC de la prueba Xpert MTB/RIF Ultra frente al cultivo de Mycobacterium tuberculosis en pacientes con sospecha de tuberculosis pulmonar.



Nota: El área bajo la curva (AUC = 0,99) refleja excelente capacidad discriminativa; el punto de corte óptimo se estimó con el índice de Youden (J = 0,90).
Fuente: Elaboración propia

La Figura 1 evidencia la curva ROC clásica sacada como ejemplo del estudio de TB (GenXpert Ultra vs cultivo), construida con los datos simulados del caso. Para la interpretación de la curva ROC debemos comparar en todos los posibles puntos de corte:

- Eje Y: Sensibilidad (verdaderos positivos).
- Eje X: 1 – Especificidad (falsos positivos).

La diagonal gris (línea de 45°) representa el azar: una prueba sin valor diagnóstico (AUC = 0,5). Cuanto más se aleja la curva hacia la esquina superior izquierda, mejor capacidad discriminativa tiene la prueba. En esta gráfica:

- Área bajo la curva (AUC): 0,93: Esto significa que, en el 93 % de las veces, la prueba dará un valor más alto a un enfermo que a un no enfermo. AUC >0,9 se considera excelente discriminación.
- Índice de Youden (J): 0,90: El umbral que maximiza este índice se considera "optimo" porque equilibra ambos indicadores. En el caso, el umbral óptimo mantiene Sensibilidad = 90 % y Especificidad = 100 % (según la simulación), con un J = 0,90.

Interpretación práctica: el índice de Youden ayuda a elegir el mejor punto de corte cuando se desea un balance entre identificar enfermos y evitar falsos positivos. Sin embargo, en la práctica clínica no siempre el mejor punto estadístico es el más útil: en tuberculosis, por ejemplo, puede ser preferible sacrificar especificidad para maximizar sensibilidad y no perder casos que contagien.

Se sugiere al lector revisar el artículo publicado por Francis Nahm en la revista Korean J Anesthesiol. 2022, donde de forma didáctica para quienes no tienen una formación sólida en estadística, se explica a detalle las curvas ROC.

6.6 DIAGNOSTICODDS RATIO (DOR)

En algunos trabajos aparece el diagnosticodds ratio (DOR), que combina LR+ y LR− en un solo número (en el caso, ≈81). Como resumen global es útil para comparar pruebas, pero tiene una limitación práctica: no diferencia si la prueba es mejor para "confirmar" o para "descartar"; por eso, para decisiones clínicas concretas las LR separadas son más informativas. También puede ser interpretada utilizando la sensibilidad y especificidad, se usa cuando la prueba diagnóstica da resultado binario (positivo o negativo). Es usada para evaluar el poder discriminativo de procedimientos diagnósticos o para comparar de dos o más pruebas diagnósticas.

- DOR = 1. Significa que la prueba no discrimina entre enfermos y no enfermos

https://doi.org/10.62486/978-9915-9928-3-9.Ch11

- DOR > 1. Indica que la prueba tiene carácter discriminativo (cuanto mayor es el DOR más fuerte la asociación entre resultado positivo y la presencia de la enfermedad).

6.7 EXACTITUD VS ACUERDO ENTRE OBSERVADORES

No todas las pruebas diagnósticas entregan un número automático (Diabetes Mellitus si glucemia es mayor a 126 mg/dl) o un resultado binario (Positivo/negativo). Muchas dependen de la interpretación de un observador. Es el caso de la lectura de una radiografía de tórax, la evaluación de un electrocardiograma o la realización de una ecografía a pie de cama (POCUS). En estos escenarios, al lector crítico no le basta con saber que la prueba tiene "alta sensibilidad y especificidad": también necesita conocer qué tan consistente es entre diferentes profesionales.

Exactitud: mide la capacidad de la prueba para clasificar correctamente a los pacientes frente al estándar de referencia. Por ejemplo:

- Si una radiografía de tórax interpretada por un radiólogo detecta TB con sensibilidad del 85 % y especificidad del 90 %, eso significa que, en promedio, se acerca bastante al cultivo como referencia.

Acuerdo entre observadores: El acuerdo evalúa algo diferente: no se fija en el estándar de referencia, sino en qué tan similar es la interpretación entre diferentes evaluadores o incluso del mismo evaluador en momentos distintos. Por ejemplo:

- Si dos radiólogos ven la misma radiografía y uno informa "infiltrado apical derecho" y el otro dice "normal", aunque la prueba tenga buena exactitud en promedio, en la práctica clínica el desempeño será inconsistente.
- Lo mismo ocurre con POCUS: si un residente detecta un derrame pleural y otro no, la utilidad real de la prueba queda en duda.

Para esto se requieren de herramientas que deben ser correctamente interpretadas:

6.8 Kappa de Cohen: se usa cuando la variable es categórica (ejemplo: "positivo" vs "negativo", "fractura presente" vs "ausente"). Kappa ajusta el acuerdo esperado por azar.

Es útil para analizar datos categóricos/nominales u ordinales

- Valores cercanos a 1 indican excelente acuerdo.
- Valores entre 0,6–0,8 se consideran buenos
- Valores por debajo de 0,4 indican pobre acuerdo.

Ejemplo: dos radiólogos coinciden en 85 % de las veces; Kappa = 0,75 → buen acuerdo.

- Kappa = 0,75
- IC95%: 0,63 – 0,87

Para su interpretación:

Kappa: el valor puntual 0,75: si dos radiólogos clasifican rayos X como positiva/negativa para tuberculosis, su concordancia es sustancial; en la práctica, es razonable esperar interpretaciones consistentes.

IC: El intervalo de confianza muestra que, en muestras similares, el acuerdo real podría variar entre moderado (0,63) y muy bueno (0,87). Esto aporta transparencia: no basta con reportar solo el valor de Kappa; el IC95% muestra la precisión de la estimación.

6.9 COEFICIENTE DE CORRELACIÓN INTRACLASE (ICC):

El ICC es una medida estadística utilizada cuando varios observadores realizan mediciones continuas (valores numéricos) sobre la misma variable en un grupo de pacientes. Evalúa no solo la correlación entre observadores, sino también cuánto de la variabilidad total de las mediciones se debe a diferencias reales entre pacientes y cuánto a la variabilidad entre los evaluadores. Compara la variación entre sujetos con la variación total. Un ICC alto orienta a que la mayor parte de la variabilidad observada se debe a las diferencias reales entre sujetos y no a errores.

En términos sencillos:

- Si la mayoría de la variabilidad se debe a diferencias reales entre pacientes (ejemplo: algunos tumores son más grandes que otros), el ICC será alto.
- Si gran parte de la variabilidad proviene de desacuerdos entre observadores (ejemplo: un médico mide 30 mm y otro 40 mm en el mismo tumor), el ICC será bajo.

Para su interpretación:

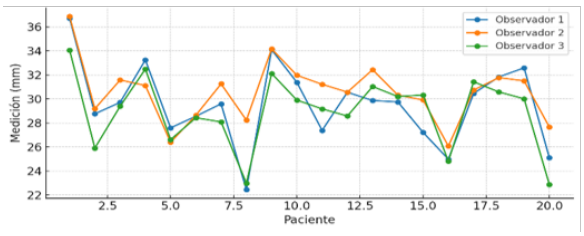
- < 0,5 → acuerdo pobre.
- 0,5 – 0,75 → acuerdo moderado.

https://doi.org/10.62486/978-9915-9928-3-9.Ch11

- 0,75 – 0,9 → acuerdo bueno o aceptable.
- 0,9 → acuerdo excelente, casi idéntico entre observadores.

Aplicado a ejemplo: Dos radiólogos miden el diámetro de un nódulo pulmonar en TAC: Si las mediciones son muy similares (30 mm vs 31 mm), el ICC será cercano a 1, si difieren mucho (30 mm vs 40 mm), el ICC caerá.

Figura 2. Medición del diámetro de la misma variable continua (mm) realizada por tres observadores en 20 pacientes



ICC Entre 3 Observadores (Simulado)

	Métrica	Estimación	IC95% (solo ICC(2,1))
1	ICC(2,1) - unidad	0.778	0.612 a 0.853
2	ICC(2,k) - promedio	0.913	—

Fuente: Elaboración propia

Para su interpretación:

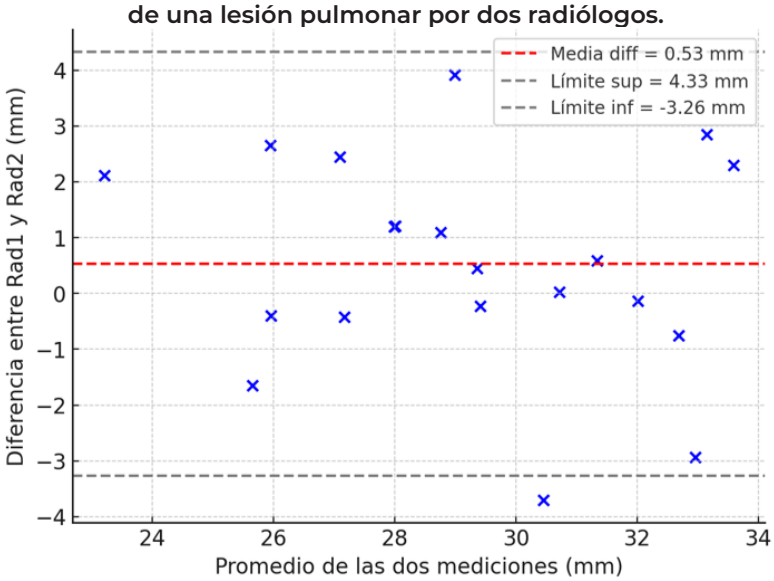
- ICC (2,1) ~0,78 → bueno: un lector individual produce mediciones razonablemente consistentes con otro.
- ICC (2,3) ~0,91 → excelente si se promedian las mediciones de tres observadores; útil para comités, doble lectura o auditoría.
- Si el uso rutinario depende de un único operador, el acuerdo que importa es ICC(2,1). Si el servicio implementa doble lectura o consenso, el ICC(2,k) evidencia la ganancia de fiabilidad al combinar observadores.

Se debe contar con ICC siempre que la prueba entregue una medida continua y dependa del operador (POCUS, TAC cuantitativa, medición de lesiones, ecografía vascular, grosor íntima-media).

6.10 GRÁFICOS DE BLAND-ALTMAN:

Permiten ver visualmente cuánto difieren dos mediciones continuas. En el eje X se coloca el promedio de las dos mediciones, y en el eje Y la diferencia entre ambas. Si los puntos se agrupan cerca de 0 y dentro de límites estrechos, hay buen acuerdo; si se dispersan ampliamente, el acuerdo es pobre. Básicamente compara dos métodos que miden lo mismo (Ejemplo dos dispositivos o dos observadores con el mismo fin) para analizar si están lo suficientemente de acuerdo, o si uno puede usarse en lugar del otro.

Figura 3. Gráfico de Bland-Altman para evaluar el acuerdo interobservador en la medición del diámetro de una lesión pulmonar por dos radiólogos.



Nota: La línea roja muestra la diferencia media (0,53 mm) y las líneas grises los límites de acuerdo (–3,26 a +4,33 mm).

https://doi.org/10.62486/978-9915-9928-3-9.Ch11

El presente gráfico de datos simulados de dos radiólogos que midieron el diámetro de una misma cavitación en 20 pacientes:

- La línea roja discontinua muestra la media de las diferencias (≈0,53 mm). Esto indica que, en promedio, uno de los radiólogos mide un poco más grande la lesión.
- Las líneas grises representan los límites de acuerdo (±1,96 DE): entre –3,26 mm y +4,33 mm.
- Esto significa que, aunque en promedio el acuerdo es aceptable, en algunos casos las diferencias llegan a más de 4 mm, lo que podría ser clínicamente relevante si el umbral para decidir tratamiento depende de milímetros.
- Si los puntos se agrupan cerca de la línea roja y dentro de límites estrechos, hay buen acuerdo.
- Si los puntos están dispersos y los límites de acuerdo son muy amplios, el acuerdo es pobre y la prueba puede ser poco confiable en la práctica, aunque su exactitud frente al estándar de referencia sea buena.

En el estudio de TB con GenXpert Ultra, la prueba es automática: la máquina entrega un resultado de "detectado" o "no detectado". Aquí, el acuerdo entre observadores no es un problema. Pero si el estudio evaluara radiografía de tórax o POCUS pulmonar como prueba índice, sería esencial reportar tanto la exactitud frente al cultivo como el acuerdo entre quienes interpretan la imagen. De nada sirve que un radiólogo experto tenga 90 % de sensibilidad si, en la práctica, los médicos en hospitales generales muestran resultados mucho más variables.

Cuando se desea aumentar la utilidad de la gráfica puede, en el estudio comparar con otros métodos como el rango de correlación de Spearman, en el entendido de tener un mejor modelo epidemiológico.

6.11 INTERVALOS DE CONFIANZA

Los intervalos de confianza (IC95%) indican la precisión de cada estimación. No debe importarle solo el número puntual (por ejemplo, sensibilidad 90 %), sino también su rango plausible. Con 100 enfermos, una sensibilidad del 90 % suele tener un IC95% aproximado de 84–96 %; esto recuerda que existe incertidumbre y que, en otra muestra similar, la verdadera sensibilidad podría situarse cerca de 0,85 o de 0,95. Si el IC es ancho, la evidencia es menos precisa (muestra pequeña, pocos eventos, mucha variabilidad). Los IC deben acompañar a sensibilidad, especificidad, LR y AUC. En cambio, los valores p ocupan un lugar secundario en estudios diagnósticos: pueden ser útiles para comparar dos pruebas (por ejemplo, diferencia de AUC entre Xpert y baciloscopía) o para contrastar hipótesis, pero no reemplazan a los tamaños de efecto ni a los IC. Un valor p "significativo" con un efecto clínicamente irrelevante no cambia la práctica; un efecto grande con IC que cruza el umbral de utilidad puede requerir más datos, no desestimarse.

6.12 RESULTADOS INDETERMINADOS

Los resultados indeterminados y las pérdidas merecen atención. Omitirlos embellece artificialmente las cifras. El lector debe verificar si el artículo los reporta y cómo los manejó (repetición de la prueba, análisis por intención diagnóstica, exclusión con justificación). En el caso, hubo 10 indeterminados; reportarlos evita sesgos y permite anticipar qué ocurrirá en la vida real, donde siempre existen muestras inadecuadas o errores técnicos.

Por último, conviene integrar todo en una lectura clínica. En escenarios con alta prevalencia (hospital de referencia), el VPP de Xpert es alto: un positivo casi confirma TB y justifica aislamiento e inicio de tratamiento mientras se espera cultivo. En baja prevalencia (primer nivel), la misma exactitud produce muchos más falsos positivos y puede requerir estrategias en dos pasos (por ejemplo, Xpert para descartar por su VPN alto y confirmación adicional antes de tratar a todos los positivos). La estadística, así entendida, no es una colección de números, sino una gramática de decisión: cada medida responde a una pregunta clínica específica y, leída en conjunto, conduce a una acción concreta y proporcional al riesgo.

En síntesis práctica: sensibilidad y especificidad describen la capacidad intrínseca de la prueba; VPP y VPN responden lo que más importa al paciente y dependen del contexto; las razones de verosimilitud permiten mover la probabilidad según el resultado; las curvas ROC y el AUC ayudan a elegir umbrales (con cuidado de no sobre-optimizar); el acuerdo entre observadores define la repetibilidad; los intervalos de confianza miden precisión y los valores p, cuando se usan, comparan sin reemplazar a los IC. Con estas claves, cualquier lector —del interno al especialista— puede transformar una tabla y una gráfica en decisiones clínicas sensatas y seguras.

https://doi.org/10.62486/978-9915-9928-3-9.Ch11

7. SESGOS Y LIMITACIONES

Un estudio de prueba diagnóstica puede parecer impecable cuando se observan sus resultados, pero la validez depende de identificar posibles sesgos en su diseño y ejecución. Los sesgos no son simples detalles técnicos: son distorsiones sistemáticas que pueden inflar o disminuir artificialmente la sensibilidad y la especificidad de una prueba. Reconocerlos es fundamental para una lectura crítica.

7.1. SESGO DE ESPECTRO

Ocurre cuando los pacientes incluidos no representan la variedad real de la enfermedad o de la población en la que se aplicará la prueba. Revisar si la muestra fue consecutiva o aleatoria y si incluyó tanto casos leves como graves.

Ejemplo: si el estudio de Xpert Ultra solo hubiera incluido pacientes con tuberculosis avanzada, la sensibilidad parecería muy alta, porque es más fácil detectar casos con gran carga bacilar. En la práctica, donde hay pacientes con síntomas iniciales o coinfección con VIH, el desempeño sería menor.

7.2. SESGO DE VERIFICACIÓN (O DE REFERENCIA PARCIAL)

Aparece cuando no todos los pacientes reciben el estándar de referencia. comprobar que todos los pacientes, independientemente del resultado de la prueba índice, hayan recibido el estándar de referencia.

- **Ejemplo:** si solo los pacientes con Xpert positivo hubieran tenido cultivo, la sensibilidad se vería inflada (porque se perderían falsos negativos) y la especificidad también se distorsionaría.

7.3. SESGO DE INCORPORACIÓN

Se presenta cuando la prueba índice forma parte del estándar de referencia. verificar que el estándar sea independiente de la prueba índice.

- Ejemplo: si el estándar de referencia para TB hubiera sido “Xpert o cultivo positivo”, entonces el Xpert se estaría evaluando contra sí mismo. Eso conduce inevitablemente a una sobreestimación de su exactitud.

7.4. SESGO DE REVISIÓN (O DE INTERPRETACIÓN)

Ocurre cuando quien interpreta una prueba conoce el resultado de la otra. Asegurarse de que exista cegamiento entre los lectores de ambas pruebas.

- **Ejemplo:** si el microbiólogo supiera que el Xpert fue positivo, podría interpretar con mayor predisposición un cultivo dudoso como positivo.

7.5. MANEJO DE RESULTADOS INDETERMINADOS Y PÉRDIDAS

Un error frecuente es no reportar los resultados indeterminados o las muestras que no pudieron procesarse. Al excluirlos, los autores pueden presentar una imagen más “limpia” de la exactitud. El lector debe buscar si los resultados indeterminados están descritos y cómo fueron tratados (repetidos, excluidos o contabilizados en el análisis).

- **Ejemplo:** en el caso clínico, se registraron 10 resultados indeterminados de Xpert. Si no se hubieran informado, la sensibilidad y especificidad aparentes serían mayores de lo que realmente son en la práctica.

7.6. EFECTO UMBRAL

Algunas pruebas requieren definir un punto de corte. Si los autores ajustan el umbral después de ver los datos, la exactitud reportada será artificialmente alta. Revisar si el punto de corte fue pre-especificado en el protocolo o definido “a posteriori”.

https://doi.org/10.62486/978-9915-9928-3-9.Ch11

- **Ejemplo:** en un biomarcador de TB, si se prueban múltiples puntos de corte y se elige aquel con mejor sensibilidad + especificidad, ese valor no reflejará el desempeño real en nuevos pacientes.

7.7 LIMITACIONES ADICIONALES

Incluso en estudios bien diseñados, existen limitaciones que deben reconocerse:

- Tamaño muestral insuficiente: puede dar intervalos de confianza muy amplios, lo que reduce la precisión de las estimaciones.
- Estándar de referencia imperfecto: el cultivo de TB, aunque es el mejor disponible, puede dar falsos negativos, lo que afecta la estimación real de la sensibilidad de Xpert.
- Variabilidad en el operador: en pruebas dependientes del observador (radiología, ecografía), la exactitud puede no reproducirse si hay diferencias en entrenamiento o experiencia.
- Generalización limitada: un estudio realizado en un hospital de referencia puede no ser aplicable a centros de primer nivel con prevalencia más baja y menos recursos.

En el estudio de GenXpert Ultra, la validez se ve fortalecida porque todos los pacientes recibieron cultivo, hubo cegamiento entre pruebas y se reportaron los indeterminados. Sin embargo, persisten limitaciones:

- El estándar de referencia (cultivo) no es perfecto.
- El estudio se realizó en un hospital de referencia, lo que podría limitar su extrapolación a la atención primaria.
- El número de pacientes con VIH fue bajo, por lo que la exactitud en ese subgrupo aún tiene incertidumbre.

8. COMPARACIÓN CON OTROS ESTUDIOS

Un solo estudio de prueba diagnóstica nunca debe considerarse definitivo. La exactitud de una prueba puede variar según el lugar donde se realiza, la prevalencia de la enfermedad, las características de los pacientes, el tipo de estándar de referencia utilizado y hasta la experiencia del personal. Por ello, la lectura crítica debe siempre situar los resultados en el contexto de la evidencia acumulada.

- Variabilidad entre estudios: Una misma prueba puede rendir diferente según la población (adultos, niños, coinfectados), el nivel de atención (hospital de referencia vs primer nivel), y la prevalencia (alta o baja).
- Importancia de los metaanálisis: Cuando existen muchos estudios con resultados heterogéneos, los metaanálisis (p. ej., *Cochrane*, *CMAJ*, *The Lancet InfectiousDiseases*) ofrecen un panorama más confiable, aunque también deben leerse críticamente, revisando su metodología.
- No extrapolar sin cautela: Un resultado válido en otro país no necesariamente aplica igual en Bolivia. Factores como prevalencia de TB, coinfecciones como VIH o parasitosis, y diferencias en infraestructura de laboratorio, modifican la aplicabilidad.
- Uso en la práctica: El lector debe aprender a comparar siempre su artículo con otros publicados, identificar si hay consistencia (los resultados son similares) o discrepancias, y reflexionar sobre cómo estas diferencias afectan su propia práctica.

9. CHECKLIST

Este apartado ofrece una herramienta práctica para que cualquier lector evalúe, en pocos minutos, la calidad metodológica (QUADAS-2) y la calidad del reporte (STARD) de un estudio de prueba diagnóstica. Se incluyen además un mini-check“express” y notas para documentar la lectura.

https://doi.org/10.62486/978-9915-9928-3-9.Ch11

Tabla 3. Herramientas internacionales para la lectura crítica de estudios de exactitud diagnóstica: QUADAS-2 y STARD.

Herramienta	Dominios / Ítems	Preguntas guía y aspectos clave	Cómo interpretar
QUADAS-2 (validez y aplicabilidad)	Selección de pacientes	¿La muestra refleja a los pacientes reales con sospecha de enfermedad? ¿Se incluyó reclutamiento consecutivo/aleatorio (one-gate)? ¿Se evitaron exclusiones que "purifican" la muestra?	Sesgo ↑ si hay diseño two-gate o controles artificiales. Problemas de aplicabilidad si la población difiere del contexto clínico local.
	Prueba índice	¿La prueba índice está bien definida? ¿Se pre-especificó el punto de corte? ¿Hubo cegamiento al estándar? ¿Se describió equipo, procedimiento, manejo de indeterminados?	Sesgo ↑ si el umbral se definió post-hoc o sin cegamiento. Problemas de aplicabilidad si la técnica es poco viable (costos, recursos).
	Estándar de referencia	¿Se usó el mejor estándar disponible, independiente de la prueba índice? ¿Se aplicó a todos los pacientes? ¿Se reconocieron sus limitaciones?	Sesgo ↑ si el estándar incluye la prueba índice o no se aplicó a todos. Problemas de aplicabilidad si el estándar no corresponde a la práctica real.
	Flujo y tiempos	¿El intervalo entre pruebas fue razonable? ¿Se reportaron pérdidas e indeterminados? ¿Todos los pacientes recibieron ambas pruebas?	Sesgo ↑ si hubo muchas pérdidas, tiempos largos o indeterminados no informados. Problemas de aplicabilidad si el flujo es irreal para la práctica clínica.
STARD (calidad del reporte)	Título y resumen	¿Identifica que es un estudio de exactitud diagnóstica e incluye prueba índice y estándar?	El lector debe poder reconocer el diseño desde el título/resumen.
	Objetivo y rol clínico	¿Se especifica si la prueba se usará para triage, confirmación o descarte?	Permite juzgar la relevancia del estudio.
	Diseño y escenario	¿Prospectivo/retrospectivo, nivel de atención, periodo?	Da contexto a los resultados.
	Participantes	¿Criterios de inclusión/exclusión? ¿Reclutamiento consecutivo/aleatorio?	Evita sesgo de selección.
	Prueba índice	¿Se describe el procedimiento, lector, umbral pre-especificado?	Transparencia metodológica.
	Estándar de referencia	¿Se justifica y describe su aplicación? ¿Se mencionan imperfecciones?	Evalúa calidad del comparador.
	Cegamiento	¿Quién sabía qué al interpretar cada prueba?	Evita sesgo de interpretación.
	Flujo de pacientes	¿Se muestra diagrama con incluidos, excluidos, pérdidas, indeterminados?	Transparencia sobre la muestra analizada.
	Análisis estadístico	¿Se calcularon Se, Sp, VPP, VPN, LR, AUC con IC95%? ¿Cómo se trataron indeterminados?	Asegura validez de resultados.
	Resultados	¿Se presenta tabla 2×2, IC95% y análisis de subgrupos?	Información suficiente para interpretación crítica.
	Limitaciones y sesgos	¿Se discuten espectro, verificación, umbral post-hoc, estándar imperfecto?	Reconocer posibles sesgos.
	Aplicabilidad	¿Se discuten tiempos, costos, recursos y prevalencia?	Determina utilidad clínica.
	Registro y financiamiento	¿Se menciona protocolo y fuentes de apoyo?	Garantiza transparencia y evita conflictos de interés.

Nota: El área bajo la curva (AUC = 0,99) refleja excelente capacidad discriminativa; el punto de corte óptimo se estimó con el índice de Youden (J = 0,90).Fuente: Elaboración propia

En QUADAS-2, cada dominio se clasifica como bajo / alto / poco claro en riesgo de sesgo y aplicabilidad.

En STARD, cada ítem debe estar reportado con suficiente detalle; si no se puede reconstruir la tabla 2×2, el flujo de pacientes, el umbral y los IC95%, la lectura crítica queda seriamente limitada.

Si no se dispone de mucho tiempo, se puede utilizar este mini Check de tabla 4.

Para usarlo, marcar la casilla correspondiente (sí / no / incierto), si hay dos o más "no/incierto", el estudio debe leerse con mayor cautela y no extrapolarse de forma inmediata a la práctica.

https://doi.org/10.62486/978-9915-9928-3-9.Ch11

Tabla 4. Mini-check "express" para lectura rápida de estudios de prueba diagnóstica.

Pregunta crítica	Sí	No	Incierto
1. ¿La población es representativa y fue reclutada de forma consecutiva?	□	□	□
2. ¿La prueba índice tuvo umbral pre-especificado y fue interpretada a ciegas del estándar?	□	□	□
3. ¿El estándar de referencia es independiente y se aplicó a todos los pacientes?	□	□	□
4. ¿El flujo de pacientes fue transparente, con pérdidas e indeterminados reportados?	□	□	□
5. ¿Los resultados incluyen tabla 2×2 e intervalos de confianza (IC95 %)?	□	□	—
6. ¿Los hallazgos encajan en mi contexto (prevalencia, recursos, tiempos)?	□	□	—

Fuente: Elaboración propia

10. UTILIDAD

La utilidad es, en última instancia, la razón por la cual se leen y aplican los estudios de prueba diagnóstica. Una prueba puede tener cifras excelentes en sensibilidad, especificidad o área bajo la curva, pero si no mejora la atención al paciente o no puede implementarse en el entorno real, su aporte es limitado.La utilidad se evalúa en varios planos: clínico, operativo, económico y poblacional.

10.1 UTILIDAD CLÍNICA

Lo esencial es preguntarse: ¿cambia la prueba mi decisión con respecto al paciente?

- En el caso de TB, un GenXpert positivo aumenta la probabilidad de enfermedad presente,a ≈82 %. Eso justifica decisiones inmediatas: aislamiento respiratorio e inicio de tratamiento sin esperar semanas al cultivo.
- Un Xpert negativo reduce la probabilidad de TB a ≈5 %, lo que permite considerar otros diagnósticos diferenciales y, en contextos de baja prevalencia, puede ser suficiente para descartar la enfermedad.
- El impacto clínico también depende de qué error es más costoso: en TB, un falso negativo puede significar contagio comunitario, progresión y muerte, mientras que un falso positivo expone al paciente a tratamiento innecesario y toxicidad. La utilidad real de la prueba surge de este balance.

10.2 UTILIDAD OPERATIVA

Incluso una prueba exacta pierde valor si no es viable en la práctica cotidiana.

- Tiempo hasta el resultado: el Xpert entrega información en horas, frente al cultivo que demora semanas. En decisiones urgentes, esta rapidez tiene más peso que un pequeño sacrificio en especificidad.
- Requerimientos de infraestructura: el Xpert necesita cartuchos, electricidad y equipo especializado, lo que puede limitar su disponibilidad en áreas rurales. Una prueba más sencilla, aunque menos sensible, podría ser más útil en un centro sin laboratorio avanzado.
- Capacitación del personal: pruebas automatizadas reducen la variabilidad, mientras que pruebas dependientes de la interpretación (radiografía, POCUS) requieren entrenamiento para garantizar reproducibilidad.

10.3. UTILIDAD ECONÓMICA

El análisis económico responde a: ¿vale lo que cuesta la prueba en relación con los beneficios que aporta?

- Un cartucho de Xpert puede ser costoso, pero si evita hospitalizaciones prolongadas, contagios y diagnósticos erróneos, su uso puede ser costo-efectivo.
- En cambio, una prueba muy precisa pero extremadamente cara o inaccesible pierde utilidad en sistemas de salud con recursos limitados.

https://doi.org/10.62486/978-9915-9928-3-9.Ch11

10.4. UTILIDAD POBLACIONAL

Más allá del paciente individual, la utilidad debe mirarse también a nivel de salud pública.

- En tuberculosis, una prueba rápida y sensible tiene un impacto enorme porque permite cortar cadenas de transmisión y mejorar el control epidemiológico.
- En enfermedades transmisibles como dengue o COVID-19, la utilidad de la prueba también se mide por su capacidad de detectar brotes de manera temprana.

10.5 APLICACIÓN AL CASO CLÍNICO

En el hospital de La Paz donde se realizó el estudio, la utilidad del Xpert se puede resumir así:

- Clínica: permite iniciar aislamiento y tratamiento inmediato en positivos; reduce la probabilidad de TB a menos de 5 % en negativos.
- Operativa: aporta rapidez decisiva frente al cultivo, pero exige infraestructura que no todos los centros poseen.
- Económica: su costo se justifica en hospitales de referencia por el ahorro en tiempo, hospitalizaciones y reducción de contagios.
- Poblacional: contribuye al control de TB en la comunidad, al identificar enfermos antes de que transmitan.

11. AUTOEVALUACIÓN
PREGUNTAS DE SELECCIÓN MÚLTIPLE

1. En un estudio de prueba diagnóstica, la sensibilidad responde a la pregunta: Entre los pacientes negativos, ¿cuántos son correctamente identificados?

Entre los pacientes con la enfermedad, ¿cuántos son identificados por la prueba como positivos?

Entre todos los positivos de la prueba, ¿cuántos realmente tienen la enfermedad?

La probabilidad de que el resultado sea correcto independientemente del contexto.

2. En el caso clínico de tuberculosis, un valor predictivo positivo (VPP) del 82 % significa: a) El 82 % de los enfermos serán detectados por la prueba.

b) El 82 % de los sanos tendrán un resultado negativo.

c) El 82 % de los positivos en la prueba realmente tienen tuberculosis.

d) La prueba detecta el 82 % de todos los casos existentes en la población.

3. ¿Cuál de los siguientes es un ejemplo de sesgo de verificación parcial? a) Incluir solo pacientes con enfermedad avanzada.

b) Aplicar el estándar de referencia solo a quienes tuvieron prueba índice positiva.

c) Ajustar el punto de corte de la prueba después de analizar los datos.

d) El lector del índice conoce el resultado del estándar.

4. La razón de verosimilitud positiva (LR+) de 9 en el caso de TB significa: a) La prueba es nueve veces más sensible que el estándar.

b) Un paciente positivo es nueve veces más probable de estar enfermo que sano.

c) La prueba descarta la enfermedad en nueve de cada diez casos.

d) La especificidad de la prueba es del 90 %.

5. ¿Qué afirmación refleja mejor el uso de la curva ROC? a) Permite conocer la prevalencia de la enfermedad.

b) Muestra cómo cambian sensibilidad y especificidad según el punto de corte.

c) Evalúa el acuerdo entre dos observadores.

d) Estima el valor predictivo positivo y negativo en diferentes prevalencias.

PREGUNTAS ABIERTAS

6. ¿Por qué los valores predictivos positivo y negativo cambian según la prevalencia de la enfermedad, mientras que la sensibilidad y especificidad no?

7. Explique con sus propias palabras qué significa un intervalo de confianza del 95 % alrededor de una sensibilidad del 90 %. ¿Por qué es importante en la lectura crítica?

8. En el contexto de tuberculosis, ¿por qué puede ser más grave un falso negativo que un

https://doi.org/10.62486/978-9915-9928-3-9.Ch11

falso positivo? ¿Qué implicaciones tiene esto al elegir un umbral de la prueba?

9. Mencione dos ejemplos de pruebas diagnósticas dependientes del observador y explique por qué es necesario evaluar el acuerdo interobservador además de la exactitud.

10. Un artículo reporta una prueba con sensibilidad de 95 % y especificidad de 95 %, pero no describe cómo manejó los resultados indeterminados. ¿Qué sospecha debería tener el lector crítico y por qué?

RESPUESTAS:

1: b) Entre los pacientes con la enfermedad, ¿cuántos son identificados por la prueba como positivos?. La sensibilidad mide la capacidad de la prueba para detectar a los enfermos. Un valor alto reduce el riesgo de falsos negativos.

2: c) El 82 % de los positivos en la prueba realmente tienen tuberculosis. El VPP indica qué proporción de los resultados positivos corresponde a enfermedad real y depende de la prevalencia del contexto clínico.

3: b) Aplicar el estándar de referencia solo a quienes tuvieron prueba índice positiva. Este sesgo ocurre cuando no todos los pacientes reciben el estándar de referencia, lo que suele sobreestimar la sensibilidad.

4: b) Un paciente positivo es nueve veces más probable de estar enfermo que sano. Una LR+ alta aumenta de forma importante la probabilidad post-prueba y es útil para confirmar enfermedad.

5: b) Muestra cómo cambian sensibilidad y especificidad según el punto de corte. La curva ROC permite evaluar el rendimiento global de la prueba y elegir umbrales según el balance entre falsos positivos y negativos.

6: Porque los valores predictivos incorporan la proporción de enfermos y sanos en la población (prevalencia), mientras que la sensibilidad y especificidad son propiedades intrínsecas de la prueba frente al estándar.

7: Significa que, si el estudio se repitiera muchas veces, el 95 % de las estimaciones incluirían el valor real de la sensibilidad. El intervalo de confianza muestra la precisión de la estimación; intervalos amplios indican mayor incertidumbre.

8: Un falso negativo puede retrasar el tratamiento y favorecer contagio y progresión de la enfermedad. Por ello se prefieren pruebas y umbrales con alta sensibilidad, aceptando más falsos positivos si es necesario.

9: Radiografía de tórax y ecografía POCUS. Aunque una prueba sea exacta, si diferentes observadores no coinciden, su desempeño no será reproducible en la práctica real.

10: Posible sesgo de exclusión o manejo selectivo de resultados. Omitir indeterminados suele inflar artificialmente el desempeño de la prueba y limita la validez del estudio.

Fuente: Elaboración propia.

https://doi.org/10.62486/978-9915-9928-3-9.Ch11

REFERENCIAS

Bossuyt PMM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. BMJ. 2015;351:h5527. Available at: https://doi.org/10.1136/bmj.h5527.
Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: Explanation and elaboration. BMJ Open. 2016;6(11):e012799. Available at: https://doi.org/10.1136/bmjopen-2016-012799.
Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529–36. Available at: https://doi.org/10.7326/0003-4819-155-8-201110180-00009.
McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, Clifford T, et al. PRISMA-DTA: Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies. JAMA. 2018;319(4):388–96. Available at: https://doi.org/10.1001/jama.2017.19163.
Hanley JA, McNeil BJ. The meaning and use of the area under a ROC curve. Radiology. 1982;143(1):29–36. Available at: https://doi.org/10.1148/radiology.143.1.7063747.
Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1(8476):307–10. Available at: https://doi.org/10.1016/S0140-6736(86)90837-8.
Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155–63. Available at: https://doi.org/10.1016/j.jcm.2016.02.012.
Deeks JJ, Altman DG. Diagnostic tests 4: Likelihood ratios. BMJ. 2004;329(7458):168–9. Available at: https://doi.org/10.1136/bmj.329.7458.168.
Dorman SE, Schumacher SG, Alland D, Nabeta P, Armstrong DT, King B, et al. Xpert MTB/RIF Ultra for detection of Mycobacterium tuberculosis and rifampicin resistance: A prospective multicentre diagnostic accuracy study. Lancet Infect Dis. 2018;18(1):76–84. Available at: https://doi.org/10.1016/S1473-3099(17)30691-6.
Haider M, Yousaf S, Zaib A, Sarfraz A, Sarfraz Z, Cherrez-Ojeda I. Diagnostic accuracy of various immunochromatographic tests for NS1 antigen and IgM antibodies detection in acute dengue virus infection: A systematic review and meta-analysis. Int J Environ Res Public Health. 2022;19(14):8756. Available at: https://doi.org/10.3390/ijerph19148756.
Molière S, Hamzaoui D, Ploussard G, Mathieu R, Fiard G, Baboudjian M, et al. A systematic review of the diagnostic accuracy of deep learning models for the automatic detection, localization, and characterization of clinically significant prostate cancer on magnetic resonance imaging. EurUrol Oncol. 2025;8(4):1182–1202. Available at: https://doi.org/10.1016/j.euo.2024.11.001.
Krag CH, Müller FC, Gandrup KL, Plesner LL, Sagar MV, Andersen MB, et al. Impact of spectrum bias on deep learning-based stroke MRI analysis. Eur J Radiol. 2025;188:112161. Available at: https://doi.org/10.1016/j.ejrad.2025.112161.
Horne DJ, Zifodya JS, Shapiro AE, Church EC, Kreniske JS, Kay AW, et al. Xpert MTB/RIF Ultra assay for pulmonary tuberculosis and rifampicin resistance in adults and adolescents. Cochrane Database Syst Rev. 2025;7(7):CD009593. Available at: https://doi.org/10.1002/14651858.CD009593.pub6.
Arifin WN, Yusof UK. Correcting for partial verification bias in diagnostic accuracy studies: A tutorial using R. Stat Med. 2022;41(9):1709–27. Available at: https://doi.org/10.1002/sim.9311.
Stahlmann K, Reitsma JB, Zapf A. Missing values and inconclusive results in diagnostic studies: A scoping review of methods. Stat Methods Med Res. 2023;32(9):1842–55. Available at: https://doi.org/10.1177/09622802231192954.
Nahm FS. Receiver operating characteristic curve: Overview and practical use for clinicians. Korean J Anesthesiol. 2022;75(1):25–36. Available at: https://doi.org/10.4097/kja.21209.
Casanova IJ, Campos M, Juarez JM, Gomariz A, Lorente-Ros M, Lorente JA. Using the diagnostic odds ratio to select patterns to build an interpretable pattern-based classifier in a clinical domain: Multivariate sequential pattern mining study. JMIR Med Inform. 2022;10(8):e32319. Available at: https://doi.org/10.2196/32319.
Krobath DM, Naumova EN, Cuevas AG, Sacheck JM, Wilson NLW, Economos CD. Use of Bland-Altman analysis to examine the racial and ethnic representativeness of study populations in community-based pediatric health research. JAMA Netw Open. 2023;6(5):e2312920. Available at: https://doi.org/10.1001/jamanetworkopen.2023.12920.

Chapter 12 / Capítulo 12

Critical reading of scientific articles on health: A practical guide (Spanish Edition)
ISBN: 978-9915-9928-3-9
DOI: 10.62486/978-9915-9928-3-9.Ch12
Pages: 156-177

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Case-Control Studies

Estudios de casos y controles

Mauricio Alejandro Paz del Rio
Isaura Oberson-Santander
María Valeria Canedo-Sánchez

1 Clínica Los Andes. La Paz, Bolivia.
2 Hôpitaux Universitaires de Genève (HUG). Genève, Suisse.
3 Ministerio de Salud y Deportes. Dirección General del Sistema Nacional de Salud. Tarija, Bolivia.

ABSTRACT

Introduction: Case-control studies are an analytical observational design used to investigate the association between an exposure and a previously occurring outcome. Their usefulness is particularly relevant in rare or long-latency diseases, allowing for the efficient exploration of risk factors by comparing individuals with the disease (cases) and without it (controls). Development: This design uses the odds ratio (OR) as the primary measure of association, retrospectively evaluating the frequency of exposure in both groups. This chapter highlights fundamental methodological aspects such as the appropriate selection of cases and controls, the objective measurement of exposure, the control of confounding factors, and the identification of biases, including selection, information, and survival biases. It emphasizes the critical interpretation of the OR, confidence intervals, and analyses adjusted using logistic regression. Conclusions: Case-control studies represent a valuable tool for generating epidemiological evidence and formulating etiological hypotheses. However, the validity of their findings depends on the methodological rigor applied and a critical interpretation that considers the potential biases and limitations inherent in the design. Their appropriate use contributes to the development of preventive strategies and evidence-based decision-making.

Keywords: Case-Control Studies; Risk Factors; Odds Ratio; Epidemiological Biases; Evidence-Based Medicine

RESUMEN

Introducción: Los estudios de casos y controles constituyen un diseño observacional analítico orientado a investigar la asociación entre una exposición y un desenlace previamente ocurrido. Su utilidad es especialmente relevante en enfermedades raras o de larga latencia, permitiendo explorar factores de riesgo de manera eficiente mediante la comparación de individuos con la enfermedad (casos) y sin ella (controles). Desarrollo: Este diseño utiliza el odds ratio (OR) como medida principal de asociación, evaluando retrospectivamente la frecuencia de exposición en ambos grupos. El capítulo destaca aspectos metodológicos fundamentales como la adecuada selección de casos y controles, la medición objetiva de la exposición, el control de factores de confusión y la identificación de sesgos, entre ellos los de selección, información y supervivencia. De esta manera, enfatiza la interpretación crítica del OR, los intervalos de confianza y los análisis ajustados mediante regresión logística. Conclusiones: los estudios de casos y controles representan una herramienta valiosa para generar evidencia epidemiológica y formular hipótesis etiológicas. Sin embargo, la validez de sus hallazgos depende del rigor metodológico aplicado y de una interpretación crítica que considere los posibles sesgos y limitaciones inherentes al diseño. Su adecuada utilización contribuye al desarrollo de estrategias preventivas y a la toma de decisiones fundamentadas en evidencia científica.

Palabras clave: Estudios de Casos y Controles; Factores de Riesgo; Odds Ratio; Sesgos Epidemiológicos; Medicina Basada en la Evidencia

1. CASO CLÍNICO

En el hospital de tercer nivel de la ciudad de La Paz, el servicio de oncología identificó un incremento inusual en la frecuencia de diagnósticos de cáncer gástrico en pacientes menores de 55 años. Durante la discusión clínica de estos casos, se observó que varios de los pacientes procedían de zonas periurbanas donde el abastecimiento de agua para consumo humano dependía principalmente de pozos sin sistemas adecuados de tratamiento o filtración. Se presentó un paciente de 48 años, agricultor, con un cuadro clínico de seis semanas de evolución caracterizado por pérdida ponderal superior a 10 kg, dolor epigástrico urente persistente y episodios recurrentes de vómitos. La endoscopia digestiva alta evidenció una lesión ulcerada localizada en el antro gástrico, mientras que el estudio histopatológico de la biopsia confirmó el diagnóstico de adenocarcinoma gástrico.

Durante la anamnesis, el paciente refirió que el agua utilizada para consumo en su comunidad presentaba frecuentemente un "sabor metálico" característico. Asimismo, indicó que, aunque algunos habitantes realizaban procesos de ebullición previos a su consumo, esta práctica no se efectuaba de manera sistemática ni uniforme en toda la población.

A partir de estas observaciones, el equipo de salud formuló las siguientes preguntas de investigación:

- ¿Existe una asociación entre la exposición crónica a concentraciones elevadas de nitratos presentes en el agua de consumo y el desarrollo de cáncer gástrico?
- ¿Factores como el tabaquismo, el consumo de alcohol, los antecedentes familiares de cáncer gástrico u otras exposiciones ambientales podrían actuar como variables de confusión o modificadores del efecto de dicha asociación?

Con el propósito de explorar estas hipótesis, los investigadores identificaron un estudio de casos y controles diseñado para evaluar la posible relación entre la exposición a nitratos y el riesgo de cáncer gástrico.

Los casos estuvieron constituidos por pacientes con diagnóstico reciente de cáncer gástrico confirmado mediante estudio histopatológico, atendidos en el mismo centro hospitalario. Los controles correspondieron a individuos sin antecedente de cáncer gástrico ni patología digestiva mayor, procedentes de la misma población de referencia y seleccionados aleatoriamente en la comunidad. El objetivo principal fue determinar si la frecuencia de exposición a agua potencialmente contaminada con nitratos era mayor entre los casos que entre los controles y, en consecuencia, evaluar si dicha exposición podría contribuir a explicar el incremento observado de cáncer gástrico en adultos jóvenes de esta región.

2. OBJETIVOS

Los estudios de casos y controles tienen como finalidad principal explorar asociaciones entre una exposición y un desenlace. A diferencia de otros diseños, no buscan estimar directamente la incidencia o el riesgo absoluto, sino comparar la frecuencia de exposición en personas con la enfermedad (casos) frente a aquellas que no la presentan (controles).

De manera general, los objetivos de un estudio de casos y controles se pueden organizar en tres niveles:

- Determinar si existe una asociación estadísticamente significativa entre una exposición sospechosa y un desenlace.
- Medir dicha asociación a través del odds ratio (OR), que estima cuántas veces es más (o menos) probable que los casos hayan estado expuestos en comparación con los controles.

OBJETIVOS SECUNDARIOS

- Identificar factores de confusión que puedan alterar la relación entre exposición y desenlace, ajustando el OR mediante técnicas estadísticas (p. ej., regresión logística).
- Explorar modificadores de efecto, es decir, variables que cambian la magnitud de la asociación (por ejemplo, edad, sexo, tabaquismo, dieta).
- Evaluar si los hallazgos se mantienen consistentes en diferentes subgrupos de la población.

https://doi.org/10.62486/978-9915-9928-3-9.Ch12

OBJETIVOS ADICIONALES SEGÚN CONTEXTO

- Generar hipótesis etiológicas para futuras investigaciones de mayor nivel de evidencia (p. ej., cohortes o ensayos clínicos).
- Valorar la magnitud potencial del impacto en salud pública si la asociación fuese causal (ejemplo: agua contaminada y cáncer gástrico).
- Diseñar estrategias de prevención primaria o secundaria basadas en factores de riesgo modificables.

3. INTRODUCCIÓN

El estudio de casos y controles es un diseño observacional, analítico y generalmente retrospectivo, cuyo propósito es evaluar la relación entre una exposición y un desenlace. A diferencia de los estudios de cohortes, que siguen a los sujetos en el tiempo, en este diseño se parte del desenlace ya ocurridoy se investiga hacia atrás para conocer si los pacientes estuvieron expuestos o no al factor de interés.

Este tipo de estudio resulta especialmente útil en situaciones donde:

- El desenlace es infrecuente o tiene una larga latencia (por ejemplo, cáncer, enfermedades raras).
- La cohorte prospectiva sería inviable por requerir muchos años de seguimiento o grandes recursos.
- Se busca generar hipótesis etiológicas que puedan guiar investigaciones posteriores de mayor nivel de evidencia.

En términos metodológicos, los casos y controles comparan la frecuencia de exposición entre dos grupos:

- Casos: personas que presentan la enfermedad o desenlace de interés.
- Controles: personas libres de la enfermedad, seleccionadas de la misma población donde aparecieron los casos.

La medida de asociación característica de este diseño es el odds ratio (OR). Cuando la enfermedad es poco frecuente, el OR puede aproximar razonablemente al riesgo relativo, lo que se conoce como la rare diseaseassumption. Sin embargo, si la enfermedad no es rara, el OR tiende a sobreestimar la magnitud de la asociación, lo que debe considerarse al interpretar resultados. Ejemplo aplicado al caso clínico:En el hospital de La Paz, se observa un incremento inusual de cáncer gástrico en adultos jóvenes. Para investigar, se seleccionan como casos a pacientes diagnosticados recientemente con cáncer gástrico confirmado por biopsia, y como controles a sujetos sanos de la misma comunidad sin historia de neoplasias gástricas. El interés principal es determinar si la exposición a agua con alto contenido de nitratos está asociada al desarrollo de cáncer gástrico.Este ejemplo ilustra por qué el diseño de casos y controles es particularmente adecuado: el desenlace (cáncer gástrico en <55 años) es relativamente raro, el tiempo de latencia es prolongado y no sería factible ni eficiente esperar la aparición de nuevos casos mediante una cohorte prospectiva.

4. METODOLOGÍA: COMO LEER CRÍTICAMENTE UN ESTUDIO DE CASOS Y CONTROLES

Al enfrentarse a un artículo científico de casos y controles, la sección de Métodos es la columna vertebral sobre la cual descansa la validez del estudio. Un lector crítico no debe limitarse a comprender lo que hicieron los investigadores, sino que debe preguntarse si lo hicieron de la manera adecuada, si justificaron sus elecciones metodológicas y si esas decisiones permiten confiar en los resultados. La lectura crítica de esta sección implica recorrer una serie de elementos clave, cada uno con posibles fortalezas o riesgos de sesgo.

4.1 POBLACIÓN FUENTE Y PERIODO DE ESTUDIO

El primer aspecto que debe revisarse es la definición de la población fuente. Un estudio de casos y controles es válido solo si los casos y los controles provienen de la misma población en riesgo. Por ello, es necesario que el artículo especifique con claridad cuál fue el marco poblacional y en qué periodo temporal se realizó la investigación. Cuando esta información no está bien delimitada, los resultados pierden aplicabilidad y pueden generar conclusiones engañosas.

https://doi.org/10.62486/978-9915-9928-3-9.Ch12

En nuestro caso clínico, por ejemplo, si los investigadores de La Paz no precisaran que tanto casos como controles provienen de las zonas periurbanas abastecidas por pozos de agua sin filtración, el lector debería cuestionar de inmediato la comparabilidad de ambos grupos.

4.2 DEFINICIÓN DE CASO

Una vez identificada la población, corresponde examinar cómo fueron definidos los casos. La pregunta crítica es si los autores utilizaron criterios diagnósticos estandarizados, válidos y reproducibles. Además, conviene verificar si los casos fueron incidentes o prevalentes.

- Los casos incidentes, son aquellos pacientes diagnosticados recientemente, dentro de un periodo definido por el estudio. Su inclusión permite acercarse de manera más fiel a la secuencia causal exposición–enfermedad, y además reduce el riesgo del llamado sesgo de supervivencia o sesgo de Neyman que aparece cuando solo se incluyen pacientes que han sobrevivido lo suficiente para ser detectados, lo que excluye a quienes murieron tempranamente o tuvieron formas más agresivas de la enfermedad.
- Casos prevalentes son los pacientes que ya presentaban la enfermedad antes del inicio del estudio, sin importar cuándo ocurrió el diagnóstico. Aunque suelen ser más fáciles de reclutar —pues basta con identificar a quienes ya están en seguimiento o en registros hospitalarios—, su uso tiene limitaciones importantes. Al incluir pacientes con distinta duración de enfermedad, se corre el riesgo de seleccionar preferentemente a aquellos con formas más benignas o de curso lento, lo que introduce un sesgo que tiende a infraestimar la fuerza de la asociación. Además, en los casos prevalentes puede resultar más difícil establecer con precisión la temporalidad entre la exposición y la aparición de la enfermedad, ya que la exposición pudo haber cambiado después del inicio de la patología.

En el ejemplo de cáncer gástrico en La Paz, un estudio de buena calidad debería trabajar con casos incidentes, es decir, pacientes recién diagnosticados y confirmados histológicamente en los últimos meses o años del periodo de estudio. Si, en cambio, se incluyeran como casos a pacientes que llevan más de cinco años conviviendo con el cáncer, la muestra podría estar compuesta por sobrevivientes con características distintas (mayor acceso a tratamiento, tumores menos agresivos), lo que sesgaría la interpretación de la relación con la exposición a nitratos.

Tabla 1: Comparación entre casos incidentes y casos prevalentes en estudios de casos y controles: definiciones, ventajas y limitaciones.

Tipo de caso	Definición	Ventajas	Desventajas / Sesgos
Casos incidentes	Pacientes diagnosticados recientemente, durante el periodo definido por el estudio.	- Reflejan mejor la secuencia causal exposición–enfermedad.	- Requieren vigilancia activa y mayor tiempo de reclutamiento.
		- Reducen el riesgo de sesgo de supervivencia (Neyman).	- Pueden ser más costosos de identificar.
		- Permiten analizar factores asociados al inicio de la enfermedad.	
Casos prevalentes	Pacientes que ya presentaban la enfermedad antes del inicio del estudio, independientemente de cuándo ocurrió el diagnóstico.	- Más fáciles de reclutar (hospitales, registros).	- Mayor riesgo de sesgo de supervivencia: sobrerrepresentan casos de evolución larga y subrepresentan los de curso letal o rápido.
		- Útiles para enfermedades crónicas de larga duración.	- Dificultan establecer la temporalidad entre exposición y enfermedad.
			- Pueden infraestimar o distorsionar la fuerza de la asociación.

Fuente: Elaboración propia

4.3 DEFINICIÓN DE CONTROL

En paralelo, es fundamental evaluar la selección de los controles, probablemente el punto más crítico de este diseño. Los controles deben representar de forma fiel la población de

https://doi.org/10.62486/978-9915-9928-3-9.Ch12

donde surgieron los casos, es decir, aquellos que podrían haber desarrollado la enfermedad si hubiesen estado expuestos. Cuando los investigadores seleccionan controles hospitalarios con otras enfermedades, el riesgo de sesgo de selección aumenta, especialmente por el conocido sesgo de Berkson: los controles hospitalizados suelen tener exposiciones distintas a la población general, lo cual distorsiona las asociaciones. Volviendo a nuestro caso clínico, sería mucho más válido que los controles se eligieran al azar en la comunidad de origen de los pacientes, y no entre enfermos hospitalizados, ya que, por ejemplo, si se elige enfermos renales, también pueden estar relacionados con exposición a nitratos.

Tabla 2. Comparación entre controles hospitalarios y controles poblacionales en estudios de casos y controles: definiciones, ventajas y principales sesgos.

Tipo de control	Definición	Ventajas	Desventajas / Sesgos principales
Controles hospitalarios	Pacientes seleccionados de hospitales o centros de salud, que no presentan la enfermedad de interés.	- Más fáciles y económicos de reclutar.	- Sesgo de Berkson: los pacientes hospitalarios no representan a la población general y pueden tener exposiciones asociadas a su motivo de hospitalización.
		- Frecuentemente más dispuestos a participar (ya están en contacto con el sistema sanitario).	- Pueden compartir factores de riesgo con los casos (ej. consumo de agua contaminada si provienen de la misma comunidad).
		- Mejor calidad de información clínica (historias médicas disponibles).	- Limitan la validez externa.
Controles poblacionales	Individuos seleccionados de la misma comunidad o área geográfica que generó los casos, sin la enfermedad de interés.	- Representan de manera más fiel la población en riesgo.	- Más costosos y complejos de reclutar (encuestas, registros comunitarios).
		- Mayor validez externa.	- Menor disponibilidad de información clínica.
		- Menor riesgo de sesgos relacionados con enfermedades distintas a la de estudio.	- Pueden tener menor tasa de participación, generando sesgo si quienes aceptan difieren de quienes rechazan.

Fuente: Elaboración propia

4.4 TIPO DE MUESTREO EMPLEADO

Otro elemento central es el tipo de muestreo empleado. No todos los estudios de casos y controles se diseñan de la misma manera, y esto determina qué parámetro puede estimarse de forma válida. Existen diferentes tipos de muestreo

- En los diseños de tipo acumulativo, los controles se seleccionan al final del periodo de estudio y el odds ratio obtenido puede aproximar el riesgo relativo únicamente cuando la enfermedad es rara.
- En los diseños de muestreo por densidad de incidencia, los controles se eligen en paralelo a la aparición de cada caso y, en ese contexto, el odds ratio estima de manera directa una razón de tasas.
- En el diseño case-cohort permite aproximarse al riesgo relativo cuando se conoce bien la base de población.

Un lector crítico debe identificar cuál de estos métodos fue usado y si los autores lo explicaron con claridad. En el ejemplo del cáncer gástrico, un artículo sólido debería indicar que se utilizó muestreo por densidad, seleccionando controles en el mismo periodo en que aparecían los casos, lo que asegura comparabilidad en el tiempo.

4.5 MEDICIÓN DE EXPOSICIÓN

La medición de la exposición es otro punto neurálgico. El lector debe preguntarse: ¿cómo se midió la exposición?, ¿se emplearon cuestionarios, entrevistas, registros médicos o pruebas objetivas? El sesgo de recuerdo es un problema frecuente en este tipo de estudios, ya que los pacientes con la enfermedad tienden a recordar o reportar la exposición con mayor detalle que los controles. Además, cuando los investigadores saben qué participante es caso o control al

https://doi.org/10.62486/978-9915-9928-3-9.Ch12

momento de medir la exposición, aparece el sesgo del observador. Por esta razón, se valoran más los estudios que utilizan métodos objetivos y ciegos.

En el caso clínico propuesto, la diferencia entre un artículo débil y uno sólido es clara: no es lo mismo preguntar al paciente si cree que el agua estaba contaminada, que medir de manera estandarizada la concentración de nitratos en muestras de agua del domicilio de cada participante.

4.6 FACTORES DE CONFUSIÓN

Otro componente clave en la lectura crítica es la gestión de los factores de confusión que distorsiona la asociación entre la exposición y el desenlace (como edad o sexo). Un estudio válido debe identificar de manera explícita los factores de confusión. El lector debe valorar si los autores seleccionaron confusores de forma justificada, por ejemplo, utilizando diagramas causales (DAGs), y no únicamente en función de pruebas estadísticas superficiales. Además, conviene analizar si se aplicó pareamiento y, de ser así, si este se manejó de manera correcta en el análisis. El pareamiento por edad, sexo o lugar de residencia puede ser útil para equilibrar grupos, pero cuando se realiza en exceso —el llamado overmatching— puede borrar la asociación real.

En el caso del cáncer gástrico, un estudio bien diseñado podría parear casos y controles por edad y sexo, ajustando luego en el análisis por tabaquismo, alcoholismo y dieta salada, factores reconocidos como potenciales confusores (porque igual podrían producir cáncer gástrico).

4.7 TAMAÑO MUESTRAL

La sección de metodología también debe explicar cómo se determinó el tamaño muestral. Un lector crítico debe comprobar si se calculó previamente el número de casos y controles necesarios para obtener resultados con suficiente poder estadístico. La ausencia de un cálculo muestral puede indicar un riesgo de resultados poco confiables. En el ejemplo de La Paz, sería tranquilizador leer que el estudio planificó incluir 150 casos y 300 controles (relación 1:2) para detectar un odds ratio de al menos 2 con un poder del 80% y un nivel de confianza del 95%.

4.8 PLAN DE ANÁLISIS

Finalmente, el plan de análisis debe estar claramente detallado. El lector debe confirmar si los autores explicaron de antemano cómo calcularon los odds ratios, cómo ajustarían por confusión, si explorarían interacciones o modificación de efecto y si consideraron análisis de sensibilidad, como el cálculo de un E-value para estimar la posible influencia de confusión no medida. La transparencia en el plan analítico ayuda a prevenir la manipulación de resultados y otorga credibilidad al artículo.

En el caso clínico, un estudio sólido informaría que se calculará primero un odds ratio crudo, seguido de un modelo de regresión logística ajustado por tabaquismo, alcoholismo y dieta, además de un análisis estratificado para explorar si el tabaco modifica el efecto de la exposición a nitratos.

La lectura crítica de la metodología en un estudio de casos y controles exige un recorrido atento y sistemático. Desde la definición de la población y de los casos, hasta la elección de los controles, la medición de la exposición y el manejo de confusores, cada detalle puede marcar la diferencia entre un estudio confiable y uno sesgado. El lector que aprende a formular estas preguntas logra no solo interpretar con precisión los resultados, sino también valorar si estos pueden o no aplicarse en la práctica clínica.

5. EVALUACIÓN DE LA VALIDEZ Y RELEVANCIA

Una vez entendida la metodología, el siguiente paso en la lectura crítica de un estudio de casos y controles es evaluar si los resultados son válidos y si tienen relevancia clínica y epidemiológica. La validez se divide en dos grandes dimensiones: validez interna, que responde a la pregunta “¿los resultados del estudio son confiables dentro de la población analizada?”, y validez externa, que pregunta “¿estos resultados pueden aplicarse a otros contextos o poblaciones?”.

https://doi.org/10.62486/978-9915-9928-3-9.Ch12

5.1 VALIDEZ INTERNA

Para juzgar la validez interna, el lector debe analizar si el estudio evitó los principales sesgos y si controló de manera adecuada la confusión. En un diseño de casos y controles, los sesgos más frecuentes se verán en el acápite 7.La validez interna se refuerza también con decisiones metodológicas como el uso de casos incidentes en lugar de prevalentes, la aplicación de análisis estratificados para verificar la consistencia de los resultados y la realización de análisis de sensibilidad (por ejemplo, calculando un E-value para explorar el posible impacto de confusores no medidos).

5.2 VALIDEZ EXTERNA

Una vez establecida la confianza en los resultados dentro del estudio, corresponde preguntarse si esos hallazgos pueden extrapolarse a otras poblaciones. La validez externa depende de cuán representativa sea la población fuente y de cuán similares sean los determinantes de salud en otros contextos.En el ejemplo, un estudio realizado en comunidades periurbanas de La Paz puede ser muy válido para entender la relación entre nitratos y cáncer gástrico en esa región, pero el lector debe cuestionar si los hallazgos pueden aplicarse a zonas rurales donde el agua proviene de manantiales de montaña, o a países donde la dieta y las fuentes de exposición a nitratos son diferentes. No es lo mismo interpretar un estudio en una comunidad de altura, con costumbres alimentarias específicas, que en una ciudad costera de Asia donde el consumo de pescados salados y nitrosaminas es mucho mayor.

5.3 RELEVANCIA CLÍNICA Y EPIDEMIOLÓGICA

Más allá de la validez, un lector crítico debe evaluar si la asociación encontrada tiene significado clínico y de salud pública. Un odds ratio elevado pero basado en un estudio pequeño, con intervalos de confianza muy amplios, puede carecer de utilidad práctica. Por el contrario, una asociación moderada pero consistente en varios estudios, ajustada por confusores y replicada en diferentes poblaciones, puede ser altamente relevante para la prevención.En el caso del cáncer gástrico, demostrar una asociación robusta entre exposición a nitratos y riesgo de enfermedad podría tener repercusiones inmediatas en políticas de salud, tales como la regulación de calidad del agua potable, campañas de educación sanitaria y vigilancia ambiental. Por ello, al leer un artículo de casos y controles, el lector debe pensar siempre más allá del valor estadístico: debe preguntarse si los hallazgos pueden contribuir a mejorar la práctica clínica o las decisiones de salud pública.

6. INTERPRETACIÓN ESTADÍSTICA

La interpretación estadística en casos y controles comienza aceptando un límite fundamental del diseño: no estima directamente riesgos absolutos, tasas de incidencia ni diferencias de riesgo. Su lenguaje nativo es la razón de momios u odds ratio (OR). Por eso, leer bien un artículo significa comprobar que lo que se informa y lo que se concluye sí es coherente con lo que el diseño puede ofrecer.

6.1 ANTES DE MIRAR NÚMEROS ¿QUÉ SI Y QUE NO PUEDE INTERPRETARSE?

Un estudio de casos y controles compara cuán frecuente fue la exposición entre quienes desarrollaron la enfermedad (casos) y quienes no (controles). Si la enfermedad es rara en la población, el OR puede aproximar el riesgo relativo (RR). Si no es rara, el OR tiende a sobreestimar la magnitud del efecto y no debe traducirse como “veces más riesgo” sin aclaraciones. La lectura crítica exige verificar que el artículo no presente RR ni riesgo absoluto como si se hubieran medido directamente.

6.2 DE “ODDS” A “ODDS RATIO”: EL MÍNIMO QUE SE DEBE DOMINAR

El odds (momio) de exposición en un grupo es la razón entre “expuestos” y “no expuestos”

https://doi.org/10.62486/978-9915-9928-3-9.Ch12

dentro de ese grupo. El odds ratio (OR) compara esos momios entre casos y controles. Se interpreta así:

- "En los casos, la probabilidad relativa de haber estado expuesto, en comparación con los controles, fue X veces mayor (o menor)."

Ejemplo: Siguiendo con el caso, se obtuvieron los siguientes resultados:

Tabla 3. Distribución de casos y controles según exposición a nitratos en el caso clínico (tabla 2×2).

	Expuestos a nitratos	No expuestos	Total
Casos (cáncer gástrico)	40	10	50
Controles	20	30	50

Fuente: Elaboración propia con apoyo de Inteligencia Artificial

ODDS RATIO ELEVADO

Cuando el OR es mayor a 1, indica que la exposición fue más frecuente entre los casos que entre los controles.

- Un OR moderado, por ejemplo 1.5 o 2.0, sugiere que los casos tuvieron entre 1.5 y 2 veces más probabilidad de estar expuestos que los controles.
- Un OR alto, como 5.0 o más, indica una asociación fuerte. Sin embargo, el lector debe recordar que un OR grande no siempre implica un efecto clínicamente importante si los intervalos de confianza son muy anchos o si hay sesgos en la selección o medición.

Tabla 4. Odds Ratio de exposición a nitratos.

Exposición	OR (IC95%)	Valor p
Agua con nitratos	6.0 (2.5–14.7)	<0.001

Fuente: Elaboración propia con apoyo de Inteligencia Artificial

Ejemplo del caso clínico: un OR de 6.0 para exposición a nitratos significa que los pacientes con cáncer gástrico tenían seis veces más probabilidad de haber estado expuestos a agua con alto nitrato respecto de los controles. Esto es una asociación fuerte y, si el intervalo de confianza es estrecho, sugiere una señal robusta. Lo que no debe decirse: "Seis veces más riesgo de cáncer" (salvo que se justifique enfermedad rara y diseño compatible).

ODDS RATIO BAJO

Cuando el OR es menor a 1, se interpreta como que la exposición es menos frecuente en los casos que en los controles. En términos clínicos, puede sugerir un efecto protector.

- Un OR de 0.7 significa que la exposición se observó 30% menos entre los casos que entre los controles.
- Un OR de 0.3 indicaría una asociación protectora más marcada.

Tabla 5. Odds Ratio de consumo de frutas.

Exposición	OR (IC95%)	Valor p
Consumo de frutas	0.4 (0.2–0.9)	0.03

Fuente: Elaboración propia con apoyo de Inteligencia Artificial

Ejemplo aplicado: si el estudio encontrara que el consumo habitual de frutas frescas tiene un OR de 0.4 frente al cáncer gástrico, esto sugeriría un posible efecto protector de la dieta, siempre que el intervalo de confianza no cruce 1.0.

¿QUÉ PASA CUANDO EL OR ESTÁ EN 1?

El valor clave de referencia en la interpretación es 1.0, que representa "no asociación" (los odds de exposición son iguales en casos y controles).

- Si el IC95% incluye el 1.0, significa que no puede descartarse el azar como explicación de los

https://doi.org/10.62486/978-9915-9928-3-9.Ch12

resultados. En ese caso, la asociación no es estadísticamente significativa, aunque el OR sea mayor o menor que 1.

- Si el IC es muy amplio, además, la estimación es poco precisa: no se puede afirmar con confianza si existe o no una relación.

Tabla 6. Odds Ratio de exposición a tabaco.

Exposición	OR (IC95%)	Valor p
Tabaquismo	1.3 (0.8–2.2)	0.03

Fuente: Elaboración propia con apoyo de Inteligencia Artificial

Ejemplo aplicado: supongamos que, al analizar el hábito tabáquico, el estudio reporta un OR de 1.3 con un IC95% de 0.8 a 2.2. El valor puntual (1.3) sugiere una ligera mayor probabilidad de exposición en casos, pero como el intervalo incluye 1.0, la asociación no es estadísticamente significativa. En lectura crítica, esto se traduce en "no hay evidencia suficiente para afirmar una relación entre el tabaco y cáncer gástrico".

Reglas de oro:

- OR > 1 (IC95% excluye 1.0): la exposición podría estar asociada a mayor probabilidad de enfermedad (posible factor de riesgo).
- OR < 1 (IC95% excluye 1.0): la exposición podría estar asociada a menor probabilidad de enfermedad (posible factor protector).
- OR ≈ 1 o IC95% incluye 1.0: no se demuestra asociación; puede deberse a azar, sesgo, confusión o falta de poder estadístico.

6.3 INTERVALOS DE CONFIANZA Y VALOR P: SEÑAL, PRECISIÓN Y AZAR

El intervalo de confianza (IC) es, probablemente, el recurso más valioso para interpretar los resultados de un estudio de casos y controles. Mientras que el OR nos da una estimación puntual de la asociación, el IC nos dice qué tan precisa es esa estimación y en qué rango de valores se podría mover la "verdadera asociación" en la población.

El valor p, por su parte, indica la probabilidad de obtener los resultados observados (o más extremos) si en realidad no hubiera asociación. Sin embargo, el valor p no mide magnitud ni relevancia clínica; simplemente ayuda a decidir si el resultado podría explicarse por azar. Por eso, la lectura crítica debe priorizar el IC por encima del p, aunque ambos se complementen.

Ejemplo con caso clínico: En el estudio de cáncer gástrico y nitratos, se presenta la siguiente tabla

Tabla 7: Odds Ratio en diferentes escenarios de exposición a nitratos.

Exposición	OR (IC95%)	Valor p
Agua con nitratos	6.0 (2.5–14.7)	<0.001
Tabaquismo	1.3 (0.8–2.2)	0.01
Consumo de frutas	0.4 (0.2–0.9)	0.03

Fuente: Elaboración propia

El OR crudo de agua con nitratos fue de 6.0.

- Si el IC95% es 2.5 a 14.7, esto significa que, con un 95% de confianza, la verdadera asociación se encuentra en algún punto entre 2.5 y 14.7.
- Como el IC no incluye 1.0, el resultado es estadísticamente significativo.
- El intervalo es amplio, lo que refleja incertidumbre en la magnitud exacta: ¿el efecto real es 2.5 veces o 14 veces mayor? Ambos valores son posibles.

El valor p (<0.001) refuerza que el resultado es muy poco compatible con el azar, pero lo más importante sigue siendo el rango del IC.

En los todos los escenarios se interpretaría de esta forma:

- OR = 6.0, IC95% 2.5–14.7, asociación fuerte y significativa.El efecto existe, pero su magnitud exacta es incierta.
- OR = 1.3, IC95% 0.8–2.2, el OR sugiere ligera asociación positiva, pero como el IC incluye 1.0, no es significativa.Interpretación crítica: "No se encontró evidencia suficiente de asociación".
- OR = 0.4, IC95% 0.2–0.9, asociación negativa significativa, posible efecto protector.Interpretación crítica: el consumo frecuente de frutas es un posible factor protector contra el cáncer

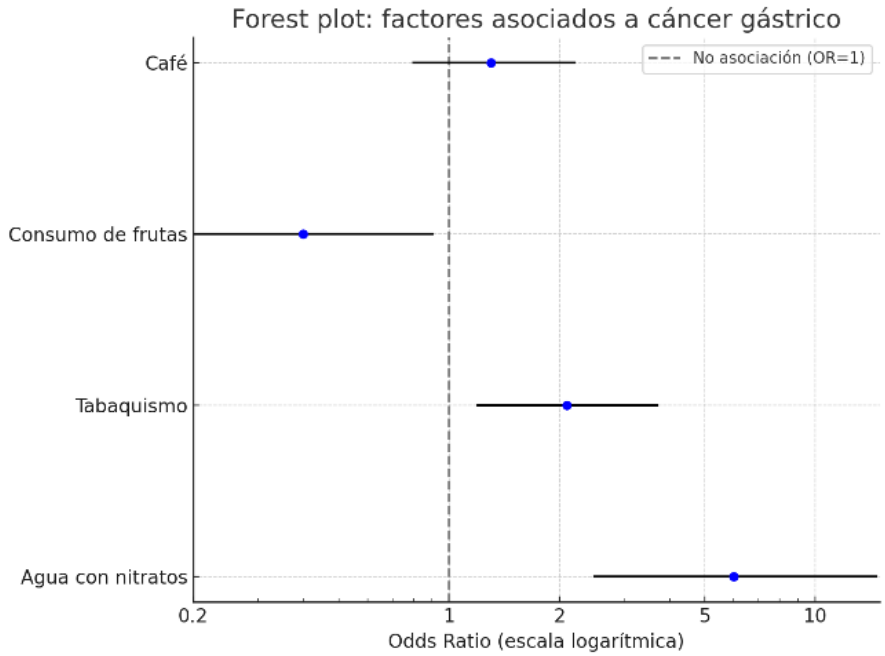
https://doi.org/10.62486/978-9915-9928-3-9.Ch12

gástrico.
- Todos cuentan con el valor p (<0.05) que refuerza que los resultados son poco compatibles con el azar.

6.4 INTERPRETACIÓN DE FOREST PLOT (GRÁFICO DE BOSQUE):

Representa visualmente cada OR como un punto (●) y su IC95% como una línea horizontal.
- La línea vertical en 1.0 representa "no asociación".
- Si el intervalo cruza esa línea, el resultado no es significativo.
- Cuanto más corta la línea, más precisa la estimación.

Figura 1 Forest plot de los factores asociados al riesgo de cáncer gástrico en el caso clínico.



Fuente: Elaboración propia

Para su interpretaciónde un Forest plot basado en nuestro caso clínico.
- Eje X: representan el OR en escala logarítmica
- Eje Y: las variables de intervención a estudiar como posible factor de riesgo (agua con nitratos, tabaquismo, consumo de frutas, café)
- La línea vertical punteada en OR = 1.0 (sin asociación).
- Cada punto azul representa el OR puntual.
- Las barras negras muestran los rangos del IC95%.
- A la derecha de línea punteada 1.0 = asociaciones positivas (factores de riesgo).
- A la izquierda de línea punteada 1.0 = asociaciones negativas (posibles factores protectores).

Ejemplo del Forest plot:Su interpretación seria de esta forma:
- Agua con nitratos: ● muy a la derecha, con línea 2.5–14.7 → asociación fuerte y significativa.
- Tabaquismo: ● también a la derecha, línea corta 1.2–3.7 → asociación moderada y precisa.
- Consumo de frutas: ● a la izquierda, línea 0.2–0.9 → efecto protector significativo.
- Café: ● cerca del 1, línea 0.8–2.2 → no significativo.

Reglas prácticas para la interpretación:
- El OR puntual da una idea del efecto, pero nunca debe leerse aislado.
- El IC95% indica precisión:Intervalo estrecho = estimación confiable, intervalo amplio = mucha incertidumbre.
- El 1.0 dentro del IC = resultado no significativo.
- El valor p solo complementa: nunca sustituye al IC.

https://doi.org/10.62486/978-9915-9928-3-9.Ch12

6.5 CAMBIOS DEL OR CUANDO SE AJUSTA

Cuando un artículo reporta un OR ajustado, significa que la asociación entre una exposición y una enfermedad fue estimada controlando estadísticamente otras variables que podrían distorsionar el resultado, en otras palabras, el OR ajustado intenta responder a la pregunta:

"¿Cuál es la asociación entre la exposición y la enfermedad si todos los grupos comparados tuvieran características similares en edad, sexo, tabaquismo u otros factores relevantes?"

"Ajustar" no significa modificar artificialmente los datos. Significa que el análisis estadístico —habitualmente mediante regresión logística en estudios de casos y controles— estima el efecto de cada exposición manteniendo constantes las demás variables incluidas en el modelo.

Por ejemplo, si el consumo de agua con nitratos está asociado con cáncer gástrico, pero las personas expuestas también fuman más, parte del efecto observado en el OR crudo podría deberse al tabaquismo, por ello el OR ajustado permite estimar la asociación independiente de los nitratos, descontando el efecto del tabaco y otras variables.

En estudios bien hechos, los autores describen los confusorespreespecificados (edad, sexo, tabaquismo, alcoholismo, dieta salada, *H. pylori*, nivel socioeconómico) y muestran el OR ajustado.

Tabla 8. Odds Ratios crudos y ajustados de factores de riesgo para cáncer gástrico en el caso clínico.

Exposición	OR crudo (IC95%)	OR ajustado (IC95%)
Agua con nitratos	6.0 (2.5–14.7)	3.2 (1.8–5.6)
Tabaquismo	2.5 (1.3–4.8)	1.9 (1.1–3.4)
Consumo de alcohol	1.8 (1.0–3.2)	1.4 (0.8–2.6)

Fuente: Elaboración propia

Interpretación:

- Agua con nitratos: el OR crudo de 6.0 sugería una asociación muy fuerte. Tras ajustar por tabaquismo y alcohol, baja a 3.2. Esto indica que parte del efecto se explicaba por confusión, pero la asociación persiste, y es significativa.
- Tabaquismo: también disminuye levemente tras el ajuste, pero sigue siendo significativo.
- Consumo de alcohol: el OR crudo parecía mostrar riesgo, pero tras el ajuste ya no es significativo (IC95% incluye 1.0). Esto muestra cómo el ajuste puede cambiar la conclusión.

6.6 ESTRATIFICACIÓN Y MANTEL–HAENSZEL (IDEA INTERPRETATIVA)

En algunos estudios de casos y controles, los autores no se limitan a presentar un Odds Ratio global. En lugar de ello, dividen el análisis en subgrupos o estratos, por ejemplo fumadores y no fumadores, hombres y mujeres, o distintos grupos etarios. Este procedimiento se denomina estratificación y tiene un objetivo claro: explorar si una tercera variable está confundiendo la asociación principal o si, por el contrario, está modificando el efecto de la exposición.

Para comprenderlo mejor, retomemos el caso clínico del cáncer gástrico y la exposición a agua con nitratos. Supongamos que el OR crudo fue de 6.0. El lector podría concluir inicialmente que la exposición aumenta seis veces la probabilidad de enfermedad. Sin embargo, si el tabaquismo está asociado tanto con el consumo de agua contaminada como con el cáncer gástrico, podría estar inflando esa estimación (tabla 8).

Si el análisis se realiza por separado en fumadores y no fumadores, pueden aparecer dos escenarios distintos.

- En el primer escenario, los OR en fumadores y en no fumadores resultan muy similares entre sí, por ejemplo 3.1 y 3.3. En este caso, el efecto es consistente en ambos grupos, pero claramente menor que el OR crudo de 6.0. Esto indica que el tabaquismo estaba distorsionando la asociación inicial. Al separar los datos por estratos, el efecto real se aproxima a 3.2. Aquí se habla de confusión. El método de Mantel–Haenszel permite entonces combinar esos OR estratificados similares en una única estimación ajustada, obteniendo un OR "combinado" que refleja la asociación controlando esa variable.
- En el segundo escenario, los resultados por estrato son muy diferentes entre sí, por ejemplo un OR de 8.0 en fumadores y de 1.5 en no fumadores. Aquí el efecto no es uniforme: la exposición parece mucho más dañina en fumadores que en no fumadores. Esto ya no es confusión, sino modificación de efecto o interacción. En esta situación no tiene sentido combinar los OR, porque no existe un único efecto verdadero; el impacto de la exposición

https://doi.org/10.62486/978-9915-9928-3-9.Ch12

depende del subgrupo. Desde el punto de vista clínico, esto es incluso más relevante, pues sugiere que ciertos pacientes son particularmente vulnerables.

El método de Mantel–Haenszel se utiliza precisamente cuando los OR por estrato son relativamente homogéneos. Antes de combinar los resultados, el investigador debe verificar que las estimaciones no difieran de manera importante entre sí. Si son similares, puede calcularse un OR ajustado global. Si son muy distintos, se debe reportar el efecto por separado para cada estrato.

Para el lector crítico, la interpretación práctica es la siguiente: si los OR estratificados son parecidos entre sí pero distintos del OR crudo, probablemente existe confusión. Si los OR estratificados difieren marcadamente entre sí, probablemente existe modificación de efecto. Reconocer esta diferencia es esencial para no interpretar de forma errónea los resultados y para comprender si una variable está distorsionando la asociación o realmente está cambiando su magnitud en determinados subgrupos.

6.7 REGRESIÓN LOGÍSTICA

En casos y controles, la regresión logística sirve para ajustar confusores y explorar interacciones. El lector debe fijarse en:

- Qué variables se ajustaron y por qué (mejor si lo justifican con diagramas causales; peor si "entró lo que dio p<0,05").
- OR ajustados con IC95% (no basta el p).
- Aclaraciones sobre no colapsabilidad: el OR puede cambiar al ajustar incluso sin confusión; es propiedad matemática del OR y no implica necesariamente "sesgo".
- Medidas de desempeño: evitar interpretar pseudoR^2 como "% de varianza explicada" (no lo es). Es preferible que el artículo muestre calibración (qué tan bien encajan probabilidades predichas con observadas) y, si construye un modelo predictivo, que informe discriminación (AUC/ROC). En estudios puramente etiológicos, AUC es secundaria; lo crucial es el OR ajustado bien interpretado.

Para entenderlo se utilizará la siguiente tabla:

Tabla 9. Resultados de la regresión logística multivariada: factores de riesgo y protección asociados a cáncer gástrico.

Variable	OR crudo (IC95%)	Valor p	OR ajustado (IC95%)	Valor p
Agua con nitratos	6.0 (2.5–14.7)	<0.001	3.2 (1.8–5.6)	<0.001
Tabaquismo	2.5 (1.3–4.8)	0.01	1.9 (1.1–3.4)	0.02
Consumo de alcohol	1.8 (1.0–3.2)	0.05	1.4 (0.8–2.6)	0.18
Consumo de frutas	0.5 (0.3–0.9)	0.03	0.6 (0.3–1.1)	0.08

Fuente: Elaboración propia

INTERPRETACIÓN PASO A PASO:

- Agua con nitratos: el OR crudo de 6.0 sugería una asociación fuerte; tras ajustar por tabaquismo, alcohol y frutas, disminuye a 3.2, pero sigue siendo significativa. Esto significa que parte de la asociación se debía a confusión, pero la exposición sigue mostrando un efecto robusto.
- Tabaquismo: mantiene significancia tras el ajuste, aunque con menor magnitud.
- Consumo de alcohol: pierde significancia al ajustar, lo que indica que su aparente efecto estaba explicado por otros factores.
- Consumo de frutas: parecía protector en el análisis crudo, pero tras el ajuste la asociación ya no alcanza significancia estadística.

Consejo para el lector critico: Al leer una tabla de regresión logística, lo más importante es comparar OR crudo vs OR ajustado. Si el OR cambia de manera relevante después del ajuste, es señal de confusión. Si se mantiene estable, la asociación es más consistente. Y si pierde significancia, probablemente la relación observada era espuria.

https://doi.org/10.62486/978-9915-9928-3-9.Ch12

Tabla 10. Estratificación (Mantel–Haenszel) vs regresión logística en estudios de casos y controles

Aspecto	Estratificación / Mantel–Haenszel	Regresión logística
¿Qué hace?	Divide el análisis en subgrupos (estratos) según una variable potencialmente confusora (ej. fumadores vs no fumadores).	Analiza simultáneamente varias variables en un modelo matemático para estimar el efecto independiente de cada una.
¿Qué problema intenta resolver?	Identificar confusión o modificación de efecto.	Controlar múltiples confusores y estimar OR ajustados.
¿Cuándo se usa?	Cuando hay una o pocas variables importantes que pueden confundir la asociación.	Cuando existen varias variables que deben ajustarse al mismo tiempo.
¿Cómo detecta confusión?	Si los OR por estrato son similares entre sí pero diferentes del OR crudo.	Si el OR ajustado difiere sustancialmente del OR crudo.
¿Cómo detecta modificación de efecto?	Si los OR por estrato son muy distintos entre sí.	Si el término de interacción es significativo en el modelo.
Resultado final	OR combinado de Mantel–Haenszel (si hay homogeneidad entre estratos).	OR ajustado multivariado.
Ventaja principal	Método simple y transparente; fácil de entender visualmente.	Permite ajustar múltiples variables simultáneamente y evaluar interacciones complejas.
Limitación	Se vuelve impráctico con muchas variables.	Requiere mayor conocimiento estadístico y supuestos del modelo.

Fuente: Elaboración propia

6.8 MODIFICACIÓN DEL EFECTO (INTERACCIÓN): CUANDO LA ASOCIACIÓN "DEPENDE DE..."

No todo cambio del OR entre subgrupos es confusión; a veces el efecto realmente varía. Si el OR de nitratos es 2.0 en no fumadores y 6.0 en fumadores, el tabaquismo podría potenciar el efecto de los nitratos. El artículo debe declararlo como interacción y, idealmente, mostrar un término de interacción en el modelo (interpretación clínica: el riesgo atribuible a la exposición no es igual para todos).

Lectura práctica:

- Si los OR por estratos son parecidos → combinar (confusión).
- Si son claramente distintos → considerar modificación de efecto y reportar por subgrupos.

6.9 ESTUDIOS PAREADOS: COMO LEER EL OR "CONDICIONAL"

Cuando los autores parean casos y controles (p. ej., por edad y sexo), el análisis correcto no usa el 2×2 simple, sino el OR de pares discordantes o una regresión logística condicional.

Ejemplo: imaginemos que, en el estudio de cáncer gástrico y nitratos, cada caso fue apareado con un control de la misma edad y sexo. Esto genera pares de individuos, y la comparación se hace "dentro de cada par".

En estos diseños, el OR no se calcula con la tabla 2×2 clásica, sino con los pares discordantes:

- Pares donde el caso estaba expuesto y el control no (b).
- Pares donde el caso no estaba expuesto y el control sí (c).

Tabla 11. Distribución de pares concordantes y discordantes en un estudio de casos y controles apareados.

	Control expuesto	Control no expuesto
Caso expuesto	12 (ambos expuestos, no cuentan)	18 (discordante)
Caso no expuesto	9 (discordante)	21 (ambos no expuestos, no cuentan)

Fuente: Elaboración propia

Interpretación: En 60 pares, hubo 18 pares "Caso expuesto / Control no expuesto" y 9 pares "Caso no expuesto / Control expuesto". El OR condicional es 18/9 = 2.0.

- El resultado indica que, dentro de los pares apareados por edad y sexo, los casos tenían el doble de probabilidad de estar expuestos a agua con nitratos que sus controles.
- Como el análisis fue condicional, la asociación se estimó independientemente de la edad y el

https://doi.org/10.62486/978-9915-9928-3-9.Ch12

sexo, porque esas variables ya estaban controladas por diseño.

Cuando el artículo dice que aplicó "regresión logística condicional", significa que analizó los datos respetando el apareamiento. El lector debe fijarse en que se reporten OR condicionales (y no simplemente OR crudos) cuando hubo apareamiento, de lo contrario los resultados podrían estar mal estimados.

6.10 OTRAS PRUEBAS QUE SUELEN APARECER

CHI CUADRADO O FISHER (TABLAS 2×2 NO PAREADAS)

Cuando se presenta una tabla simple de exposición/no exposición entre casos y controles, los autores suelen usar la prueba de Chi cuadrado para evaluar si las diferencias observadas podrían deberse al azar. Si el número de observaciones es pequeño (por ejemplo, cuando alguna celda tiene menos de 5 casos), se usa la prueba exacta de Fisher, que es más confiable en muestras pequeñas.

Ejemplo: En el estudio, si de 50 casos, 40 estaban expuestos y en los 50 controles solo 20 lo estaban, un Chi cuadrado o Fisher mostrará si esa diferencia es estadísticamente significativa. El OR nos dice la magnitud de la asociación (ej. OR=6.0), mientras que Chi cuadrado/Fisher nos dicen si esa diferencia podría explicarse solo por azar (<0.001).

MCNEMAR (PARA ESTUDIOS APAREADOS)

Cuando los casos y controles fueron apareados (pareados), por ejemplo por edad y sexo, el análisis se centra en los pares discordantes: aquellos en los que el caso estuvo expuesto y el control no, o viceversa.

La prueba de McNemar evalúa si la diferencia en estos pares discordantes es significativa. Va de la mano con el cálculo del OR condicional.

Ejemplo:Si en 60 pares, 18 pares muestran exposición solo en el caso y 9 pares exposición solo en el control, el OR condicional es 2.0. La prueba de McNemar dirá si esa diferencia (18 vs 9) es suficientemente grande para no atribuirla al azar.

PRUEBAS DE INTERACCIÓN (MODELOS MULTIVARIABLES)

En los modelos de regresión logística, los investigadores pueden añadir un término de interacción entre dos variables. Esto sirve para saber si el efecto de una exposición cambia dependiendo de otra variable.

Ejemplo**:** En nuestro estudio, el efecto de los nitratos podría ser diferente en fumadores y no fumadores. Si el modelo muestra un término de interacción significativo, eso significa que el tabaquismo modifica el efecto de los nitratos.

- En no fumadores: OR=2.5.
- En fumadores: OR=6.5.

Esto no es confusión, sino modificación de efecto real. El lector debe interpretar que el riesgo asociado a nitratos es mucho más alto en fumadores, y por lo tanto, los resultados deben reportarse por separado en cada subgrupo.

Tabla 12. Principales pruebas estadísticas utilizadas en estudios de casos y controles y su interpretación.

Prueba	¿Qué evalúa?	¿Cuándo se usa?	¿Cómo interpretarlo?
Chi cuadrado	Si la distribución de exposición difiere entre casos y controles.	Tablas 2×2 con muestras grandes.	Si el valor p < 0.05, la diferencia en exposición no se debe solo al azar. Complementa al OR/IC95%.
Exacta de Fisher	Igual que Chi cuadrado, pero más precisa con números pequeños.	Tablas 2×2 con frecuencias bajas (<5 en alguna celda).	Si p < 0.05, la asociación es significativa aunque la muestra sea pequeña.
Mantel–Haenszel	Calcula un OR "combinado" ajustado por una variable de estratificación.	Cuando se sospecha confusión y se dividen los datos en estratos (ej. fumadores vs no fumadores).	Si los OR por estrato son similares, se reporta el OR combinado. Si son muy distintos, hay modificación de efecto.

https://doi.org/10.62486/978-9915-9928-3-9.Ch12

Prueba	¿Qué evalúa?	¿Cuándo se usa?	¿Cómo interpretarlo?
McNemar	Si existe diferencia entre pares discordantes en estudios apareados.	Estudios de casos y controles apareados (1:1 o más).	Un $p < 0.05$ indica que la exposición fue más frecuente en casos que en controles dentro de los pares.
Pruebas de interacción (en regresión)	Si el efecto de la exposición cambia según otra variable.	Modelos multivariables (ej. regresión logística).	Un término de interacción significativo indica modificación de efecto real, no confusión. Ejemplo: nitratos con efecto más fuerte en fumadores que en no fumadores.

Fuente: Elaboración propia

Consejo para el lector critico: Cuando un artículo presenta estas pruebas, lo importante no es memorizar fórmulas, sino reconocer qué pregunta responde cada una. Así, la lectura crítica se vuelve más ordenada: Chi cuadrado/Fisher dicen si hay diferencia, Mantel–Haenszel muestra un OR ajustado por estratos, McNemar evalúa discordancia en pares apareados, y las pruebas de interacción responden si el efecto depende de otro factor.

6.11 ERRORES DE INTERPRETACIÓN FRECUENTES (Y CÓMO EVITARLOS)

Tabla 13. Errores frecuentes en la interpretación de estudios de casos y controles y cómo evitarlos

Error frecuente	Consecuencia	Cómo evitarlo
Traducir el OR como Riesgo Relativo (RR) sin que la enfermedad sea rara o sin justificación	Exagera la fuerza de la asociación	Recordar: OR ≈ RR solo cuando la enfermedad es poco frecuente
Confiar únicamente en el valor p	Puede llevar a conclusiones engañosas; ignora la magnitud y precisión del efecto	Revisar siempre el IC95% y su amplitud junto al p
Ignorar el diseño del estudio (ej. apareamiento)	Análisis inadecuado; resultados sesgados	Verificar si hubo apareamiento y buscar OR condicional
Ajustar por todas las variables o por las "significativas"	Riesgo de ajustar por colisionadores y distorsionar la asociación	Seleccionar variables de ajuste a priori con criterio causal
Olvidar la precisión de la muestra	OR extremos con IC muy amplios → resultados inciertos	Valorar el tamaño de muestra y la amplitud del IC95%

Fuente: Elaboración propia

7. SESGOS Y LIMITACIONES

Uno de los mayores desafíos al leer un estudio de casos y controles es identificar los sesgos que pueden distorsionar los resultados. Un sesgo es un error sistemático que afecta la validez interna del estudio y puede llevar a sobreestimar, infraestimar o incluso invertir la asociación real entre exposición y enfermedad. El lector crítico debe entrenarse para reconocerlos en la sección de métodos y discusión del artículo.

7.1 SESGO DE SELECCIÓN

Aparece cuando los casos y los controles no provienen de la misma población fuente. Esto hace que las diferencias observadas no reflejen una verdadera asociación, sino la forma en que fueron elegidos los participantes.

- Ejemplo aplicado: si los casos de cáncer gástrico se reclutan en La Paz pero los controles provienen de otra ciudad con agua tratada, la comparación es inválida. Igualmente, si los controles se seleccionan solo entre pacientes hospitalizados por enfermedades renales (también asociadas a nitratos), se produce el sesgo de Berkson, que tiende a diluir la asociación real.

7.2 SESGO DE INFORMACIÓN

Ocurre cuando la medición de la exposición o del desenlace no es precisa o difiere entre casos y controles.En estudios retrospectivos, el sesgo de recuerdo es frecuente: los casos tienden

https://doi.org/10.62486/978-9915-9928-3-9.Ch12

a recordar más su historia de exposición (ej. agua con mal sabor, consumo de alimentos salados) que los controles.También puede presentarse sesgo del observador, cuando el investigador sabe quién es caso y quién es control y mide la exposición con cierta expectativa.

- Ejemplo aplicado: si la exposición se evaluara solo preguntando al paciente si creía que el agua estaba contaminada, los casos tenderían a recordarlo con más detalle, exagerando la asociación. Si en cambio se midieran niveles de nitratos en laboratorio, el riesgo de este sesgo se reduciría.

7.3 SESGO DE SUPERVIVENCIA (NEYMAN)

Aparece cuando el estudio incluye casos prevalentes en lugar de incidentes. Como los casos prevalentes son aquellos que han sobrevivido cierto tiempo con la enfermedad, suelen representar formas más benignas.

- Ejemplo aplicado: si en el estudio se incluyen pacientes diagnosticados con cáncer gástrico hace 5 años y que aún viven, se excluiría a los pacientes con tumores agresivos que fallecieron rápidamente. El resultado sería una subestimación de la asociación entre nitratos y cáncer.

7.4 SESGO POR SOBREEMPAREJAMIENTO (OVERMATCHING)

El pareamiento puede ser útil para equilibrar grupos, pero si se aplica en exceso puede eliminar diferencias reales y hacer que la asociación desaparezca.

- Ejemplo aplicado: si los investigadores parearan casos y controles no solo por edad y sexo, sino también por características relacionadas con la exposición (ej. zona exacta de residencia y tipo de agua consumida), ambos grupos tendrían prácticamente la misma probabilidad de exposición, borrando la señal real del efecto de los nitratos.

7.5 CONFUSIÓN RESIDUAL

Incluso con buen diseño, algunos confusores pueden no medirse o no ajustarse de manera adecuada. Esto se llama confusión residual.

- Ejemplo aplicado: si el estudio no midiera adecuadamente la infección por *Helicobacter pylori*, un factor de riesgo fuerte para cáncer gástrico, la asociación entre nitratos y cáncer podría estar parcialmente distorsionada.

7.6 LIMITACIONES INHERENTES AL DISEÑO

Más allá de los sesgos específicos, el lector debe recordar las limitaciones propias de los estudios de casos y controles:

- No permiten calcular incidencia ni riesgo absoluto. Solo se puede estimar un OR.
- La temporalidad puede ser difícil de establecer. Aun cuando el OR sea alto, no siempre queda claro si la exposición precedió al inicio de la enfermedad.
- Susceptibilidad a sesgos de recuerdo y selección. Esto obliga a ser especialmente crítico al evaluar cómo se midió la exposición y cómo se eligieron los controles.
- Generalización limitada. Sus resultados suelen aplicarse con cautela a otras poblaciones, salvo que los contextos sean muy similares.

8. CHECKLIST

Los checklistson andamios que ayudan al lector a verificar, de forma ordenada, si un artículo reporta lo esencial y si lo hace con suficiente calidad. A nivel internacional, las guías más empleadas son:

- **STROBE (EQUATOR)**: pauta de reporte para estudios observacionales (incluye una lista específica para casos y controles, además de la combinada) y un documento de Explicación y Elaboración que aclara el porqué de cada ítem. Es el estándar de referencia para autores, revisores y lectores.
- **JBI CriticalAppraisalChecklist (2020)**: lista de evaluación crítica enfocada en riesgo de sesgo

https://doi.org/10.62486/978-9915-9928-3-9.Ch12

y validez del estudio, muy útil para lectores y para revisiones sistemáticas.

- **CASP Case–Control Checklist(edición vigente, 2024)**: 11 preguntas prácticas para valorar validez, resultados y aplicabilidad. Gran herramienta docente y clínica.
- **RECORD / RECORD-PE (extensión de STROBE)**: si el estudio usa datos rutinarios (registros, bases administrativas, historias clínicas electrónicas) o es farmacoepidemiológico, añade ítems de transparencia, enlace de bases, calidad de datos y trazabilidad del código.

A partir de estas guías, se propone el siguiente checklist práctico para lectura crítica:

Tabla 13. Checklist para la lectura crítica de estudios de casos y controles

Sección	Pregunta clave	Fuente principal
Identificación y objetivo	¿El título/resumen dejan claro que es un estudio de casos y controles y cuál es la pregunta principal?	STROBE
	¿El artículo justifica el diseño casos–controles (desenlace raro, latencia, factibilidad)?	STROBE–E&E
Métodos (validez interna)	¿Población fuente y periodo claramente definidos? ¿Casos y controles provienen de la misma población en riesgo?	STROBE, JBI
	¿Definición de casos estandarizada (idealmente incidentes)? ¿Se evitó sesgo de supervivencia (Neyman)?	STROBE–E&E, JBI
	¿Selección de controles adecuada? ¿Se evitó sesgo de Berkson?	STROBE–E&E, JBI, CASP
	¿Tipo de muestreo declarado y congruente con la medida reportada (OR vs IRR/RR)?	STROBE–E&E
	¿Medición de la exposición válida y objetiva/cegada? Si fue autorreporte, ¿se discutió sesgo de recuerdo?	STROBE, JBI, CASP
	Si hubo pareamiento: ¿se explican variables, criterios y análisis condicional? ¿Se evitó el overmatching?	STROBE–E&E, JBI
	¿Tamaño muestral y supuestos justificados (prevalencia, OR detectable, poder)?	STROBE
	¿Plan analítico preespecificado (OR crudo, ajuste por confusión, modificación de efecto, análisis de sensibilidad)?	STROBE–E&E, CASP
Resultados	¿Tablas 2×2 y OR con IC95% y p-valores presentados claramente (crudos y ajustados)?	STROBE
	Si hubo pareamiento: ¿se reportó OR condicional/logística condicional? Si estratificación: ¿OR de Mantel–Haenszel y homogeneidad?	STROBE–E&E
	Si se usaron datos rutinarios (HCE, registros), ¿se informó procedencia, códigos, enlaces y calidad?	RECORD/RECORD-PE
Discusión y aplicabilidad	¿Se identificaron sesgos plausibles (selección, recuerdo, supervivencia, overmatching) y su dirección probable?	CASP, JBI
	¿Se discutió validez externa y aplicabilidad clínica/salud pública?	STROBE
	¿Conclusiones prudentes? Evitar traducir OR a RR sin enfermedad rara y evitar inferencias causales sin soporte.	STROBE–E&E, CASP
Uso rápido (60 segundos)	1. Revisar título/resumen: diseño, población, exposición/desenlace.	Adaptado de STROBE, CASP, JBI
	2. Métodos: casos incidentes, controles comparables, exposición medida objetivamente, tipo de muestreo.	
	3. Resultados: OR con IC95% (¿cruza 1.0?), ajustes (¿cambian mucho?), análisis condicional si hubo pareamiento.	
	4. Discusión: sesgos reconocidos y aplicabilidad; si datos rutinarios, buscar ítems RECORD.	

Fuente: Elaboración propia

9. UTILIDAD

El diseño de casos y controles, pese a sus limitaciones, conserva un lugar central en la investigación clínica y epidemiológica. Su mayor fortaleza radica en la eficiencia: permite explorar asociaciones entre exposiciones y enfermedades sin necesidad de seguir a grandes cohortes durante años. Esto lo convierte en una herramienta particularmente útil en tres escenarios:

Enfermedades raras o de baja incidencia: Cuando el desenlace es infrecuente —como ciertos cánceres, enfermedades autoinmunes poco comunes o patologías raras, los estudios de cohorte requerirían decenas de miles de participantes para acumular suficientes casos. El diseño

https://doi.org/10.62486/978-9915-9928-3-9.Ch12

de casos y controles, en cambio, comienza reclutando directamente los casos ya ocurridos, lo que lo vuelve mucho más factible.

Enfermedades de larga latencia: En condiciones donde existe un período prolongado entre la exposición y el desarrollo de la enfermedad como la exposición a carcinógenos y la aparición de cáncer, un estudio prospectivo implicaría décadas de seguimiento. El enfoque de casos y controles acorta este tiempo, permitiendo evaluar retrospectivamente las exposiciones relevantes.

Exploración inicial de asociaciones: Frecuentemente, los estudios de casos y controles sirven como punto de partida para identificar hipótesis etiológicas que luego se confirmarán con diseños más robustos (cohortes o ensayos clínicos). Al detectar señales tempranas, pueden orientar la agenda de investigación y las políticas de salud.

En la práctica clínica: Para el médico lector, la utilidad está en la lectura crítica: al encontrarse con un estudio de casos y controles en una revista, debe reconocer que sus resultados no prueban causalidad, pero sí pueden aportar evidencia sugestiva y coherente con la biología de la enfermedad y otros estudios. En temas como el papel del *Helicobacter pylori* en cáncer gástrico, el consumo de tabaco en EPOC o la asociación entre fármacos y efectos adversos poco frecuentes, este diseño ha sido crucial para establecer vínculos iniciales que luego transformaron la práctica clínica.

En salud pública: Los hallazgos de los estudios de casos y controles pueden tener un impacto inmediato en políticas de prevención. Identificar factores de riesgo modificables (ej. consumo de agua con nitratos, exposición a humo de leña, hábitos dietéticos) permite diseñar intervenciones comunitarias y campañas educativas incluso antes de contar con grandes estudios prospectivos.

10. AUTOEVALUACIÓN

PREGUNTAS DE SELECCIÓN MÚLTIPLE

1. En un estudio de casos y controles, ¿cuál es la medida de asociación característica y más adecuada?

a) Riesgo relativo (RR)
b) Incidencia acumulada
c) Odds ratio (OR)
d) Diferencia de riesgos

2. Un artículo reporta OR = 2.5 (IC95% 0.9–6.0) para exposición a pesticidas y leucemia infantil. ¿Cuál es la interpretación correcta?

a) Existe asociación significativa porque el OR es > 1.

https://doi.org/10.62486/978-9915-9928-3-9.Ch12

b) No hay evidencia suficiente de asociación porque el IC95% incluye 1.
c) La exposición duplica el riesgo de leucemia en todos los casos.
d) El valor p debe ser <0.05 porque el OR es elevado.
3. El sesgo de Berkson ocurre típicamente cuando:
a) Los casos son seleccionados de otra población distinta a los controles.
b) Los controles hospitalarios tienen exposiciones relacionadas con su motivo de ingreso.
c) Se incluyen casos prevalentes en lugar de incidentes.
d) El investigador conoce la exposición antes de clasificar los grupos.
4. Si en un estudio apareado 1:1 se encuentran 18 pares con caso expuesto/control no expuesto y 9 pares con caso no expuesto/control expuesto, el OR condicional es:
a) 0.5
b) 1.0
c) 2.0
d) 3.0
5. ¿Cuál de las siguientes afirmaciones refleja mejor el valor de un OR ajustado frente al crudo?
a) El OR ajustado es siempre más alto que el crudo.
b) El OR ajustado refleja el efecto de la exposición controlando por posibles confusores.
c) El OR crudo es suficiente y no se necesita ajustar.
d) El ajuste elimina toda posibilidad de sesgo.

PREGUNTAS ABIERTAS

6. Explique con sus palabras la diferencia entre un OR crudo y un OR ajustado. ¿Por qué es importante mirar ambos al leer un artículo?
7. Describa un ejemplo de cómo el sesgo de recuerdo puede distorsionar los resultados de un estudio de casos y controles.
8. ¿Por qué los casos incidentes son preferibles a los casos prevalentes en este tipo de diseño?
9. Imagine que un artículo muestra un OR de 0.4 para consumo de frutas y cáncer gástrico, con IC95% 0.2–0.9. ¿Cómo lo interpretaría críticamente?
10. ¿En qué situaciones el OR puede aproximar al riesgo relativo (RR), y qué debe tener en cuenta el lector al hacer esa interpretación?

RESPUESTAS:

1: c) Odds ratio (OR). El Odds Ratio es la medida característica de los estudios de casos y controles porque no es posible calcular incidencia ni riesgo directamente, ya que los grupos se definen por el desenlace y no por la exposición.
2: b) No hay evidencia suficiente de asociación porque el IC95% incluye 1. Aunque el OR es mayor que 1, el intervalo de confianza cruza el valor nulo (1.0), lo que indica que la asociación no es estadísticamente significativa al 95%.
3: b) Los controles hospitalarios tienen exposiciones relacionadas con su motivo de ingreso. El sesgo de Berkson ocurre cuando los controles hospitalarios no representan adecuadamente la población fuente, generando una distorsión en la estimación del OR.
4: c) 2.0. En estudios apareados, el OR condicional se calcula como la razón entre pares discordantes: 18/9 = 2.0. Solo los pares discordantes aportan información a la estimación.
5: b) El OR ajustado refleja el efecto de la exposición controlando por posibles confusores. El ajuste permite estimar la asociación independiente de otras variables que podrían distorsionar el resultado.
6: El OR crudo muestra la asociación entre exposición y enfermedad sin controlar otras variables. El OR ajustado estima esa misma asociación pero controlando posibles factores de confusión (como edad, sexo o tabaquismo). Es importante mirar ambos porque si el OR cambia mucho tras el ajuste, probablemente existía confusión. Si casi no cambia, la asociación es más estable. La comparación permite evaluar la validez interna del estudio.
7: El sesgo de recuerdo ocurre cuando los casos recuerdan mejor o con mayor detalle exposiciones pasadas que los controles. Por ejemplo, en un estudio sobre pesticidas y leucemia infantil, los padres de niños enfermos pueden esforzarse más por recordar exposiciones

https://doi.org/10.62486/978-9915-9928-3-9.Ch12

ambientales, mientras que los padres de niños sanos pueden no hacerlo con la misma intensidad. Esto puede sobreestimar la exposición en los casos y exagerar el OR.

8: Los casos incidentes son preferibles porque representan diagnósticos recientes y reducen el sesgo de supervivencia (sesgo de Neyman). Los casos prevalentes incluyen solo pacientes que han sobrevivido, lo que puede excluir formas graves o letales de la enfermedad y distorsionar la asociación. Además, los casos incidentes permiten evaluar mejor la temporalidad entre exposición y enfermedad.

9: Un OR de 0.4 indica que la exposición (consumo de frutas) se asocia con una reducción aproximada del 60% en la probabilidad de cáncer gástrico. Como el IC95% (0.2–0.9) no incluye 1.0, la asociación es estadísticamente significativa. Sin embargo, debe analizarse si el OR es ajustado y qué variables se controlaron, ya que podría existir confusión por factores como dieta global, nivel socioeconómico o tabaquismo.

10: El OR puede aproximarse al riesgo relativo cuando la enfermedad es poco frecuente en la población (generalmente <10%). En ese contexto, la diferencia matemática entre odds y riesgo es pequeña. Sin embargo, si la enfermedad es común, el OR puede sobreestimar el efecto real. El lector debe verificar la frecuencia del desenlace antes de interpretar el OR como si fuera un RR.

https://doi.org/10.62486/978-9915-9928-3-9.Ch12

REFERENCIAS

Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. Epidemiology. 2007;18(6):800–4. Available at: https://doi.org/10.1097/EDE.0b013e3181577654.
Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and elaboration. PLoS Med. 2007;4(10):e297. Available at: https://doi.org/10.1371/journal.pmed.0040297.
Greenland S, Thomas DC. On the need for the rare disease assumption in case-control studies. Am J Epidemiol. 1982;116(3):547–53. Available at: https://doi.org/10.1093/oxfordjournals.aje.a113439.
Labrecque JA, Knol MJ. Do case-control studies always estimate odds ratios? Int J Epidemiol. 2020;49(5):1763–71. Available at: https://doi.org/10.1093/ije/dyaa052.
VanderWeele TJ, Ding P. Sensitivity analysis in observational research: Introducing the E-value. Ann Intern Med. 2017;167(4):268–74. Available at: https://doi.org/10.7326/M16-2607.
Pearce N. Analysis of matched case-control studies. BMJ. 2016;352:i969. Available at: https://doi.org/10.1136/bmj.i969.
Schneeweiss S, Rassen JA, Brown JS, Rothman KJ, Happe L, Arlett P, et al. Graphical depiction of longitudinal study designs in health care databases. Ann Intern Med. 2019;170(6):398–406. Available from: https://doi.org/10.7326/M18-3079.
Lash TL, VanderWeele TJ, Haneuse S, Rothman KJ. Modern epidemiology. 4th ed. Philadelphia: Wolters Kluwer; 2021.
Delgado-Rodríguez M, Llorca J. Bias. J Epidemiol Community Health. 2004;58(8):635–41. Available at: https://doi.org/10.1136/jech.2003.008466.
Alexander LK. Case–control studies: Basic concepts and rare disease assumption. Chapel Hill: North Carolina Institute for Public Health, University of North Carolina; 2015. Available at: https://sph.unc.edu/wp-content/uploads/sites/112/2015/07/nciph_ERIC5.pdf.

https://doi.org/10.62486/978-9915-9928-3-9.Ch12

Chapter 13 / Capítulo 13

Critical reading of scientific articles on health: A practical guide (Spanish Edition)
ISBN: 978-9915-9928-3-9
DOI: 10.62486/978-9915-9928-3-9.Ch13
Pages: 178-208

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Cohort Studies

Estudios de Cohorte

Jorge Marquez Molina
Isis Scarleth Funes Galindo
Stanley Aguirre
Giovanni Callizaya-Macedo

1 Hospital del Norte. Cochabamba, Bolivia.
2 Clínica municipal de emergencias. Talanga. Honduras.
3 Investigador independiente.
4 Hospital del Niño "Dr. Ovidio Aliaga Uría". Departamento de Cirugía Pediátrica. La Paz, Bolivia.

ABSTRACT

Introduction: Cohort studies are among the most robust analytical observational designs for assessing the association between an exposure and the occurrence of an outcome. Their main strength lies in establishing the temporal sequence between these two events, allowing for the estimation of incidences and the quantification of risks using measures such as relative risk (RR), incidence rate, and hazard ratio (HR). Development: This chapter describes the methodological foundations of prospective, retrospective, and ambispective cohorts, emphasizing the importance of appropriate population selection, precise definition of the exposure and outcomes, longitudinal follow-up, and control for confounding factors. It addresses the main measures of frequency and association, including cumulative incidence, incidence rate, attributable risk, RR, and HR, complemented by survival analysis tools such as Kaplan-Meier curves and Cox regression. Criteria for evaluating the internal and external validity and clinical relevance of the findings are also highlighted. Conclusions: Cohort studies provide fundamental evidence for understanding the natural history of diseases and estimating the impact of risk and protective factors. Their proper interpretation requires assessing the time frame, the quality of follow-up, the control of confounding factors, and the true magnitude of the observed effects. Beyond statistical results, these studies allow for the translation of evidence into clinical decisions and public health strategies aimed at prevention and improving healthcare.

Keywords: Cohort Studies; Relative Risk; Hazard Ratio; Survival Analysis; Confounding Factors

RESUMEN

Introducción: Los estudios de cohorte constituyen uno de los diseños observacionales analíticos más sólidos para evaluar la asociación entre una exposición y la ocurrencia de un desenlace. Su principal fortaleza radica en establecer la secuencia temporal entre ambos eventos, permitiendo estimar incidencias y cuantificar riesgos mediante medidas como el riesgo relativo (RR), la tasa de incidencia y el hazard ratio (HR). Desarrollo: El capítulo describe los fundamentos metodológicos de las cohortes prospectivas, retrospectivas y ambispectivas, enfatizando la importancia de una adecuada selección de la población, la definición precisa de la exposición y los desenlaces, el seguimiento longitudinal y el control de factores de confusión. De esta manera, aborda las principales medidas de frecuencia y asociación, incluyendo incidencia acumulada, tasa de incidencia, riesgo atribuible, RR y HR, complementadas con herramientas de análisis de supervivencia como las curvas de Kaplan–Meier y la regresión de Cox. También se destacan los criterios para evaluar la validez interna, externa y la relevancia clínica de los hallazgos. Conclusiones: Los estudios de cohorte proporcionan evidencia fundamental para comprender la historia natural de las enfermedades y estimar el impacto de factores de riesgo y protección. Su adecuada interpretación exige valorar la temporalidad, la calidad del seguimiento, el control de la confusión y la magnitud real de los efectos observados. Más allá de los resultados estadísticos, estos estudios permiten traducir la evidencia en decisiones clínicas y estrategias de salud pública orientadas a la prevención y al mejoramiento de la atención sanitaria.

Palabras clave: Estudios de Cohorte; Riesgo Relativo; Hazard Ratio; Análisis de Supervivencia; Factores de Confusión

1. CASO CLÍNICO

En una comunidad rural del altiplano boliviano, donde la mayoría de los hogares utilizan leña como principal fuente de energía para cocinar, un médico internista nota un número creciente de pacientes con tos crónica, disnea progresiva y diagnóstico de enfermedad pulmonar obstructiva crónica (EPOC). Surge la pregunta clínica: ¿la exposición prolongada al humo de leña aumenta el riesgo de desarrollar EPOC en comparación con la no exposición?

Para responder, se diseña un estudio de cohorte prospectivo:

- Se reclutan 1,200 adultos sin EPOC al inicio del estudio (600 expuestos al humo de leña diariamente por ≥10 años y 600 no expuestos que utilizan gas natural o electricidad).
- Se los sigue durante 12 años, evaluando anualmente la función pulmonar mediante espirometría y registrando aparición de EPOC.
- Durante el seguimiento, algunos participantes cambian de fuente de energía, lo que obliga a registrar la exposición de forma dinámica.
- El desenlace principal es la incidencia acumulada de EPOC, y se planea calcular el riesgo relativo (RR), la tasa de incidencia (TI) y el hazard ratio (HR) mediante modelos de Cox.

2. OBJETIVOS

Los estudios de cohorte son diseños investigativos longitudinales con el propósito de responder preguntas clínicas u epidemiológicas estableciendo la secuencia temporal en tiempo entre la exposición y el desenlace sobre los individuos expuestos y no expuestos.

2.1 OBJETIVOS ESPECIFICOS

- Evaluar la asociación entre exposición y desenlace.
- Estimar la incidencia de la enfermedad en cada grupo.
- Calcular medidas de asociación: Riesgo Relativo, Tasa de Incidencia, Hazard Ratio.

2.2 FINALIDAD DE LOS ESTUDIOS DE COHORTE

- Determinar la evaluación de la causalidad.
- Cuantificar el impacto de la exposición: Permite calcular el riesgo atribuible (RA) y la fracción atribuible poblacional (FAP).
- Generar evidencia informativa para la prevención.
- Analizar desenlaces secundarios.

3. INTRODUCCIÓN

Los estudios de cohorte son estudios observacionales, longitudinales, analíticos, generalmente prospectivos que constituyen uno de los pilares de la epidemiología clínica porque permiten observar cómo evoluciona una población en relación con una exposición específica a lo largo del tiempo. A diferencia de los estudios transversales, que solo muestran una “fotografía” de la realidad, las cohortes proporcionan una película en movimiento, siguiendo a los individuos desde que están libres del desenlace hasta que éste ocurre.En términos metodológicos, una cohorte se define como un grupo de personas que comparte una característica común en un momento inicial (por ejemplo, estar expuestos al humo de leña, como en nuestro caso clínico) y que se sigue en el tiempo para evaluar la aparición de un desenlace (EPOC). El rasgo distintivo es que la exposición se mide antes de que ocurra el desenlace, lo que garantiza la secuencia temporal, requisito fundamental para aproximarse a inferencias causales.

Importancia clínica y epidemiológica:

- Los estudios de cohorte permiten estimar la incidencia de enfermedades, una medida clave para entender cuán frecuente es que aparezca un desenlace en personas inicialmente sanas.
- Al comparar expuestos y no expuestos, se obtiene el RR y otras métricas (tasa de incidencia, hazard ratio), que son las medidas de asociación más usadas en investigación clínica.
- Una sola cohorte puede analizar múltiples desenlaces para una misma exposición (por

https://doi.org/10.62486/978-9915-9928-3-9.Ch13

ejemplo, humo de leña → EPOC, hipertensión pulmonar, cáncer pulmonar, mortalidad).

- De igual forma, permite estudiar tanto factores de riesgo como de protección (ejemplo: uso de cocina a gas como medida protectora frente a EPOC).

En países de ingresos medios y bajos, especialmente en zonas rurales, los estudios de cohorte son herramientas valiosas para documentar la carga de enfermedad relacionada con exposiciones ambientales y ocupacionales. El caso clínico que plantea la asociación entre humo de leña y EPOC refleja un problema real de salud pública en Bolivia y América Latina, donde más del 30% de los hogares rurales aún utilizan biomasa como principal fuente de energía.

Para el personal de salud, la lectura crítica de cohortes es esencial porque:

- Aclara el riesgo real de que un paciente desarrolle una enfermedad tras estar expuesto a un factor.
- Informa decisiones preventivas: identificar factores de riesgo modificables permite orientar recomendaciones prácticas (dejar de fumar, cambiar hábitos dietéticos, reducir exposición ambiental).
- Fortalece la medicina basada en evidencia (MBE): una cohorte bien diseñada ofrece evidencia de alta calidad entre los estudios observacionales analíticos, especialmente cuando se complementa con análisis estadísticos robustos que controlan confusores.

4. METODOLOGÍA

La sección de metodología es el corazón de un estudio de cohorte y constituye la parte más importante a la hora de evaluarlo críticamente. Allí se decide si los resultados serán creíbles, válidos y aplicables a la práctica clínica. Al leer un artículo, el personal de salud debe preguntarse si la población fue bien seleccionada, si la exposición y los desenlaces fueron medidos de manera confiable, si el seguimiento fue suficiente y si el análisis estadístico fue el adecuado para la pregunta planteada. A continuación, se presentan los principales componentes metodológicos que deben evaluarse al leer un artículo de cohorte.

4.1 TIPOS DE ESTUDIOS DE COHORTE

Los estudios de cohorte pueden clasificarse en prospectivos, retrospectivos o ambispectivos según el momento en que se recolectan los datos y la forma en que se sigue a los participantes. Identificar correctamente el tipo de cohorte en un artículo es esencial, porque de ello dependerán la validez interna, el tiempo necesario, los costos y el riesgo de sesgos.

COHORTE PROSPECTIVA

Una cohorte prospectiva, también llamada concurrente, comienza en el presente: los investigadores reclutan a los participantes, miden su exposición antes de que ocurra el desenlace, y los siguen hacia adelante en el tiempo. En este diseño, la secuencia temporal es muy clara, ya que la exposición se registra antes de la aparición del evento.

Su fortaleza principal radica en que permite estandarizar la medición de la exposición y los desenlaces, con criterios diagnósticos definidos desde el inicio. Miden características en puntos de tiempo predeterminados para entender la relación entre la cohorte y resultado del estudio. Además, facilita el control de calidad de los datos y la verificación de la temporalidad. Por estas razones, suele considerarse el “estándar de oro” dentro de los estudios observacionales.

Sin embargo, tiene limitaciones importantes: es costoso, requiere un gran tamaño de muestra, bastante tiempo de seguimiento y suele enfrentar pérdidas de participantes que, si son diferenciales entre expuestos y no expuestos, introducen sesgos (de desgaste).

En nuestro caso clínico: reclutar en la actualidad a 600 adultos expuestos al humo de leña y 600 no expuestos en comunidades rurales del altiplano, asegurándose de que ninguno tenga EPOC al inicio. Luego se realiza un seguimiento anual por 12 años, con espirometrías estandarizadas, hasta observar cuántos desarrollan EPOC en cada grupo.

COHORTE RETROSPECTIVA

En una cohorte retrospectiva, también llamada histórica, los análisis de datos son en

https://doi.org/10.62486/978-9915-9928-3-9.Ch13

el presente, pero las mediciones en la línea de base de los participantes y sus seguimientos ocurrieron en el pasado. Se utilizan expedientes clínicos, bases de datos hospitalarias, censos o registros administrativos para clasificar a los participantes según su exposición y determinar si desarrollaron o no el desenlace.

La fortaleza de este diseño es que es mucho más rápido y económico que una cohorte prospectiva. Permite obtener resultados en menor tiempo al revisar variables predictoras (datos pasados) para examinar eventos o resultados, útil cuando se dispone de registros de buena calidad.

No obstante, su limitación principal es la dependencia absoluta de los registros existentes. Si la exposición o el desenlace no fueron medidos de forma precisa o estandarizada, el riesgo de sesgo de información es elevado. Además, la temporalidad puede estar menos clara, ya que los datos se reconstruyen de manera retrospectiva.

En nuestro caso clínico: revisar historias clínicas y registros de salud entre 2010 y 2020 para identificar a personas que estuvieron expuestas al humo de leña y compararlas con aquellas que no lo estuvieron. A partir de esos registros, se determina cuántos desarrollaron EPOC en el período. La calidad del estudio dependerá de si los registros incluyeron espirometrías y exposición detallada al tipo de combustible utilizado.

COHORTE AMBISPECTIVA

La cohorte ambispectiva, también llamada mixta, combina elementos de los dos diseños anteriores. Se inicia utilizando datos retrospectivos ya disponibles para clasificar la exposición y recoger parte de la información histórica, y a partir de ese punto se continúa con un seguimiento prospectivo hacia adelante.

La fortaleza de este diseño es que aprovecha la rapidez de los datos históricos, pero al mismo tiempo permite continuar la observación con métodos estandarizados en el presente. De esta forma, combina eficiencia con mayor control sobre la calidad de la información.

La limitación, sin embargo, es que hereda los problemas de ambas modalidades: puede tener registros incompletos del pasado y enfrenta pérdidas de seguimiento en la etapa prospectiva.

En nuestro caso clínico: los investigadores podrían revisar los expedientes de un hospital rural de los últimos 7 años para identificar la exposición al humo de leña y diagnosticar casos de EPOC durante ese período. A partir de hoy, se decide continuar el seguimiento prospectivo durante cinco años más, con visitas anuales y mediciones estandarizadas de función pulmonar, para registrar nuevos casos de EPOC.

En resumen, al leer un artículo de cohorte es indispensable identificar el tipo de diseño. Una cohorte prospectiva suele ser más válida pero costosa; una retrospectiva es más rápida y barata, pero depende de la calidad de los registros; y una ambispectiva busca un equilibrio entre ambos enfoques. Reconocer estas diferencias permite al lector critico juzgar de inmediato las fortalezas y debilidades del estudio que tiene frente a sí, evitando interpretaciones simplistas o erróneas.

Tabla 1. Comparación entre los tipos de estudios de cohorte

Característica	Cohorte prospectiva	Cohorte retrospectiva	Cohorte ambispectiva
Momento de inicio del estudio	El estudio comienza en el presente y el seguimiento ocurre hacia el futuro.	El estudio se realiza en el presente, pero la exposición y el seguimiento ocurrieron en el pasado.	Combina datos del pasado y seguimiento prospectivo desde el presente.
Medición de la exposición	Se mide antes de que ocurra el desenlace.	Se reconstruye a partir de registros existentes.	Parte retrospectiva reconstruida + medición prospectiva posterior.
Temporalidad exposición–desenlace	Muy clara y bien establecida.	Puede ser menos clara si los registros son incompletos.	Clara en la fase prospectiva; depende de registros en la fase retrospectiva.
Control sobre calidad de datos	Alto (mediciones estandarizadas desde el inicio).	Variable (depende de la calidad de los registros históricos).	Moderado (hereda limitaciones del pasado, mejora en fase prospectiva).
Tiempo para obtener resultados	Largo (requiere años de seguimiento).	Corto (los eventos ya ocurrieron).	Intermedio.
Costo	Elevado.	Bajo o moderado.	Intermedio.
Riesgo de sesgo de información	Bajo si hay buen control metodológico.	Mayor (dependencia de registros previos).	Variable (alto en fase retrospectiva, menor en fase prospectiva).
Riesgo de pérdidas de seguimiento	Alto si el seguimiento es prolongado.	Bajo (los eventos ya ocurrieron).	Presente en fase prospectiva.

https://doi.org/10.62486/978-9915-9928-3-9.Ch13

Característica	Cohorte prospectiva	Cohorte retrospectiva	Cohorte ambispectiva
Ejemplo aplicado al caso clínico (EPOC y humo de leña)	Reclutar hoy 1,200 adultos sin EPOC y seguirlos 12 años con espirometría anual.	Revisar registros 2010–2020 para identificar exposición y diagnóstico de EPOC.	Revisar registros 2015–2022 y continuar seguimiento 5 años más con mediciones estandarizadas.
Fortaleza principal	Mayor validez interna y mejor control de medición.	Rapidez y menor costo.	Equilibrio entre eficiencia y control metodológico.
Limitación principal	Costosa y susceptible a pérdidas diferenciales.	Dependencia absoluta de registros previos.	Combina limitaciones de ambos diseños.
Fuente: Elaboración propia			

4.2 SELECCIÓN DE LA POBLACIÓN

La validez de un estudio de cohorte depende en gran medida de cómo se selecciona la población. Para el lector crítico es fundamental preguntarse si los participantes incluidos realmente representan a la población de interés y si la muestra inicial estaba libre del desenlace en estudio.

En primer lugar, la población de estudio debe estar libre de la enfermedad al inicio del seguimiento. De lo contrario, lo que se estaría midiendo no es la incidencia de la enfermedad, sino su prevalencia, lo cual distorsionaría los resultados. En el caso, si se investiga la relación entre humo de leña y riesgo de EPOC, ningún participante debería tener un diagnóstico previo de EPOC al momento de ser reclutado.

Además, los autores deben describir claramente los criterios de inclusión y exclusión. Un buen artículo especifica quiénes fueron elegibles y quiénes quedaron fuera del estudio. En nuestro caso, podrían incluirse adultos mayores de 25 años residentes en comunidades rurales que utilicen leña o gas como fuente de energía. En cambio, podrían excluirse personas con enfermedades pulmonares intersticiales, cáncer de pulmón o con antecedentes de exposición ocupacional intensa a polvos industriales, ya que estas condiciones confundirían la relación entre humo de leña y EPOC.

Un aspecto clave es la representatividad de la muestra. El lector debe preguntarse: ¿los participantes incluidos reflejan a la población general a la que se quieren aplicar los resultados? Una cohorte que reclute únicamente pacientes de un hospital puede no ser representativa de toda la comunidad rural, porque quienes llegan a la atención médica suelen diferir en características socioeconómicas o en gravedad de síntomas. Por eso, en el caso clínico, la validez sería mayor si los investigadores reclutan participantes directamente en las comunidades, visitando hogares y aplicando cuestionarios estandarizados.

Otro criterio esencial es la comparabilidad entre grupos. En estudios de cohorte, los expuestos y los no expuestos deben ser similares en características importantes como edad, sexo, nivel socioeconómico, tabaquismo y comorbilidades. Si un grupo está compuesto mayoritariamente por personas mayores y el otro por jóvenes, la diferencia de riesgo podría deberse a la edad y no a la exposición estudiada. Aquí entra en juego el concepto de confusión, que se analizará en detalle más adelante, pero el lector debe reconocer desde la metodología si existe un desequilibrio inicial entre los grupos.

Finalmente, es importante considerar la tasa de participación. Un artículo debería indicar qué porcentaje de las personas elegibles aceptó participar. Una tasa muy baja podría introducir sesgo de selección, ya que quienes aceptan participar pueden diferir en motivaciones, salud o conductas respecto de quienes rechazan hacerlo.

En síntesis, al evaluar la selección de la población en un estudio de cohorte, el lector debe asegurarse de que los participantes estuvieran libres de la enfermedad al inicio, que los criterios de inclusión y exclusión estén bien definidos, que la muestra sea representativa de la población objetivo, que los grupos expuestos y no expuestos sean comparables, y que la tasa de participación sea adecuada. Solo así podrá confiar en que los resultados reflejan la realidad del problema en estudio y no una distorsión generada por una muestra sesgada.

4.3 DEFINICIÓN Y MEDICIÓN DE LA EXPOSICIÓN

La forma en que se define y mide la exposición es determinante para la validez de los resultados, el lector critico debe prestar atención a cómo los investigadores conceptualizaron la exposición y con qué herramientas la evaluaron. Una exposición mal definida, evaluada de

https://doi.org/10.62486/978-9915-9928-3-9.Ch13

manera imprecisa o heterogénea entre los grupos, puede generar sesgos graves que conduzcan a conclusiones equivocadas.

La definición operacional de la exposición es el primer aspecto a revisar. Es necesario que este definido su concepto general con claridad sobre lo que se entiende por exposición, así puede ser transformada en una variable concreta, medible y reproducible. No basta con mencionar que alguien estuvo "expuesto al humo de leña"; Por ejemplo, en nuestro caso clínico debería ser definido como "utilizar leña como principal combustible para cocinar al menos cuatro horas al día durante diez años consecutivos".

En segundo lugar, es fundamental examinar los métodos de medición de la exposición. Estos pueden ser objetivos o subjetivos. Los métodos objetivos incluyen, por ejemplo, el uso de sensores ambientales para medir partículas contaminantes, biomarcadores de exposición (como niveles de monóxido de carbono en sangre o carboxihemoglobina), o registros oficiales sobre el tipo de combustible utilizado en los hogares. Los métodos subjetivos, en cambio, se basan en cuestionarios o autoinformes, donde los participantes declaran sus hábitos de uso de combustibles. Aunque más fáciles de aplicar, son susceptibles a errores de memoria o a respuestas socialmente deseables.

El lector crítico debe preguntarse si los investigadores utilizaron la misma técnica de medición para todos los participantes. La homogeneidad en la medición es crucial: si en un grupo se utilizaron cuestionarios y en otros biomarcadores, las diferencias encontradas podrían reflejar más el método de medición que la exposición en sí.

Otro punto a considerar es la clasificación de la exposición. Esta puede ser dicotómica (expuesto/no expuesto) o categórica, estableciendo niveles de intensidad. La estratificación no solo permite analizar si existe una relación dosis–respuesta, sino que fortalece la interpretación causal de los hallazgos. Por ejemplo, en el caso, se clasifica a los participantes como de exposición baja (menos de dos horas al día), moderada (dos a cuatro horas) o alta (más de cuatro horas diarias), así a mayor intensidad de exposición aumenta el riesgo de EPOC y la asociación gana consistencia.

Un aspecto avanzado que cada vez más estudios incluyen es la exposición dinámica. En cohortes largas, es habitual que los hábitos cambien con el tiempo. Los mejores estudios actualizan la exposición durante el seguimiento y la analizan como una variable dependiente del tiempo. Por ejemplo, una familia que inicialmente usaba leña puede adoptar gas o electricidad en el transcurso del estudio. Si el artículo clasifica a los participantes únicamente en función de la exposición inicial, se corre el riesgo de cometer errores de clasificación que subestimen o sobreestimen el efecto.

En síntesis, cuando se revisa un estudio de cohorte, el lector debe asegurarse de que la exposición haya sido definida de manera precisa, medida con métodos confiables, evaluada con el mismo procedimiento en todos los participantes y, cuando sea posible, actualizada a lo largo del seguimiento. Si la definición es vaga, la medición subjetiva o la clasificación inadecuada, las conclusiones del estudio pierden fuerza y su aplicabilidad clínica se ve comprometida.

4.4 DEFINICIÓN Y MEDICIÓN DE LOS DESENLACES

El desenlace es la variable de mayor relevancia, pues representa el evento clínico cuya relación con la exposición se busca establecer. La claridad en su definición y la precisión en su medición determinan la calidad de las conclusiones del estudio. Para el lector crítico, esta sección debe ser revisada con especial atención, ya que los errores en la definición del desenlace son una de las fuentes más frecuentes de sesgo.

Lo primero que debe verificarse es si el desenlace principal (resultado primario) está definido de manera explícita y objetiva. La definición debe basarse en criterios diagnósticos reconocidos internacionalmente. Si un estudio no aclara con qué criterio se diagnosticó el desenlace principal, o si utiliza definiciones vagas como "síntomas respiratorios crónicos", su validez científica queda seriamente limitada. En nuestro caso clínico, el desenlace de interés es el diagnóstico de enfermedad pulmonar obstructiva crónica (EPOC), como el establecido por la Iniciativa Global para la EPOC (GOLD): un cociente FEV1/FVC menor al 70% confirmado mediante espirometría post-broncodilatador.

Además del desenlace principal, muchos estudios de cohorte evalúan desenlaces secundarios (resultados secundarios), que enriquecen el análisis y permiten una visión más amplia de la carga de enfermedad. Estos desenlaces secundarios ofrecen información complementaria sobre la evolución clínica y la severidad de la enfermedad. En el caso, los investigadores

https://doi.org/10.62486/978-9915-9928-3-9.Ch13

registraron exacerbaciones graves de EPOC, hospitalizaciones, mortalidad por causa respiratoria o disminución progresiva del volumen espiratorio forzado (FEV1).

Otro aspecto esencial es la homogeneidad en la medición. Todos los participantes deben ser evaluados con los mismos métodos diagnósticos, en las mismas condiciones y con protocolos estandarizados. El lector debe comprobar que el artículo especifique cómo, cuándo y con qué instrumentos se midieron los desenlaces. En nuestro caso si algunos participantes son diagnosticados mediante espirometría y otros solo a partir de síntomas clínicos, el riesgo de sesgo de información aumenta considerablemente.

Los estudios también pueden incluir desenlaces compuestos, es decir, una combinación de varios eventos clínicos, como EPOC, hospitalización y mortalidad respiratoria. Aunque esto aumenta la potencia estadística, el lector crítico debe valorar si todos los componentes del desenlace compuesto son de igual relevancia clínica. Si se combinan eventos leves (tos crónica) con eventos graves (muerte por insuficiencia respiratoria), la interpretación puede resultar engañosa.

Finalmente, en cohortes largas, es fundamental que los desenlaces sean evaluados en intervalos regulares y definidos. El lector debe identificar si el artículo reporta estos intervalos y si hubo consistencia en la recolección de datos. En el caso, sería razonable realizar controles anuales con espirometría, complementados con visitas clínicas en caso de síntomas o exacerbaciones.

En conclusión, para que los resultados de una cohorte sean válidos, los desenlaces deben estar claramente definidos, medidos con métodos estandarizados, aplicados de forma homogénea en todos los participantes y evaluados en intervalos regulares durante el seguimiento. De no cumplirse estas condiciones, la confianza en los hallazgos se ve comprometida, sin importar cuán sofisticado sea el análisis estadístico posterior.

4.5 SEGUIMIENTO DE LA COHORTE

El seguimiento constituye la columna vertebral de un estudio de cohorte. Es en esta fase donde se observa si los participantes desarrollan o no el desenlace en cuestión, y de la calidad del seguimiento dependerá que las medidas de incidencia y asociación reflejen la realidad. Para el lector crítico, evaluar cómo se llevó a cabo este proceso es tan importante como revisar la definición de exposición o desenlace.

El primer aspecto a considerar es la duración del seguimiento. Ésta debe ser lo suficientemente larga como para que aparezca el desenlace de interés. Si un estudio presenta un tiempo de seguimiento demasiado corto, sus conclusiones sobre causalidad deben ser interpretadas con cautela. En el caso, un seguimiento breve de dos o tres años sería insuficiente para detectar la aparición de EPOC, una enfermedad de evolución crónica y lenta. En cambio, un período de 10 a 12 años, ofrece mayor probabilidad de registrar nuevos diagnósticos y de establecer diferencias claras entre expuestos y no expuestos.

En segundo lugar, es esencial examinar los intervalos de evaluación. Un estudio serio no se limita a observar a los participantes al inicio y al final, sino que programa controles periódicos para recolectar datos de forma continua. Estos pueden ser anuales, semestrales o adaptados al desenlace de interés. Cuando los intervalos son irregulares o no se explicitan en el artículo, el riesgo de datos incompletos o de diagnósticos tardíos aumenta. En el caso, los controles anuales con espirometría y cuestionarios respiratorios serían un estándar apropiado.

Un tercer aspecto crucial es el problema de las pérdidas de seguimiento. En cualquier cohorte es esperable que algunos participantes abandonen el estudio, se muden, fallezcan por causas no relacionadas o simplemente dejen de acudir a los controles. El lector debe fijarse en cuántas personas se perdieron y en qué grupo ocurrieron esas pérdidas. Una pérdida global inferior al 10% se considera aceptable; cuando supera el 20%, la validez de los resultados queda seriamente comprometida. Más grave aún es cuando las pérdidas son diferenciales: si, por ejemplo, abandonan más participantes del grupo expuesto porque desarrollaron síntomas respiratorios severos, los resultados podrían subestimar la verdadera incidencia del desenlace. Los autores deben reportar no solo el número de pérdidas, sino también las razones, y describir qué métodos utilizaron para minimizar su impacto (contacto telefónico, visitas domiciliarias, análisis de sensibilidad).

Otro elemento cada vez más importante es el manejo de la exposición dinámica. En estudios prolongados, es frecuente que los hábitos de los participantes cambien con el tiempo como se mencionó previamente. Los mejores artículos actualizan la información de exposición en cada visita de seguimiento y utilizan modelos estadísticos que permiten tratar la exposición como una variable dependiente del tiempo. En nuestro caso, algunas familias inicialmente expuestas al

https://doi.org/10.62486/978-9915-9928-3-9.Ch13

humo de leña pueden migrar al uso de gas natural o electricidad, mientras que otras, por razones económicas, pueden volver a usar leña tras un período sin hacerlo.

Finalmente, el lector debe considerar la calidad de los registros durante el seguimiento. Es común que algunos estudios mencionen el tiempo de observación, pero no detallen cómo se documentaron los eventos. Un buen artículo debe especificar si los desenlaces fueron confirmados mediante pruebas objetivas (como espirometrías), si hubo un comité independiente que validó los diagnósticos y si se utilizaron fuentes múltiples de información (historias clínicas, entrevistas, registros de defunción).

En conclusión, el seguimiento en una cohorte debe ser suficiente en duración, realizado en intervalos regulares, con pérdidas mínimas y bien documentadas, considerando los cambios en la exposición y asegurando la calidad en la recolección de datos. Para el lector crítico, un estudio con seguimiento incompleto o deficiente es como una película mal editada: puede tener un inicio prometedor, pero carece de un final confiable que permita interpretar la historia en su totalidad.

4.6 CONTROL DE LA CONFUSIÓN

Uno de los grandes retos de los estudios de cohorte es el manejo de la confusión. El lector crítico debe estar atento a cómo el artículo identifica, mide y controla los posibles confusores, ya que de ello depende la validez de la asociación observada.

Se entiende por confusión a la distorsión de la relación entre exposición y desenlace producida por una variable (externa). Las variables confusoras deben tener 3 condiciones, estar asociadas con la exposición investigada; estar asociadas con el resultado investigado; no deben estar en el camino causal entre exposición y resultado. Variables a considerar como confusores son: la edad, el sexo, el nivel socioeconómico o las condiciones ocupacionales. Todas ellas deben estar claramente registradas en la cohorte y reportadas en la metodología. En el caso clínico de humo de leña y riesgo de EPOC, el tabaquismo es un confusor evidente: los fumadores tienen más probabilidad de estar expuestos a humo en el hogar (por factores culturales o socioeconómicos) y, al mismo tiempo, el tabaco es un factor de riesgo reconocido para desarrollar EPOC. Si no se mide y ajusta adecuadamente, se podría concluir erróneamente que el humo de leña tiene un efecto mayor al real.

La confusión podría dar lugar a una distorsión de los efectos; puede conducir a sobreestimación o subestimación de un efecto, o incluso invertir la dirección de un efecto/resultado.

El lector debe analizar si el artículo aplicó estrategias de control de confusión. Estas pueden plantearse en dos niveles:

- Durante el diseño del estudio. Algunos autores emplean técnicas como la restricción (incluir solo no fumadores para eliminar la influencia del tabaco), el emparejamiento (hacer que los grupos expuestos y no expuestos sean comparables en edad, sexo u otras variables) o la selección de una muestra homogénea. Estas medidas mejoran la comparabilidad desde el inicio, aunque reducen la generalización de los resultados.
- Durante el análisis estadístico. Aquí se aplican métodos más sofisticados que buscan ajustar los resultados a los confusores medidos. Los más frecuentes son la regresión multivariable (incluyendo edad, sexo, tabaquismo, ocupación), la estratificación por categorías (calcular riesgos separados en fumadores y no fumadores) o el uso de puntuaciones de propensión (propensity scores) que equilibran los grupos mediante técnicas estadísticas. En cohortes con análisis de supervivencia, la regresión de Cox ajustada es una de las herramientas más utilizadas.

Un aspecto importante que el lector crítico debe valorar es cómo los autores seleccionaron las variables de ajuste. Si eligieron ajustar únicamente aquellas que resultaron “significativas” en un análisis preliminar, existe riesgo de sobreajuste o de dejar fuera variables realmente importantes. La mejor práctica consiste en justificar la selección de confusores con base en la literatura científica y en modelos conceptuales, como los diagramas causales dirigidos (DAGs).

Por último, es esencial recordar que ningún estudio observacional está completamente libre de confusión. Aun con los mejores métodos estadísticos, siempre puede quedar un grado de confusión residual debido a factores no medidos o mal clasificados. Por eso, un artículo sólido debe reconocer esta limitación en su discusión.

En resumen, al leer un estudio de cohorte, el lector critico debe preguntarse: ¿los autores identificaron los confusores más relevantes? ¿Los midieron adecuadamente? ¿Aplicaron técnicas de ajuste apropiadas? Y, sobre todo, ¿reconocieron la posibilidad de confusión residual? Si la respuesta es afirmativa, la confianza en los hallazgos aumenta significativamente.

https://doi.org/10.62486/978-9915-9928-3-9.Ch13

4.7 ANÁLISIS ESTADÍSTICO

El análisis estadístico tiene como propósito comparar la frecuencia de los desenlaces entre los grupos expuestos y no expuestos. Para ello se utilizan medidas de ocurrencia y asociación que permiten al lector cuantificar el riesgo y entender la magnitud de la relación observada.

El punto de partida suelen ser las medidas de incidencia, que informan cuántos casos nuevos de la enfermedad aparecen en un periodo determinado. A partir de estas incidencias, se calculan las medidas de asociación más utilizadas en cohortes: el riesgo relativo y la tasa de incidencia. En estudios donde interesa el tiempo hasta que ocurre el evento, se emplean análisis de supervivencia, como las curvas de Kaplan–Meier, y modelos más complejos como la regresión de Cox, que generan el hazard ratio.

Al evaluar un artículo, el lector debe fijarse en que estas medidas estén acompañadas de intervalos de confianza (IC95%). Un valor p aislado nunca es suficiente. Además, es recomendable que los autores presenten no solo medidas relativas (RR o HR), sino también medidas absolutas, como la diferencia de riesgos, que aportan mayor relevancia clínica.

En este apartado basta con reconocer que las cohortes requieren análisis que consideren la temporalidad y el tiempo de seguimiento. Más adelante, en el capítulo de interpretación estadística, se abordará en detalle cómo calcular estas medidas, cómo interpretarlas en la práctica clínica y qué errores comunes deben evitarse.

5. EVALUACIÓN DE LA VALIDEZ Y RELEVANCIA

Una vez comprendida la metodología, el siguiente paso para el lector crítico es evaluar si el estudio tiene validez interna (es decir, si los resultados reflejan la realidad dentro de la población estudiada) y validez externa (si esos resultados pueden aplicarse a otros contextos, incluidos los pacientes propios del lector). Al mismo tiempo, debe juzgar si los hallazgos son clínicamente relevantes, más allá de la significancia estadística.

5.1 VALIDEZ INTERNA

La validez interna depende de que el estudio haya controlado adecuadamente los sesgos y la confusión. Un artículo de cohorte es sólido cuando:

- Los grupos comparados estaban libres del desenlace al inicio y fueron definidos de manera clara.
- La exposición fue medida de forma precisa y homogénea.
- Los desenlaces fueron definidos con criterios diagnósticos objetivos y evaluados de manera uniforme.
- El seguimiento fue suficientemente largo, con pérdidas mínimas y explicadas en detalle.
- Los confusores más relevantes fueron identificados, medidos y ajustados en el análisis.

5.2 VALIDEZ EXTERNA

La validez externa hace referencia a la generalización de los hallazgos. El lector debe preguntarse si los participantes de la cohorte se parecen a los pacientes de su práctica clínica. Un estudio realizado en un hospital especializado de Estados Unidos puede tener resultados difíciles de aplicar en una comunidad rural de Bolivia, y viceversa. En nuestro caso clínico, los hallazgos de una cohorte rural sobre humo de leña y EPOC tendrían una validez externa muy alta para poblaciones latinoamericanas.

5.3 RELEVANCIA CLÍNICA

No todo hallazgo estadísticamente significativo es clínicamente importante. El lector debe fijarse en la magnitud del efecto y en las medidas absolutas de riesgo. Un riesgo relativo de 2.0 puede sonar impactante, pero si la incidencia absoluta es muy baja, el impacto real en salud pública puede ser limitado. Por el contrario, una diferencia de riesgos de 15% entre expuestos y no expuestos al humo de leña tendría implicaciones directas en prevención y políticas sanitarias, incluso si el intervalo de confianza es relativamente amplio, esto lo veremos a detalle en el acápite

https://doi.org/10.62486/978-9915-9928-3-9.Ch13

de interpretación estadística.

5.4 INTERPRETACIÓN EQUILIBRADA

Un buen lector crítico no se deja llevar únicamente por la significancia estadística. Analiza los intervalos de confianza (¿son estrechos o demasiado amplios?), evalúa si los resultados son consistentes con estudios previos, y considera la plausibilidad biológica del hallazgo. También se pregunta si los autores reconocieron las limitaciones de su diseño y si realizaron análisis de sensibilidad para explorar la robustez de sus resultados.

5.5 EN LA PRÁCTICA CLÍNICA

Finalmente, la evaluación de validez y relevancia se traduce en una pregunta práctica: ¿los resultados de este estudio cambian la manera en que debo aconsejar, diagnosticar o tratar a mis pacientes? En el caso de nuestro ejemplo, si la cohorte muestra que el humo de leña triplica el riesgo de EPOC y que la fracción atribuible poblacional es alta, el médico debería incorporar este conocimiento a la educación de sus pacientes y apoyar programas de transición energética en comunidades rurales.

En conclusión, un estudio de cohorte solo será útil si es válido y relevante. La validez asegura que los resultados no son producto del azar ni de sesgos metodológicos; la relevancia garantiza que esos resultados tienen sentido en la práctica clínica y en la salud pública.

6. INTERPRETACIÓN ESTADÍSTICA

Se debe comprender dos ideas fundamentales:

- Riesgo (incidencia acumulada): es la proporción de personas que enferman durante un periodo.
- Tasa de incidencia (persona-tiempo): es la frecuencia a la que ocurren casos por unidad de tiempo en riesgo (casos por persona-años).

Ambas miden “aparición de enfermedad”, pero no son iguales: el riesgo ignora el detalle fino del tiempo (solo mira el período completo), mientras la tasa lo incorpora (si alguien se va antes, aporta menos tiempo).

También necesitaremos tres ideas más:

- Censura: alguien deja de ser observado antes de presentar el evento. No es “fallo”; es tiempo conocido sin evento y luego desconocido.
- Supervivencia libre de evento: probabilidad de no haber presentado el desenlace a lo largo del tiempo (se ve con Kaplan–Meier).
- Riesgo instantáneo (hazard): velocidad a la que ocurre el evento en un instante; comparar hazards da el hazard ratio (HR).

Tabla 2. Incidencia acumulada, tasa de incidencia y razón de tasas de EPOC según exposición al humo de leña durante 12 años de seguimiento

Exposición	N al inicio	Casos nuevos de EPOC	Persona-años	Incidencia acumulada (%)	Tasa de incidencia (por 100 persona-años)	Razón de tasas (IC95%)
Expuestos	600	132	5,28	22.0	2.5	3.0 (2.1–4.3)
No expuestos	600	48	5,76	8.0	0.83	Ref.

Fuente: Elaboración propia

Para fijar ideas, usaremos la tabla 2 asociado al caso, a partir de esta podremos calcular las medidas fundamentales.

6.2 INCIDENCIA ACUMULADA (RIESGO)

La incidencia acumulada es la medida fundamental en los estudios de cohorte. Se define

https://doi.org/10.62486/978-9915-9928-3-9.Ch13

como la proporción de individuos inicialmente libres de enfermedad que desarrollan el desenlace durante un período determinado de seguimiento.

A diferencia de la prevalencia, que mide cuántos tienen la enfermedad en un momento dado, la incidencia acumulada refleja el riesgo de enfermar en el futuro. Por eso, es la medida más intuitiva para hablar de probabilidad y para establecer relaciones temporales entre una exposición y un desenlace.

Su fórmula básica es:

$$\text{Incidencia acumulada} = \frac{\text{Número de casos nuevos en el período}}{\text{Número de personas en riesgo al inicio}}$$

Sus características principales son:

- Periodo definido: siempre se calcula en relación a un lapso concreto (un año, cinco años, diez años, etc.).
- Población en riesgo: el denominador debe estar libre del desenlace al inicio.
- Medida de riesgo absoluto: expresa directamente la probabilidad de que un individuo desarrolle la enfermedad en el tiempo estudiado.
- Fácil comunicación: más comprensibles, responde a la pregunta: "De cada 100 personas, ¿cuántas enfermarán en un tiempo dado?".

La principal limitación de la incidencia acumulada es que no toma en cuenta las diferencias en el tiempo de seguimiento individual. Se asume que todos los participantes han sido observados durante todo el periodo, aunque en la realidad muchos pueden haberse perdido en el seguimiento, haber cambiado de exposición o haber fallecido por otras causas. En estos casos, la incidencia acumulada puede dar una idea parcial de la realidad. Cuando estas variaciones son importantes, es más recomendable utilizar la tasa de incidencia, que incorpora el tiempo exacto de observación de cada persona.

En nuestro caso clínico, siguiendo la tabla 2:

- La columna N al inicio indica el número de personas libres de enfermedad al comenzar el seguimiento; este es el denominador de la incidencia acumulada.
- La columna de casos nuevos representa el numerador tanto de la incidencia acumulada como de la tasa de incidencia.
- La incidencia acumulada expresa qué proporción de la cohorte enfermó al final de los 12 años
- Incidencia acumulada en expuestos = 132 ÷ 600 = 0.22 → 22%.
- Incidencia acumulada en no expuestos = 48 ÷ 600 = 0.08 → 8%.
- Interpretación clínica: durante 12 años de seguimiento, 22 de cada 100 personas expuestas al humo de leña desarrollaron EPOC, mientras que entre quienes no estaban expuestos solo 8 de cada 100 enfermaron.

Este resultado ya permite al lector crítico detectar una diferencia clara entre los grupos. Pero aún no dice nada sobre cuán grande es esa diferencia en términos relativos o absolutos, ni sobre su velocidad en el tiempo. Para eso necesitamos avanzar hacia otras medidas.

6.3 TASA DE INCIDENCIA (PERSONA-TIEMPO) O DENSIDAD DE INCIDENCIA

La tasa de incidencia, también llamada densidad de incidencia, mide la frecuencia con la que ocurren nuevos casos de una enfermedad en relación al tiempo total de observación aportado por la población en riesgo. Su fórmula es:

$$\text{Tasa de incidencia} = \frac{\text{Número de casos nuevos}}{\text{Suma del tiempo en riesgo de cada persona}}$$

A diferencia de la incidencia acumulada, que trata a todos los participantes como si hubieran sido seguidos exactamente el mismo periodo, la tasa de incidencia reconoce que en los estudios longitudinales no todos permanecen en observación el mismo tiempo. Algunas personas pueden abandonar el estudio, cambiar de exposición o morir por otras causas antes de presentar el desenlace. La tasa ajusta el denominador sumando únicamente el tiempo efectivamente observado.

El denominador, entonces, no es "número de personas", sino persona-tiempo: por ejemplo, 100 personas seguidas en 1 año equivalen a 100 persona-años. Si una persona se pierde al cabo de 6 meses, aporta solo 0.5 persona-año.

https://doi.org/10.62486/978-9915-9928-3-9.Ch13

Sus características principales son:

- Mide velocidad de ocurrencia: refleja qué tan rápido aparecen los eventos en la población.
- Incorpora pérdidas y censura: es especialmente útil cuando hay tiempos de seguimiento heterogéneos.
- Permite comparaciones válidas: cuando los grupos no tuvieron exactamente el mismo tiempo de observación, la tasa es más justa que el riesgo acumulado.
- Expresada en unidades de tiempo: casos por 100 persona-años, casos por 1,000 persona-meses, etc.

Su principal limitación es que es menos intuitiva que la incidencia acumulada. Decir que hay "2.5 casos por 100 persona-años" no se comprende tan fácilmente como "22 de cada 100 enfermaron en 12 años". Por eso, en la práctica clínica suele complementarse ambas medidas: el riesgo absoluto (incidencia acumulada) comunica probabilidad, mientras que la tasa de incidencia informa sobre velocidad.

En nuestro caso clínico retomando la tabla 2:

- El dato de persona-años refleja la suma del tiempo que cada participante estuvo realmente en riesgo, es decir, el período de observación efectivo aportado por cada individuo hasta enfermar, abandonar el estudio o concluir el seguimiento.
- La tasa de incidencia nos muestra la frecuencia de nuevos casos ajustada por tiempo de seguimiento.
- La razón de tasas nos muestra cuántas veces más rápido ocurre la enfermedad en un grupo respecto al otro, considerando el tiempo de observación aportado por cada participante.

La etiqueta "**Tasa de incidencia (por 100 persona-años)**" indica cuántos casos nuevos aparecen, en promedio, cuando sumas 100 años de observación entre las personas del grupo (por ejemplo: 100 personas seguidas 1 año, o 50 personas seguidas 2 años, etc.). No es una probabilidad, sino una velocidad de aparición del evento.

Analogía clínica para reforzar la comprensión: Piensa en la tasa de incidencia como en la velocidad de un auto:

- "60 km/h" no significa que el auto recorra 60 km cada hora siempre.
- Significa que, si mantiene esa velocidad promedio, tras 1 hora recorrerá unos 60 km; tras 2 horas, 120 km; y así sucesivamente.
- De la misma forma, "2.5 por 100 persona-años" es la velocidad promedio de aparición de casos en la cohorte. Nos dice que aparecerán 2.5 casos por cada 100 personas seguidas en 1 año.

Bajo esto, nuestros resultados se interpretarían de esta forma:

- **Expuestos:** 2.5 por 100 persona-años. Si se acumulan 100 años de observación en personas expuestas, se esperarían en promedio entre 2 y 3 casos nuevos de EPOC en ese período total, independientemente de cómo se distribuyan esos años entre los individuos.
- **No expuestos:** 0.83 por 100 persona-años. Si se acumulan 100 años de observación en personas no expuestas, se esperarían en promedio ≈0.8 casos (menos de 1).
- **Comparación directa (razón de tasas).**Dividir 2.5 entre 0.83 da ≈ 3.0. Eso significa que la enfermedad aparece tres veces más rápida entre los expuestos que entre los no expuestos, teniendo en cuenta el tiempo efectivo que cada persona estuvo en riesgo.

Por qué es útil esta lectura: La tasa integra pérdidas de seguimiento y diferentes duraciones de observación (cada quien aporta solo el tiempo real que estuvo "en riesgo"). Por eso es una comparación más justa cuando los tiempos no son iguales entre grupos.

Precauciones al interpretar:

- Una tasa no es una probabilidad acumulada: no debes "multiplicarla por años" para obtener riesgo a 12 años; para eso se usan curvas de Kaplan–Meier o modelos de supervivencia.
- En un artículo completo, busca los intervalos de confianza de estas tasas o de la razón de tasas (idealmente estimados con métodos de Poisson u otros), y confirma cómo se construyó el denominador de persona-años (quién se perdió, cuándo enfermó, si hubo cambios de exposición).
- Recuerda que esta comparación aún es cruda; más adelante debes ver si el estudio ajustó por confusores clave (p. ej., tabaquismo, edad, sexo, ocupación) y si la conclusión se mantiene.

Finalmente concluyendo la lectura critica correcta de la tabla 2 es:

- El riesgo absoluto de EPOC a 12 años fue del 22% en los expuestos frente al 8% en los no expuestos.
- La tasa de incidencia muestra con claridad que la enfermedad aparece con mayor velocidad en quienes usan leña.
- Ambas medidas —riesgo e incidencia— son consistentes: la exposición al humo de leña prácticamente triplica la carga de enfermedad.

https://doi.org/10.62486/978-9915-9928-3-9.Ch13

No obstante, el lector crítico debe ir más allá de la lectura numérica. Una pregunta obligada es si los autores ajustaron o no por posibles confusores. Esta información no suele aparecer en la tabla básica, sino en los análisis multivariables y en la discusión del artículo. Aun así, la tabla ya ofrece una orientación clara sobre el efecto crudo de la exposición.

Finalmente, conviene destacar que la tasa de incidencia es particularmente útil en estudios con tiempos de observación variables o con pérdidas durante el seguimiento, porque ofrece una comparación más justa que la incidencia acumulada. Sin embargo, para comunicar estos hallazgos a pacientes o a la comunidad, sigue siendo preferible traducirlos a probabilidades acumuladas o a cifras absolutas, que resultan más fáciles de comprender y de aplicar en la práctica clínica.

6.4 RIESGO RELATIVO (RR)

En estudios de cohorte, una de las medidas más utilizadas para comparar la ocurrencia de un evento entre expuestos y no expuestos es el RR, también llamado Razón de Riesgos.

El RR expresa cuántas veces más (o menos) probable es que ocurra el evento en los expuestos en comparación con los no expuestos. Se calcula dividiendo la incidencia acumulada del grupo expuesto entre la incidencia acumulada del grupo no expuesto:

$$RR = \frac{Riesgo\ en\ expuestos}{Riesgo\ en\ no\ expuestos}$$

El valor resultante se interpreta de la siguiente manera:

- RR = 1: no existe diferencia en el riesgo entre expuestos y no expuestos (la exposición no parece asociarse con la enfermedad).
- RR > 1: la exposición se asocia con un mayor riesgo de enfermedad. Por ejemplo, un RR = 2 indica que los expuestos tienen el doble de probabilidad de enfermar que los no expuestos.
- RR < 1: la exposición parece tener un efecto protector. Por ejemplo, un RR = 0.5 significa que los expuestos tienen la mitad del riesgo de enfermar que los no expuestos.

El RR se expresa casi siempre acompañado de un intervalo de confianza (IC95%), que nos indica la precisión de la estimación y su p valor que indica si la asociación podría deberse al azar. Es especialmente útil porque se basa en proporciones absolutas (incidencia acumulada) y permite comunicar el riesgo de forma sencilla:

- En investigación clínica, ayuda a cuantificar la fuerza de la asociación entre una exposición y un desenlace.
- En salud pública, permite estimar el impacto que tendría eliminar una exposición nociva o promover una intervención protectora.
- En práctica clínica, traduce la evidencia en probabilidades fáciles de explicar a pacientes y comunidades.

Sus principales ventajas radican en que es intuitivo: responde a la pregunta "¿cuántas veces más riesgo tienen los expuestos?". Es fácil de entender y comunicar. Sin embargo, sus limitaciones son: no tiene en cuenta el tiempo de seguimiento (a diferencia de la tasa de incidencia). En presencia de pérdidas de seguimiento o censura importante, puede subestimar o sobrestimar el efecto. Por todo ello siempre debe interpretarse junto con medidas absolutas (como la diferencia de riesgos) para tener una visión completa del impacto real.

En nuestro caso, se presenta la siguiente tabla:

Tabla 3. Riesgo acumulado y riesgo relativo (RR) de EPOC según exposición al humo de leña durante 12 años de seguimiento

Exposición	N al inicio	Casos nuevos	Incidencia acumulada (%)	Riesgo Relativo (IC95%)
Expuestos	600	132	22.0	2.75 (2.0–3.7)
No expuestos	600	48	8.0	Ref.

Nota: El intervalo de confianza (IC95%) es ilustrativo; en un artículo real se obtiene con métodos estadísticos (generalmente regresión binomial o modelos de Poisson con varianza robusta).

Para su Interpretación: Un RR de 2.75 significa que, en el periodo de 12 años, los expuestos al humo de leña tuvieron casi tres veces más riesgo de desarrollar EPOC que quienes no estaban expuestos, además su IC95% (2.0–3.7) no incluye el 1, por lo que podríamos concluir que la

https://doi.org/10.62486/978-9915-9928-3-9.Ch13

asociación no solo es fuerte, sino también estadísticamente significativa.

Como lector crítico, este resultado nos orienta de forma clara hacia la existencia de un riesgo aumentado, pero siempre debemos revisar si el análisis fue ajustado por factores de confusión (como tabaquismo, edad, nivel socioeconómico), ya que podrían explicar parte del efecto observado.

6.5 DIFERENCIA ABSOLUTA DE RIESGOS (RIESGO ATRIBUIBLE ABSOLUTO, RAR)

La Diferencia Absoluta de Riesgos (DAR), también llamada Riesgo Atribuible Absoluto (RAR), mide el exceso de riesgo que se observa en el grupo expuesto en comparación con el grupo no expuesto.

Se calcula restando el riesgo acumulado de los no expuestos al de los expuestos:

$$RAR = Riesgo\ en\ expuestos - Riesgo\ en\ no\ expuestos$$

Su interpretación general es de la siguiente forma:

- Si RAR > 0, la exposición añade riesgo: existe un exceso de casos atribuibles a la exposición.
- Si RAR < 0, la exposición reduce riesgo: refleja un efecto protector.
- Si RAR = 0, no hay diferencia entre los grupos.

Esta medida es muy valiosa porque expresa el impacto absoluto de la exposición. Mientras que el RR indica la "fuerza de la asociación" (por ejemplo, triplica el riesgo), la RAR muestra cuántos casos adicionales ocurren realmente por efecto de la exposición.

La RAR también se utiliza para calcular el Número Necesario a Tratar (NNT) o el Número Necesario para Dañar (NNH), conceptos muy útiles en la práctica clínica:

$$NNT/NNH = \frac{1}{RAR}$$

Estos valores responden a preguntas como: ¿cuántas personas debo tratar (o cuántas estarán expuestas) para que aparezca o se evite un caso adicional de la enfermedad?

Las ventajas de la RAR son:

- Muestra el impacto absoluto de la exposición (número de casos adicionales).
- Es muy útil para decisiones en salud pública, porque permite dimensionar el problema a nivel poblacional.
- Facilita la comunicación con pacientes: es más comprensible decir *"14 casos adicionales por cada 100 adultos"* que *"un riesgo 2.75 veces mayor"*.

Entre sus limitaciones:

- Depende de la magnitud del riesgo basal: el mismo RR puede dar diferentes RAR en poblaciones con distinta incidencia.
- Puede variar según el tiempo de seguimiento; si la cohorte fuera de 20 años, la RAR probablemente sería mayor.
- En presencia de censura importante o pérdida de seguimiento, puede estar sesgada si no se ajusta adecuadamente.

Para que lo apliquemos en nuestra cohorte de ejemplo, se presenta la siguiente tabla:

Tabla 4. Riesgo acumulado, diferencia absoluta de riesgos (RAR) y número necesario a dañar (NNH) de EPOC según exposición al humo de leña durante 12 años de seguimiento

Exposición	N al inicio	Casos nuevos	Incidencia acumulada (%)	RAR (%)	NNH
Expuestos	600	132	22.0	+14.0	7
No expuestos	600	48	8.0	Ref.	–

Fuente: Elaboración propia

Para su interpretación de RAR: 14% significa que, en 12 años de seguimiento, 14 de cada 100 adultos expuestos al humo de leña desarrollaron EPOC como exceso atribuible a la exposición, en comparación con los no expuestos.

Este dato se puede traducir a la práctica clínica de la siguiente forma:

- Por cada 100 adultos expuestos al humo de leña, habrá 14 casos de EPOC que no se hubieran

https://doi.org/10.62486/978-9915-9928-3-9.Ch13

producido si hubieran tenido condiciones de no exposición.

- Para su interpretación de NNH: por cada 7 adultos expuestos al humo de leña, 1 desarrollará EPOC atribuible a la exposición en 12 años.

Ahora si Imaginamos un ensayo clínico hipotético en el que se implementa ventilación mejorada en cocinas para reducir la exposición al humo de leña. Tras 12 años nos reporta un Numero Necesario a tratar (NNT) de 8, esto se interpreta que por cada 8 mujeres que reciben ventilación mejorada, se previene 1 caso de EPOC en 12 años.

Reflexión para el lector crítico:

- El NNH comunica el daño atribuible a una exposición en términos absolutos y prácticos.
- El NNT comunica el beneficio de una intervención de la misma manera.
- Ambas medidas son muy poderosas porque permiten a médicos, pacientes y autoridades traducir los números de un artículo a decisiones concretas.

6.6 HAZARD Y HAZARD RATIO (HR)

HAZARD

El hazard (también llamado tasa de riesgo instantánea) es la probabilidad de que ocurra el evento de interés en un instante de tiempo determinado, entre quienes todavía no lo han presentado.

- A diferencia de la incidencia acumulada (que mira la proporción de personas que enferman al final de un periodo) o de la tasa de incidencia (que mide la velocidad promedio de aparición de casos por unidad de tiempo), el hazard se enfoca en el momento exacto en que ocurre el evento.
- Puede pensarse como la "intensidad instantánea del riesgo": ¿qué tan probable es que una persona que hasta ahora está sana enferme justo en el siguiente momento?

Matemáticamente, el hazard se expresa como una tasa por unidad de tiempo, pero en la práctica casi nunca lo calculamos manualmente, sino que se estima a través de modelos de supervivencia, como el modelo de Cox.

HAZARD RATIO (HR)

El HR compara el hazard entre dos grupos (expuestos vs. no expuestos) y se lo interpreta de esta forma:

- HR = 1: No hay diferencia en el riesgo instantáneo entre grupos.
- HR > 1: El evento ocurre más rápido (o con mayor intensidad) en el grupo expuesto.
- HR < 1: El evento ocurre más lentamente en el grupo expuesto (efecto protector).

Por todo esto, debemos diferenciarlo con RR y tasa de incidencia

- RR (Riesgo Relativo): compara proporciones acumuladas de enfermos al final de un seguimiento.
- Tasa de incidencia: compara la velocidad promedio de aparición de casos.
- HR: compara la velocidad *instantánea* de ocurrencia a lo largo del tiempo.

Por eso, el HR es más adecuado cuando:

- El tiempo de seguimiento es largo.
- Hay censura (pérdida de seguimiento, muertes por otras causas).
- El riesgo no se distribuye uniformemente en el tiempo (por ejemplo, mayor riesgo en los primeros años).

Tabla 5. Análisis de supervivencia mediante modelo de Cox para riesgo de EPOC según exposición al humo de leña durante 12 años de seguimiento

Variable	HR	IC95%	p-valor
Exposición al humo de leña	2.8	2.0 – 3.9	<0.001
Edad (>50 años)	1.5	1.1 – 2.1	0.02
Tabaquismo	2.2	1.6 – 3.1	<0.001

Fuente: Elaboración propia

En nuestra cohorte: Supongamos que al aplicar un análisis de supervivencia con modelo de

https://doi.org/10.62486/978-9915-9928-3-9.Ch13

Cox encontramos los resultados de la tabla 5**,** su interpretación clínica será:

- En cada momento del seguimiento, los pacientes expuestos al humo de leña tienen 2.8 veces más probabilidad de desarrollar EPOC que los no expuestos.
- A diferencia del RR (que solo nos dice que al final del seguimiento las expuestas tienen casi 3 veces más riesgo), el HR nos informa sobre la dinámica temporal: no solo enferman más, sino que además lo hacen antes.

Reflexión crítica para el lector

- El HR es una de las medidas más poderosas porque integra la información del tiempo y permite análisis ajustados por múltiples variables.
- Sin embargo, su interpretación requiere entender el supuesto del modelo de Cox y Kaplan – Meier.
- En lectura crítica, si un artículo reporta un HR, conviene verificar que los autores hayan evaluado la proporcionalidad de riesgos; de lo contrario, la estimación podría ser engañosa.

6.7 KAPLAN–MEIER Y REGRESIÓN DE COX EN ESTUDIOS DE COHORTE

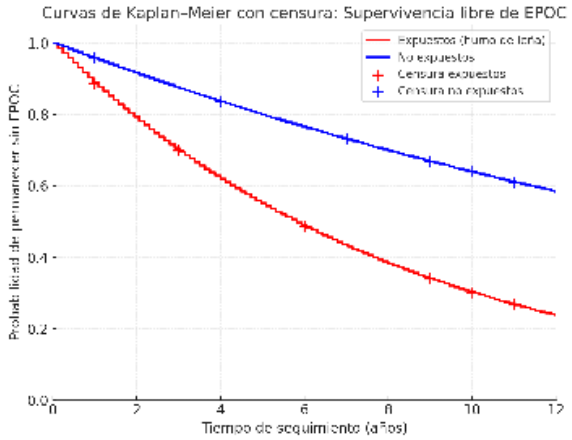
En los estudios de cohorte, cuando el interés se centra en analizar no solo cuántos participantes enferman, sino cuándo lo hacen, entran en juego dos herramientas estadísticas fundamentales: las curvas de Kaplan–Meier y el modelo de regresión de Cox. Ambas se utilizan de manera complementaria y permiten al lector crítico comprender la dinámica temporal del evento.

Las curvas de Kaplan–Meier son un método descriptivo que estima la probabilidad de permanecer libre del evento a lo largo del tiempo. Cada vez que ocurre un caso, la curva desciende en forma de escalón, mostrando el momento en que alguien desarrolla la enfermedad. Si un paciente se pierde en el seguimiento o termina el estudio sin enfermar, se registra como "censura" y se marca en la curva, generalmente con un pequeño símbolo. Esta técnica tiene la virtud de aprovechar al máximo la información del tiempo hasta el evento y de incluir a quienes no completan el seguimiento, algo muy común en cohortes largas. Cuando se comparan dos grupos, por ejemplo, expuestos y no expuestos, se construyen dos curvas y la diferencia global entre ellas se evalúa mediante la prueba de log-rank, que responde a la pregunta: ¿la supervivencia libre de enfermedad es distinta entre ambos grupos a lo largo del tiempo?

La regresión de Cox, también llamada modelo de riesgos proporcionales, va un paso más allá. No se limita a mostrar curvas descriptivas, sino que estima de manera cuantitativa la magnitud de la diferencia entre los grupos. Su resultado central es el HR, que indica cuántas veces más rápido ocurre el evento en un grupo comparado con otro en cualquier momento del seguimiento. Además, permite ajustar por múltiples variables de confusión (como edad, tabaquismo o nivel socioeconómico), lo cual le otorga gran poder analítico. Sin embargo, su validez depende de un supuesto fundamental: la proporcionalidad de riesgos, es decir, que la relación entre los hazards de los grupos se mantenga relativamente constante a lo largo del tiempo. Si este supuesto se viola, la interpretación del HR puede ser engañosa.

Veamos cómo estas herramientas se aplican a nuestro caso:

Figura 1. Curvas de Kaplan–Meier con censura para supervivencia libre de EPOC según exposición al humo de leña durante 12 años de seguimiento



https://doi.org/10.62486/978-9915-9928-3-9.Ch13

Al observar una curva de Kaplan–Meier, lo primero es identificar el **eje X**, que representa el tiempo de seguimiento (en nuestro caso, los años del estudio), y el **eje Y**, que indica la probabilidad de permanecer libre del evento (EPOC). La curva comienza siempre en 1.0, es decir, el 100% de los participantes sin enfermedad al inicio.

Cada vez que un participante desarrolla el evento, la curva desciende en un escalón. La magnitud del descenso depende de cuántos enferman en ese momento en relación con el número de personas que aún estaban en riesgo. Por ejemplo, un gran salto hacia abajo indica que varios individuos presentaron EPOC en ese intervalo.

Los símbolos de censura, representados aquí como cruces, marcan el momento en que un individuo deja de aportar información al estudio sin haber desarrollado el evento. Esto puede deberse a que abandonó el seguimiento, falleció por otra causa o simplemente terminó el periodo de observación sin enfermar. El lector debe notar que la presencia de censuras significa que esos datos se interrumpieron en ese instante.

La comparación entre curvas de diferentes grupos es clave. En nuestro caso, la curva de los pacientes expuestos al humo de leña desciende más rápidamente y se separa visiblemente de la curva de las no expuestos desde los primeros años. Esto significa que la probabilidad de permanecer libre de EPOC es consistentemente menor en los expuestos. El espacio entre curvas se mantiene e incluso se amplía hacia el final, lo que indica un efecto sostenido de la exposición.

El análisis estadístico complementario suele ser el log-rank test, que evalúa si la diferencia observada entre curvas es estadísticamente significativa. Si el valor p es menor de 0.05, podemos afirmar que las curvas difieren más de lo que cabría esperar por azar.

Como lector crítico, es importante preguntarse: ¿cuántas censuras hay y en qué momentos ocurren? Si la mayoría de los participantes fueron censurados en los últimos años, la curva en ese tramo tiene menos precisión y las conclusiones sobre el largo plazo son menos seguras. También conviene revisar si los autores reportan un HR ajustado por regresión de Cox, como mencionaremos a continuación.

Tabla 6. Modelo de Cox de riesgos proporcionales para tiempo a EPOC durante 12 años de seguimiento

Variable	Categoría (referencia)	HR ajustado	IC95%	p
Exposición al humo de leña	Sí (vs No)	2.70	2.00–3.65	<0.001
Edad	≥50 años (vs <50)	1.55	1.12–2.14	0.008
Tabaquismo	Actual/exfumador (vs nunca)	2.15	1.54–3.00	<0.001
Exposición ocupacional a polvos	Sí (vs No)	1.38	1.01–1.90	0.045
Sexo	Mujer (vs Varón)	1.20	0.86–1.68	0.27
Nivel socioeconómico	Bajo (vs Medio/alto)	1.22	0.90–1.66	0.19

Notas del modelo: n = 1,200; eventos = 180; mediana de seguimiento ≈ 11 años; prueba de log-rank global entre expuestos vs no expuestos p < 0.001. Supuesto de proporcionalidad de riesgos evaluado con residuos de Schoenfeld (global p = 0.21), sin violaciones relevantes.

La interpretación comienza por la fila de la exposición principal. Un HR ajustado de 2.70 indica que, en cualquier momento del seguimiento, las personas expuestas al humo de leña presentan una velocidad de ocurrencia de EPOC aproximadamente 2.7 veces mayor que las no expuestas, tras controlar por edad, tabaquismo, exposición ocupacional, sexo y nivel socioeconómico. Dicho de forma clínica, no solo hay más casos entre los expuestos; además, los casos aparecen antes. Esta conclusión no depende de un único momento del estudio, sino que resume todo el proceso temporal, que ya pudimos visualizar con las curvas de Kaplan–Meier.

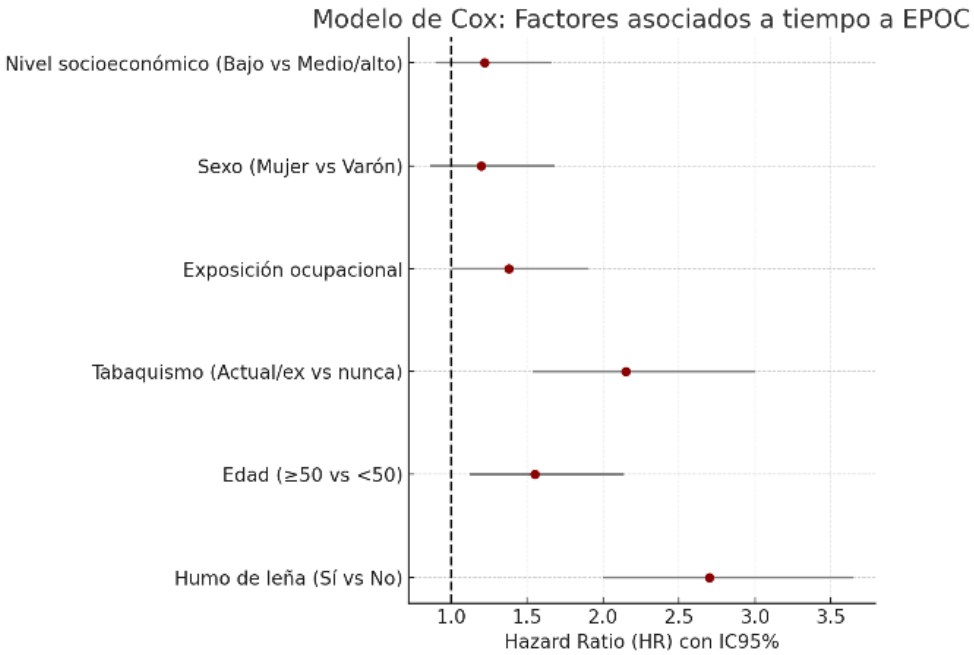
La edad muestra un efecto independiente: los participantes de 50 años o más tienen una aceleración del riesgo del 1.55 veces con respecto a los menores de 50, coherente con la historia natural de EPOC. El tabaquismo, como era esperable, duplica el hazard de EPOC; que el efecto del humo de leña se mantenga elevado aun ajustando por tabaco refuerza la plausibilidad etiológica de la exposición ambiental del hogar. La exposición ocupacional a polvos añade un incremento modesto pero significativo del hazard, lo que sugiere efectos aditivos en vías respiratorias cuando confluyen fuentes de partículas. Sexo y nivel socioeconómico, en este ejemplo, no alcanzan significación estadística; eso no implica que carezcan de relevancia clínica (sus intervalos de confianza IC cuentan con el 1), sino que, con estos datos y covariables, su contribución independiente al hazard no se demostró con precisión suficiente.

Es crucial recordar que un HR no es un riesgo relativo acumulado: cuantifica la relación entre tasas instantáneas de ocurrencia. Por ello, su interpretación siempre debe ir acompañada de la verificación del supuesto de proporcionalidad de riesgos. Aquí, la comprobación con residuos de Schoenfeld no sugiere violaciones; de haberlas, correspondería informar modelos

https://doi.org/10.62486/978-9915-9928-3-9.Ch13

estratificados, términos dependientes del tiempo o alternativas como tiempos medios de supervivencia restringidos (RMST). También es buena práctica que el artículo especifique cómo se codificaron las variables (por ejemplo, tabaquismo agrupado como nunca vs actual/ex), cómo se manejaron empates en el tiempo (método de Efron o Breslow) y si hubo estratificación por centro o comunidad para controlar confusión no medida a nivel contextual.

Figura 2. Modelo de Cox: factores asociados al tiempo a EPOC con estimaciones de hazard ratio (HR) e intervalos de confianza al 95% durante 12 años de seguimiento



El gráfico de bosque o *Forest plot* es una representación visual de los resultados de la regresión de Cox. En el eje vertical aparecen las variables analizadas, mientras que en el eje horizontal se ubica la escala de los HR. Cada punto corresponde al valor puntual del HR ajustado, y las barras horizontales representan su IC95%.

La línea vertical en 1.0 marca el valor de "no efecto". Si el intervalo de confianza cruza esa línea, significa que el efecto estimado podría ser nulo y, por lo tanto, el resultado no es estadísticamente significativo.

En nuestro ejemplo, la variable exposición al humo de leña presenta un HR de 2.70, con IC95% entre 2.00 y 3.65, como la barra de confianza no toca la línea de 1, este hallazgo es estadísticamente sólido. La variable sexo (HR 1.20, IC95% 0.86–1.68) y nivel socioeconómico bajo (HR 1.22, IC95% 0.90–1.66) no alcanzan significación, pues sus intervalos cruzan la línea de 1. Eso no implica que sean irrelevantes, sino que en este estudio los datos disponibles no permiten concluir con certeza sobre su efecto independiente.

Lo más importante para el lector crítico es entender que este gráfico resume en una sola mirada la dirección, magnitud y precisión de cada asociación. Mientras la tabla muestra los números, el Forest plot permite comparar rápidamente cuáles variables son relevantes, cuáles tienen gran incertidumbre y cuáles no muestran un efecto claro.

Finalmente, conviene recordar que estos HR son ajustados por las demás variables incluidas en el modelo. Esto significa que la asociación entre humo de leña y EPOC que se observa aquí no se debe a la edad, al tabaquismo u otros factores considerados, sino que refleja un efecto independiente de la exposición.

6.8 RAZÓN DE TASAS VS. RIESGO RELATIVO: ¿CUÁL USO?

https://doi.org/10.62486/978-9915-9928-3-9.Ch13

Tabla 7. Comparación entre riesgo relativo (RR) y razón de tasas de incidencia (IRR) en estudios de cohorte

Medida	¿Qué compara?	Ventajas	Limitaciones	Ejemplo aplicado al caso clínico
Riesgo Relativo (RR)	Probabilidad acumulada de enfermar en un periodo fijo (incidencia acumulada)	- Fácil de entender y comunicar a pacientes y responsables de salud. - Se traduce directamente en cifras absolutas y en medidas como NNT/NNH. - Útil en cohortes cerradas con seguimiento uniforme.	- No incorpora diferencias en el tiempo de seguimiento. - Sesgado si hay muchas pérdidas o censuras. - Menos útil si la población en riesgo varía mucho en el tiempo.	Tras 12 años: 22% expuestas vs 8% no expuestas → RR ≈ 2.75
Razón de Tasas de Incidencia (IRR)	Velocidad de aparición de nuevos casos, ajustada por persona-tiempo (persona-años)	- Más precisa cuando los seguimientos son desiguales o hay censura. - Permite comparar "la rapidez" del evento. - Útil en cohortes dinámicas o con tiempos de observación variables.	- Menos intuitiva para pacientes/comunidad. - Requiere cálculo detallado de persona-tiempo. - No siempre se reporta en artículos clínicos.	Tasa en expuestas: 2.5/100 persona-años vs 0.83/100 en no expuestas → IRR ≈ 3.0

Fuente: Elaboración propia

Aunque los valores son parecidos, no siempre ocurre así. En estudios con alta censura o tiempos de observación desiguales, las diferencias entre RR e IRR pueden ser notables. Como lector crítico, conviene identificar qué medida se reporta, si corresponde al diseño y, sobre todo, cuál es más útil para interpretar la aplicabilidad clínica.

En la práctica clínica, para conversar con un paciente o la comunidad, el RR (o medidas derivadas como el NNT/NNH) suele ser más comprensible. En cambio, para evaluar políticas de salud, comparaciones internacionales o estudios con tiempos de exposición muy variables, la IRR aporta una visión más ajustada.

6.9 ATRIBUCIÓN DE RIESGO: FAE Y FAP

Además de estimar cuánto aumenta el riesgo con una exposición, podemos calcular qué proporción de casos se debe realmente a esa exposición. Esto se conoce como atribución del riesgo, y es una herramienta poderosa tanto para la práctica clínica como para la salud pública. Existen dos formas principales de expresarlo:

a) Fracción Atribuible en Expuestos (FAE): indica qué fracción de la enfermedad observada en los expuestos puede atribuirse directamente a la exposición, suponiendo causalidad. En otras palabras, responde: *"¿Si eliminámos la exposición, qué porcentaje de casos en el grupo expuesto no ocurrirían?"*

Se calcula con la fórmula:

$$FAE = \frac{RR - 1}{RR} \times 100$$

b) Fracción Atribuible Poblacional (FAP): amplía la mirada a toda la población, considerando tanto expuestos como no expuestos. Responde: *"¿Si eliminámos la exposición en la población, qué proporción total de casos se evitarían?"*

Se calcula con:

$$FAP = \frac{Pe \times (RR - 1)}{Pe \times (RR - 1) + 1} \times 100$$

Donde Pe es la proporción de población expuesta.

La FAP depende no solo de la fuerza de asociación (RR), sino también de cuán frecuente es la exposición en la comunidad. Así, una exposición moderadamente riesgosa pero muy frecuente puede tener mayor impacto poblacional que una muy peligrosa pero rara.

En nuestro caso se presenta la tabla de esta forma:

https://doi.org/10.62486/978-9915-9928-3-9.Ch13

Tabla 8. Riesgo relativo, fracción atribuible en expuestos (FAE) y fracción atribuible poblacional (FAP) para EPOC según exposición al humo de leña

Medida	Estimación	IC95%
Riesgo relativo (RR)	2.75	1.90 – 3.98
Fracción atribuible en expuestos (FAE)	64%	–
Fracción atribuible poblacional (FAP)	47%	–

Fuente: Elaboración propia

Para interpretar la tabla: Durante el seguimiento de 12 años la FAE fue del 64%, indicando que casi dos tercios de los casos de EPOC entre los adultos expuestos podrían atribuirse al humo de leña.

A nivel poblacional, considerando que el 50% de la cohorte estuvo expuesta, la FAP fue del 47%. Esto implica que la eliminación de la exposición podría prevenir aproximadamente la mitad de los casos de EPOC en esta población.

Nota: En la mayoría de artículos, la FAE y la FAP rara vez aparecen con todos los detalles; normalmente se presentan directamente los porcentajes, a veces en tablas secundarias. El lector crítico debe saber de dónde vienen estos números (RR y prevalencia de la exposición) para verificar si están correctamente estimados y si su interpretación es válida.

6.10 RIESGOS COMPETITIVOS: PORQUE IMPORTAN

En la práctica clínica los pacientes pueden experimentar otros eventos que impidan o modifiquen la probabilidad de que ocurra el desenlace principal. Estos son los llamados riesgos competitivos. Un riesgo competitivo es cualquier evento que:

- Previene la ocurrencia del desenlace principal (ejemplo: un paciente fallece por infarto antes de desarrollar EPOC).
- Modifica la probabilidad de que ocurra el desenlace (ejemplo: un paciente es diagnosticado de otra enfermedad pulmonar crónica y cambia de ambiente o de exposición, reduciendo la probabilidad de llegar a EPOC clínico).

En teoría estadística, los riesgos competitivos afectan la incidencia acumulada: si se ignoran, se asume que todos los pacientes "siguen en riesgo" hasta el final, lo que sobreestima la probabilidad de enfermar.

En nuestra cohorte de 1,200 pacientes:

- La incidencia acumulada de EPOC fue 22% en expuestos y 8% en no expuestos.
- Sin embargo, durante el seguimiento algunos pacientes fallecieron por otras causas (ej. cáncer, enfermedad cardiovascular) antes de desarrollar EPOC.
- Estas muertes actúan como riesgos competitivos: esos pacientes ya no podían enfermar de EPOC, por lo que el análisis que no los considere podría inflar la incidencia real atribuida a la exposición.
- En cohortes largas, con pacientes adultos o con comorbilidades, este problema es frecuente y relevante.

Como lector crítico de un artículo, hay que revisar si los autores reconocieron la existencia de riesgos competitivos y usaron métodos adecuados, como curvas de incidencia acumulada en presencia de riesgos competitivos o modelos de Fine & Gray, en lugar de limitarse a Kaplan–Meier o a incidencias acumuladas crudas.

Ignorar los riesgos competitivos puede llevar a:

- Sobreestimación del riesgo de EPOC.
- Distorsión del efecto real de la exposición.
- Poca aplicabilidad clínica, ya que en la práctica muchos pacientes no alcanzarán a desarrollar el evento de interés porque otro evento se interpuso antes.

En nuestro ejemplo: los expuestos al humo de leña siguen mostrando mayor riesgo de EPOC, pero si un 10% muere por enfermedad cardiovascular antes, el "impacto real" de la exposición es distinto al que se calcula ignorando esas muertes.

6.11 AJUSTE POR CONFUSIÓN: MEDIDAS CRUDAS VERSUS AJUSTADAS

Cuando se interpreta un estudio de cohorte, es fundamental diferenciar entre las medidas crudas (sin ajuste) y las medidas ajustadas (tras controlar por confusores).

https://doi.org/10.62486/978-9915-9928-3-9.Ch13

Medidas crudas: son las primeras que reporta un estudio: riesgos relativos, tasas de incidencia o hazard ratios calculados sin considerar confusores. Estas medidas nos muestran el efecto "total" observado, pero pueden estar distorsionadas por variables externas.

Ejemplo con nuestros datos de tabla 3:

- RR crudo de EPOC en expuestos vs. no expuestos = 2.75 (IC95%: 2.0–3.7), indica casi 3 veces más riesgo de desarrollar EPOC.
- Sin embargo, todavía no sabemos si este efecto se debe exclusivamente al humo de leña o si parte de la asociación se explica por tabaquismo, edad, sexo u otros factores.

Medidas ajustadas: Para acercarse a la verdadera magnitud de la asociación, los investigadores usan modelos multivariables (regresión de Cox, regresión de Poisson, regresión logística según el diseño). Estos modelos permiten ajustar por confusores y reportar un efecto "depurado".

Ejemplo (hipotético, usando el modelo de Cox de nuestro caso en tabla 6):

- HR ajustado por edad, sexo y tabaquismo = 2.1 (IC95%: 2.0–3.65).
- Aquí vemos que el efecto del humo de leña sigue siendo significativo, pero menor que en la estimación cruda (de 2.75 baja a 2.10). Esto sugiere que parte del riesgo observado inicialmente estaba explicado por otros factores, en particular el tabaquismo.

Reflexión crítica: El lector crítico debe preguntarse siempre:

- ¿Los autores reportaron tanto medidas crudas como ajustadas?
- ¿Qué variables incluyeron en el ajuste? ¿Tienen base causal o fueron seleccionadas solo por significancia estadística?
- ¿Cambió mucho el estimador al ajustar? Un cambio importante indica confusión relevante.
- ¿Persistió la asociación tras el ajuste? Si se desvanece, es posible que el efecto crudo fuera explicado casi por completo por confusores.

6.12 MODIFICADORES DEL EFECTO

En un estudio de cohorte, no siempre la asociación entre exposición y desenlace es uniforme en todos los subgrupos de pacientes. A veces, la magnitud del efecto cambia según otra variable. En esos casos hablamos de modificación del efecto, también conocida como interacción.

Un modificador de efecto es una variable que altera la fuerza o dirección de la asociación entre exposición y desenlace.

- Si el efecto de la exposición es mayor en un grupo que en otro, decimos que hay interacción.
- A diferencia de un confusor, el modificador de efecto no distorsiona la relación: forma parte de la realidad clínica y es un hallazgo importante por sí mismo.

En términos prácticos:

- Confusión: un "artefacto" que debe eliminarse (ajuste estadístico).
- Modificación de efecto: una característica real de cómo actúa la exposición, que debe describirse y reportarse.

Identificación en estudios: La modificación de efecto se explora de dos maneras:

- Análisis estratificado: se calcula el RR o HR en distintos subgrupos (por ejemplo, hombres vs. mujeres, fumadores vs. no fumadores). Si los estimadores difieren sustancialmente, puede existir interacción.
- Modelos con términos de interacción: en regresión de Cox o Poisson se incluyen variables producto (ej. exposición × sexo). Si resultan significativas, indican modificación de efecto.

Tabla 9. Riesgo relativo (RR) de EPOC asociado a exposición al humo de leña: análisis global y por estratos

Estratificación	RR crudo	IC95%	p de interacción
Análisis global	2.75	1.90 – 3.98	–
Sexo			
Hombres	1.90	1.30 – 2.80	0.02
Mujeres	3.40	2.30 – 5.00	
Tabaquismo			
No fumadores	3.20	2.10 – 4.80	0.01
Fumadores	1.50	1.00 – 2.20	

Fuente: Elaboración propia

https://doi.org/10.62486/978-9915-9928-3-9.Ch13

En nuestro caso el análisis global mencionaba un RR crudo = 2.75, tras la estratificación por sexo, se evidencia una modificación de RR en hombres y mujeres, su interpretación: el efecto del humo de leña sobre el riesgo de EPOC es más fuerte en mujeres que en hombres. Esto refleja una modificación real del efecto: probablemente porque en contextos rurales son las mujeres quienes permanecen más tiempo en la cocina, acumulando mayor exposición. Así mismo se hizo la estratificación por tabaquismo, en esto se evidencia como interpretación: el humo de leña tiene un mayor impacto en no fumadores, ya que en fumadores parte del riesgo ya está "ocupado" por el tabaco.

Reflexión crítica: El lector crítico debe distinguir entre confusión y modificación de efecto.

- Si el efecto desaparece al ajustar, probablemente era confusión.
- Si el efecto se mantiene, pero varía en magnitud según un subgrupo, hablamos de modificación del efecto.

Este concepto es de enorme importancia clínica y en salud pública, porque permite identificar poblaciones particularmente vulnerables. En nuestro ejemplo, reconocer que las mujeres y los no fumadores tienen mayor susceptibilidad al humo de leña orienta las estrategias preventivas: programas de sustitución de combustibles y ventilación deben priorizar estos grupos.

Tabla 10. Comparación entre confusión y modificación del efecto en estudios de cohorte

Aspecto	Confusión	Modificación del efecto (interacción)
Definición	Distorsión aparente de la asociación exposición–desenlace causado por una tercera variable relacionada con ambos.	Cambio real en la magnitud o dirección de la asociación según los niveles de otra variable.
Naturaleza	Es un artefacto metodológico. No refleja la realidad.	Es un fenómeno real de la relación causal.
Efecto sobre la asociación	Hace que la asociación observada sea más fuerte, más débil o incluso cambie de dirección respecto a la asociación verdadera.	La asociación existe en todos los grupos, pero su magnitud varía (ej.: más fuerte en mujeres que en hombres).
Cómo se detecta	Comparando medidas crudas vs. ajustadas.	Estratificación (comparar riesgos relativos en subgrupos).
	Si cambian sustancialmente, había confusión.	Inclusión de términos de interacción en modelos multivariables.
Qué hacer	Debe controlarse en el análisis (ajuste estadístico, restricción, emparejamiento).	Debe reportarse y destacarse, pues refleja una característica real de la asociación.
Ejemplo en el caso clínico	Tabaquismo: si no se ajusta, el efecto del humo de leña sobre EPOC se sobrestima. Tras ajustar, el HR baja de 2.75 a 2.10.	Sexo: el RR del humo de leña es 3.4 en mujeres y 1.9 en hombres. El efecto existe en ambos, pero es más fuerte en mujeres.

Fuente: Elaboración propia

6.13: P-VALOR, INTERVALOS DE CONFIANZA Y RELEVANCIA CLÍNICA

En la lectura crítica de un estudio de cohorte, los números crudos (riesgos, tasas, RR o HR) no bastan por sí solos. Es indispensable interpretar p-valores e intervalos de confianza (IC95%), siempre en conjunto con la relevancia clínica de los hallazgos.

EL P-VALOR

Indica la probabilidad de observar los datos (o algo más extremo) si en realidad no hubiera asociación entre la exposición y el desenlace.

- Si $p < 0.05$, se considera que la asociación observada es poco compatible con el azar y se denomina "estadísticamente significativa".
- Si $p \geq 0.05$, no hay evidencia suficiente para descartar el azar

Sin embargo, el p-valor tiene limitaciones claras:

- No mide la magnitud del efecto.
- No refleja la relevancia clínica.
- Puede ser "significativo" incluso con efectos muy pequeños si el tamaño muestral es grande, o "no significativo" pese a efectos importantes si la muestra es pequeña.

Por eso, un lector crítico nunca debe basar su interpretación solo en el p-valor.

https://doi.org/10.62486/978-9915-9928-3-9.Ch13

INTERVALOS DE CONFIANZA (IC95%)

Proporciona un rango de valores dentro del cual es razonable que se encuentre la verdadera magnitud de la asociación, considerando la variabilidad muestral.

- Un IC95% estrecho indica precisión en la estimación.
- Un IC95% amplio indica incertidumbre.
- Si el IC95% incluye el valor nulo (RR=1 o HR=1), la asociación puede no ser estadísticamente significativa.

El intervalo de confianza es más informativo que el p-valor porque muestra simultáneamente:

- La dirección del efecto (mayor, menor o nula).
- La magnitud de la asociación.
- La precisión de la estimación.

RELEVANCIA CLÍNICA

Más allá de la significación estadística, el lector debe preguntarse si el hallazgo tiene impacto clínico o de salud pública.

- Un RR de 1.2 con p<0.001 puede ser estadísticamente sólido, pero clínicamente irrelevante si el exceso de riesgo es mínimo.
- En cambio, un RR de 2.5 con p=0.07 puede ser clínicamente relevante, aunque no alcance la significación estadística por falta de tamaño muestral.

La combinación de magnitud del efecto, intervalos de confianza y plausibilidad biológica es lo que determina si el hallazgo es útil para la práctica clínica.

Reflexión crítica: El mensaje para el lector crítico es claro:

- El p-valor solo responde a la pregunta "¿es probable que esto sea azar?".
- El IC95% responde a "¿qué tan grande y preciso es el efecto?".
- La relevancia clínica responde a "¿este hallazgo importa en la práctica?".

Solo integrando los tres niveles se logra una interpretación madura de los resultados de una cohorte.

Tabla 11. Comparación entre p-valor, intervalo de confianza e importancia clínica en estudios de cohorte

Aspecto	p-valor	Intervalo de confianza (IC95%)	Relevancia clínica
Definición	Probabilidad de observar los datos (o más extremos) si la hipótesis nula fuera cierta.	Rango de valores plausibles para la verdadera magnitud del efecto, considerando la variabilidad muestral.	Valor práctico o impacto real del hallazgo en la salud de pacientes y comunidades.
Pregunta que responde	¿Es probable que el hallazgo sea producto del azar?	¿Cuál es la magnitud y precisión del efecto observado?	¿Importa este hallazgo para la práctica clínica o la salud pública?
Valor nulo	p < 0.05 suele considerarse significativo (umbral convencional).	El efecto no es significativo si el IC incluye el valor nulo (RR=1, HR=1, diferencia=0).	No depende de umbrales fijos; requiere juicio clínico, epidemiológico y social.
Fortalezas	Fácil de calcular e interpretar en términos de significación estadística.	Aporta dirección, magnitud y precisión.	Permite contextualizar si un efecto "estadísticamente significativo" realmente cambia la práctica o las políticas.
		Más informativo que el p-valor solo.	
Limitaciones	No informa magnitud ni relevancia clínica.	Puede ser amplio si la muestra es pequeña, reflejando incertidumbre.	Subjetiva: depende del contexto, del desenlace y de la factibilidad de intervenir.
	Puede ser significativo con efectos mínimos en muestras grandes, o no significativo con efectos importantes en muestras pequeñas.	Requiere interpretación conjunta con la clínica.	
Ejemplo aplicado al caso	p < 0.001 → muy improbable que la asociación entre humo de leña y EPOC sea por el azar.	RR = 2.75 (IC95%: 1.90–3.98) → el riesgo verdadero se encuentra entre 1.9 y 4.0 veces mayor en expuestos.	Un incremento de casi 3 veces en el riesgo de EPOC es clínicamente muy relevante, pues justifica intervenciones preventivas.

Fuente: Elaboración propia

6.14 ERRORES FRECUENTES Y COMO EVITARLOS

Al leer un estudio de cohorte, incluso médicos con experiencia pueden caer en errores

https://doi.org/10.62486/978-9915-9928-3-9.Ch13

comunes de interpretación. Estos errores no siempre son técnicos: a veces provienen de trasladar de forma automática los números a la práctica clínica sin analizar el contexto. Reconocerlos es clave para evitar conclusiones engañosas.

1. Traducir el RR como si fuera un riesgo absoluto

Uno de los errores más frecuentes. Por ejemplo, decir que *"el humo de leña aumenta en 275% el riesgo de EPOC"* puede sonar alarmante, pero sin indicar la incidencia absoluta se pierde perspectiva. En nuestro caso, el riesgo pasó de 8% en no expuestos a 22% en expuestos: la diferencia real es del 14%. Lo correcto es buscar ambos: medida relativa y absoluta.

2. Confiar solo en el p-valor

Esto lleva a catalogar hallazgos como "significativos" o "no significativos" sin considerar su magnitud clínica. Un RR de 2.5 con p=0.06 en una cohorte pequeña puede ser muy relevante clínicamente, aunque no cruce el umbral estadístico. El lector crítico debe mirar también el intervalo de confianza y la plausibilidad clínica.

3. Ignorar el diseño y los tiempos de seguimiento

Muchos lectores pasan por alto si el seguimiento fue prospectivo o retrospectivo, o si hubo pérdidas importantes de pacientes. Estos detalles son fundamentales: un estudio con seguimiento desigual puede sesgar la incidencia acumulada y exagerar los riesgos. Siempre se debe preguntar: *¿cuánto tiempo estuvieron realmente los pacientes en riesgo?*

4. Ajustar por todo (o solo por lo "significativo")

Un error frecuente es pensar que mientras más variables se ajusten, mejor. Sin embargo, ajustar por variables irrelevantes puede distorsionar el resultado. El ajuste debe basarse en criterios causales, no en caprichos estadísticos.

5. Confundir confusión con modificación del efecto

A diferencia de la confusión, de cuyos efectos queremos deshacernos en nuestro análisis, la modificación de efectos es un hallazgo interesante en sí mismo, y lo reportamos. Para verificar la modificación del efecto, realizar un análisis estratificado.

6. Olvidar la precisión de las estimaciones

Un RR elevado con un intervalo de confianza muy amplio (ej. 1.2–9.0) es mucho menos confiable que un RR de 2.5 con IC95% 2.0–3.0. Los lectores suelen fijarse solo en el valor puntual y olvidan que la amplitud del intervalo refleja la incertidumbre de la estimación.

7. Desestimar riesgos competitivos

En cohortes largas, algunos pacientes mueren por otras causas antes de desarrollar el evento de interés. Si se ignora este hecho, se sobreestima la incidencia. El lector debe revisar si se aplicaron métodos para riesgos competitivos o al menos si se reconoce esta limitación.

8. Generalizar sin evaluar validez externa

Un error clásico es extrapolar resultados de una cohorte específica a toda la población. Si el estudio se hizo en comunidades rurales expuestas a biomasa, no podemos generalizar sin cautela a pacientes urbanos que usan gas. El lector crítico debe siempre contextualizar: *¿se parece esta cohorte a mis pacientes?*

Reflexión crítica: Los errores de interpretación no son detalles menores: pueden cambiar la manera en que aplicamos la evidencia en la clínica. Por eso, el lector crítico debe adoptar una actitud sistemática: verificar si el estudio reporta medidas absolutas y relativas, revisar los intervalos de confianza, distinguir confusión de modificación del efecto, y valorar si los hallazgos son clínicamente relevantes y aplicables a su población.

7. SESGOS Y LIMITACIONES EN ESTUDIOS DE COHORTE

Ningún estudio está libre de imperfecciones. Incluso las cohortes más cuidadosamente diseñadas pueden presentar sesgos (errores sistemáticos que distorsionan la asociación real) y limitaciones (factores que restringen la interpretación de los resultados). El lector crítico debe reconocerlos, porque de ello depende decidir si los hallazgos son confiables y aplicables.

7.1. SESGO DE SELECCIÓN

Ocurre cuando los grupos de comparación (expuestos y no expuestos) no son realmente comparables desde el inicio, o cuando las pérdidas en el seguimiento no ocurren al azar.

- Ejemplo en nuestro caso: si los pacientes expuestos al humo de leña eran en promedio de mayor edad o con peor estado de salud inicial, el riesgo de EPOC se vería aumentado por

https://doi.org/10.62486/978-9915-9928-3-9.Ch13

factores distintos a la exposición. Asimismo, si durante el seguimiento se pierde más del 20% de los expuestos porque migran a áreas urbanas, la estimación del riesgo puede ser sesgada.

7.2. SESGO DE INFORMACIÓN

Se produce cuando la exposición o el desenlace se miden de forma incorrecta. Puede ser diferencial (más error en un grupo que en otro) o no diferencial (igual error en ambos grupos).

- Ejemplo: si la exposición al humo de leña se evaluó solo con entrevistas y no con mediciones objetivas de partículas, puede haber clasificación errónea. Del mismo modo, si el diagnóstico de EPOC en zonas rurales se basó en síntomas sin espirometría, podrían clasificarse falsamente enfermos y sanos.

7.3. CONFUSIÓN

Ya hemos visto que ocurre cuando una tercera variable se relaciona tanto con la exposición como con el desenlace.

- Ejemplo: el tabaquismo es un confusor clave. Si no se mide y ajusta, parte del riesgo atribuido al humo de leña en realidad se debe al tabaco.

7.4. SESGO DE SUPERVIVENCIA O PREVALENCIA

En cohortes retrospectivas o ambispectivas, puede haber una “paradoja de supervivencia”: quienes ingresan al estudio ya sobrevivieron cierto tiempo a la exposición, lo que puede subestimar el efecto real.

- Ejemplo: incluir solo pacientes que aún viven en áreas rurales y llevan décadas expuestos al humo de leña puede excluir a los que murieron antes por EPOC severo.

7.5. RIESGOS COMPETITIVOS

Como vimos, los pacientes pueden fallecer por otras causas antes de presentar EPOC. Si se ignora este fenómeno, la incidencia acumulada puede sobreestimarse.

LIMITACIONES HABITUALES DE LAS COHORTES

- **Tiempo y costo:** Las cohortes prospectivas requieren años de seguimiento y gran inversión. Esto limita su factibilidad y tamaño de muestra.
- **Pérdidas de seguimiento:** Una limitación casi inevitable. Cuando son numerosas, la validez interna se debilita.
- **Exposiciones cambiantes:** Los pacientes pueden modificar hábitos durante el seguimiento (ej. cambiar de leña a gas), lo cual complica la clasificación.
- **Generalización restringida:** La validez externa depende de cuánto se parezca la cohorte a la población de interés. Cohortes rurales pueden no ser comparables a poblaciones urbanas.
- **Medición de variables:** No siempre es posible medir con precisión exposición, desenlace o confusores.

8. CHECKLIST

La lectura crítica de un estudio de cohorte requiere una estructura clara que guíe al lector paso a paso. Para elaborar este checklist hemos integrado las principales guías internacionales:

- STROBE (StrengtheningtheReportingofObservationalStudies in Epidemiology): que define los elementos mínimos que debe reportar un estudio observacional.
- CASP (CriticalAppraisalSkillsProgramme): que propone preguntas prácticas para valorar validez, resultados y aplicabilidad.
- JAMA Users’ Guides tothe Medical Literature: una de las guías más influyentes de la medicina basada en evidencia, centrada en utilidad clínica.

https://doi.org/10.62486/978-9915-9928-3-9.Ch13

- EQUATOR Network: iniciativa internacional que reúne estándares de calidad en publicación científica.

El resultado es una herramienta didáctica y práctica que permite, por un lado, revisar de manera exhaustiva todos los aspectos metodológicos de un estudio de cohorte, y por otro, contar con un esquema resumido para la consulta rápida en la práctica clínica o docente.

Tabla 13. Checklist completo para lectura crítica de estudios de cohorte

Dimensión	Preguntas clave
1. Pregunta de investigación	¿Está claramente definida y es clínicamente relevante?
	¿La estructura sigue un esquema PICO (Población, Exposición, Comparador, Desenlace)?
2. Selección de la cohorte	¿Se describen claramente los criterios de inclusión y exclusión?
	¿Los grupos de expuestos y no expuestos eran comparables al inicio?
	¿Se justifica el tamaño muestral y la potencia del estudio?
3. Medición de la exposición	¿La exposición se midió con métodos válidos y fiables (ej. cuestionario validado, biomarcadores, registros)?
	¿Se definieron claramente las categorías de exposición (sí/no, intensidad, duración)?
	¿Se consideraron cambios de exposición durante el seguimiento (ej. exposición dependiente del tiempo)?
4. Medición del desenlace	¿El desenlace está claramente definido y medido con criterios estándar (ej. diagnóstico de EPOC por espirometría)?
	¿La medición fue ciega o independiente respecto a la exposición?
	¿Los desenlaces son clínicamente relevantes?
5. Seguimiento	¿La duración del seguimiento fue adecuada para que aparezca el evento?
	¿Se reportó la proporción de pérdidas de seguimiento? ¿Fue <20%?
	¿Se analizaron las pérdidas para verificar si fueron diferenciales entre grupos?
	¿Se consideraron riesgos competitivos (ej. muerte por otra causa antes del desenlace)?
6. Análisis estadístico	¿Se presentaron medidas crudas (RR, HR, tasas) y ajustadas?
	¿Se ajustó por confusores clave (edad, sexo, tabaquismo, comorbilidades)?
	¿Se exploró la modificación de efecto (estratificación o interacción)?
	¿Se verificaron los supuestos de los modelos (ej. proporcionalidad de riesgos en Cox)?
7. Resultados	¿Se reportan p-valores e intervalos de confianza (IC95%)?
	¿Los resultados son clínicamente relevantes además de estadísticamente significativos?
	¿Se reportan medidas absolutas además de relativas (ej. diferencia de riesgos, NNT/NNH)?
8. Comparación y validez externa	¿Los resultados son consistentes con otros estudios?
	¿La población estudiada es comparable a la de mi práctica clínica?
	¿Las conclusiones pueden generalizarse con cautela?
9. Sesgos y limitaciones	¿El estudio reconoce sus limitaciones (sesgo de selección, información, confusión)?
	¿Los autores discuten cómo estas limitaciones pueden haber afectado los resultados?
Fuente: Elaboración propia	

Tabla 14. Checklist breve para lectura crítica de estudios de cohorte (versión aplicable)

✓	Pregunta clave
□	¿Es válida la cohorte?
□	¿La población fue bien definida, con exposición y desenlace medidos de forma confiable?
□	¿Hubo buen seguimiento (tiempo suficiente, pérdidas <20%)?
□	¿Se controlaron los confusores (edad, sexo, tabaquismo u otras variables clave)?
□	¿Cuál fue la magnitud y precisión del efecto (RR/HR con IC95% estrecho y plausible)?
□	¿Son aplicables los resultados a mis pacientes? ¿El efecto observado es clínicamente relevante?

Fuente: Elaboración propia

https://doi.org/10.62486/978-9915-9928-3-9.Ch13

9. UTILIDAD DE LOS ESTUDIOS DE COHORTE

Los estudios de cohorte son una de las herramientas más valiosas en la investigación clínica y epidemiológica. Aunque no alcanzan el mismo nivel de evidencia que los ensayos clínicos aleatorizados para demostrar causalidad, poseen características únicas que los convierten en instrumentos indispensables tanto para comprender la historia natural de las enfermedades como para orientar políticas de salud.

Una primera utilidad es que permiten seguir a los pacientes a lo largo del tiempo y describir cómo surge y progresa una enfermedad desde la exposición hasta el desenlace. Esto aporta información esencial sobre la historia natural. El ejemplo clásico es el Estudio de Framingham, que identificó factores de riesgo cardiovascular como la hipertensión, el tabaquismo y la hipercolesterolemia. En nuestro caso, el seguimiento de pacientes expuestos al humo de leña permite observar cómo la exposición continua produce inflamación crónica de la vía aérea, deterioro progresivo de la función pulmonar y finalmente EPOC.

Además, las cohortes permiten identificar factores de riesgo y también factores protectores. Una sola cohorte puede arrojar información sobre múltiples desenlaces relacionados con una misma exposición. En el caso de la biomasa, un estudio no solo puede medir su impacto en la EPOC, sino también su relación con cáncer de pulmón, neumonía, exacerbaciones respiratorias y mortalidad general. Esta capacidad de explorar múltiples desenlaces a partir de una misma población hace que el diseño sea especialmente valioso en entornos de salud pública.

Otra fortaleza central es que las cohortes permiten calcular directamente medidas de ocurrencia y asociación, como la incidencia acumulada, la tasa de incidencia, el riesgo relativo y el hazard ratio. Estas medidas son la base de toda la epidemiología analítica porque aportan información tanto absoluta como relativa sobre la magnitud del problema. Gracias a esto, los médicos y los decisores de salud pueden estimar no solo cuánto riesgo tiene un paciente, sino también qué impacto tendría una intervención preventiva en la comunidad.

Un aspecto clave es que el diseño de cohorte asegura la temporalidad: la exposición se registra antes del desenlace, lo que refuerza la plausibilidad causal. Ningún otro estudio observacional aporta con tanta claridad esta dimensión temporal. En nuestro ejemplo, el hecho de demostrar que los pacientes expuestos al humo de leña enferman más y además lo hacen antes que los no expuestos fortalece la hipótesis causal y reduce la posibilidad de que la asociación observada sea espuria.

Los estudios de cohorte también tienen un enorme valor para la salud pública, ya que generan evidencia que fundamenta políticas y programas preventivos. Así como las cohortes sobre tabaquismo fueron determinantes para las leyes de espacios libres de humo, las cohortes sobre biomasa pueden sustentar intervenciones como la sustitución de combustibles y la ventilación adecuada en viviendas rurales. Cuando se demuestra que casi la mitad de los casos de EPOC en una comunidad pueden atribuirse al humo de leña, se establece una base sólida para la acción sanitaria.

Siguiendo a individuos durante periodos prolongados, los estudios de cohorte aclaran como diversos factores contribuyen a los resultados de salud con el tiempo. Proporcionan pruebas sólidas que informa la formulación de políticas más efectivas y específicas de intervenciones y estrategias preventivas de salud.

Las cohortes son insumos esenciales para revisiones sistemáticas y metaanálisis, especialmente en áreas donde un ensayo clínico sería poco ético o inviable. La acumulación de cohortes de diferentes contextos refuerza la consistencia de los hallazgos y permite establecer conclusiones de mayor alcance.

10. AUTOEVALUACIÓN

PREGUNTAS DE SELECCIÓN MÚLTIPLE

1. ¿Cuál es la principal característica que diferencia a los estudios de cohorte de otros diseños observacionales?

a) Se realizan en un solo momento en el tiempo.
b) Se parte de la exposición y se sigue a los individuos hasta observar el desenlace.
c) Seleccionan participantes según su desenlace y analizan retrospectivamente la exposición.
d) No requieren grupo comparador.

2. ¿Qué medida de asociación puede calcularse directamente en una cohorte prospectiva?

https://doi.org/10.62486/978-9915-9928-3-9.Ch13

a) Odds ratio (OR).
b) Incidencia acumulada.
c) Sensibilidad y especificidad.
d) Prevalencia.

3. En una cohorte sobre humo de leña y EPOC se encontró un RR de 2.75 (IC95% 1.9–3.9; p<0.001). ¿Cuál es la interpretación más adecuada?
a) La asociación es estadísticamente significativa, pero clínicamente irrelevante.
b) Los expuestos tienen 2.75 veces más riesgo de desarrollar EPOC que los no expuestos.
c) Los expuestos tienen un 275% de riesgo absoluto de desarrollar EPOC.
d) El humo de leña no es un factor de riesgo independiente.

4. ¿Qué diferencia principal existe entre confusión y modificación del efecto?
a) La confusión refleja una característica real de la asociación y la modificación es un error metodológico.
b) Ambas deben eliminarse mediante ajuste estadístico.
c) La confusión es un artefacto que debe controlarse; la modificación del efecto es real y debe reportarse.
d) No existe diferencia entre ambos conceptos.

5. ¿Cuál de los siguientes es un ejemplo de riesgo competitivo en una cohorte de EPOC?
a) Un paciente cambia de exposición durante el seguimiento.
b) Un paciente muere por infarto antes de desarrollar EPOC.
c) Un paciente presenta exacerbaciones repetidas de EPOC.
d) Una paciente se pierde en el seguimiento por migración.

PREGUNTAS ABIERTAS

6. Explique con sus palabras la diferencia entre incidencia acumulada, tasa de incidencia y hazard. ¿En qué situación es más útil cada medida?

7. En el caso clínico del humo de leña y EPOC, ¿por qué es importante reportar tanto el riesgo relativo como la diferencia absoluta de riesgos?

8. ¿Qué información adicional aportan las curvas de Kaplan–Meier frente a la incidencia acumulada simple en un estudio de cohorte?

9. Mencione tres sesgos frecuentes en estudios de cohorte y explique cómo podrían afectar los resultados.

10. Imagine que un artículo reporta un RR crudo de 3.0, pero tras ajustar por tabaquismo y edad el RR baja a 1.5. ¿Cómo interpretaría esta diferencia?

RESPUESTAS:

1: b) Se parte de la exposición y se sigue a los individuos hasta observar el desenlace.

La característica fundamental del estudio de cohorte es que los participantes se clasifican según su exposición inicial, y posteriormente se observan en el tiempo para identificar la aparición del desenlace. Esto garantiza la secuencia temporal exposición → evento, lo que permite calcular incidencia y riesgo relativo

2: b) Incidencia acumulada.

En una cohorte, los individuos están libres de enfermedad al inicio y se siguen en el tiempo. Por ello, es posible calcular directamente la incidencia acumulada, la tasa de incidencia y el riesgo relativo. El Odds Ratio es característico de estudios de casos y controles, no de cohortes.

3: b) Los expuestos tienen 2.75 veces más riesgo de desarrollar EPOC que los no expuestos.

Un RR de 2.75 significa que el riesgo en el grupo expuesto es 2.75 veces el riesgo en el grupo no expuesto. El intervalo de confianza no incluye el valor nulo (1), por lo que la asociación es estadísticamente significativa. No implica un 275% de riesgo absoluto, sino una comparación relativa entre grupos.

4: c) La confusión es un artefacto que debe controlarse; la modificación del efecto es real y debe reportarse.

La confusión distorsiona la asociación real y debe ajustarse mediante estratificación o modelos multivariables. La modificación del efecto, en cambio, indica que la magnitud del efecto varía entre subgrupos (por ejemplo, mayor efecto en mujeres que en hombres), y debe informarse como hallazgo relevante.

https://doi.org/10.62486/978-9915-9928-3-9.Ch13

5: b) Un paciente muere por infarto antes de desarrollar EPOC.

Un riesgo competitivo es un evento que impide la ocurrencia del desenlace principal. Si un paciente fallece antes de desarrollar EPOC, ya no puede experimentar ese evento. Esto modifica la probabilidad acumulada de aparición del desenlace y debe considerarse en el análisis.

6. La incidencia acumulada representa la proporción de individuos que desarrollan el evento durante un período determinado. Es una medida de riesgo absoluto y responde a la pregunta: ¿qué probabilidad tiene una persona de enfermar en un intervalo específico? Es más útil cuando el seguimiento es relativamente uniforme y completo.

La tasa de incidencia incorpora el tiempo de observación y se expresa en persona-tiempo (por ejemplo, casos por 100 persona-años). Es especialmente útil cuando los participantes tienen tiempos de seguimiento variables o existe censura, ya que ajusta la frecuencia del evento por el tiempo real en riesgo. El hazard describe la tasa instantánea de ocurrencia del evento en un momento específico, condicionado a haber permanecido libre del evento hasta ese punto. Se utiliza en análisis de supervivencia y modelos como la regresión de Cox, siendo particularmente útil cuando interesa evaluar cómo cambia el riesgo a lo largo del tiempo.

7. El RR cuantifica la magnitud comparativa del efecto de la exposición. Indica cuántas veces mayor es el riesgo en el grupo expuesto respecto al no expuesto. La diferencia absoluta de riesgos (RAR) muestra el aumento real del riesgo en términos absolutos. Es fundamental para estimar el impacto clínico y poblacional, y permite calcular medidas como el número necesario a dañar (NNH). En el caso del humo de leña y EPOC, el RR informa sobre la fuerza de la asociación, mientras que la RAR permite dimensionar el impacto real en la población. Ambas medidas son complementarias y necesarias para una interpretación completa.

8. La incidencia acumulada simple ofrece una estimación global al final del seguimiento, pero no informa cuándo ocurren los eventos. Las curvas de Kaplan–Meier permiten visualizar la evolución temporal del riesgo, comparar grupos a lo largo del tiempo, identificar la velocidad de aparición de los eventos y mostrar la distribución de censuras. Además, permiten aplicar pruebas estadísticas como el log-rank para evaluar diferencias entre grupos. Por tanto, aportan una perspectiva dinámica del riesgo que la incidencia acumulada aislada no proporciona.

9. Un sesgo frecuente es el sesgo de selección, que ocurre cuando los grupos no son comparables al inicio. Puede generar estimaciones exageradas o atenuadas del efecto. El sesgo de información se produce cuando la exposición o el desenlace se miden de forma inexacta o diferente entre grupos, lo que puede distorsionar la asociación. La confusión ocurre cuando una variable externa, como el tabaquismo, está asociada tanto con la exposición como con el desenlace. Si no se controla adecuadamente, puede generar asociaciones espurias o infladas. Estos sesgos comprometen la validez interna del estudio.

10. Si el RR crudo es 3.0 y, tras ajustar por edad y tabaquismo, disminuye a 1.5, esto indica que parte de la asociación inicial estaba influenciada por confusión. El ajuste elimina la influencia de variables que distorsionaban la relación entre exposición y desenlace. El RR ajustado representa una estimación más precisa del efecto independiente de la exposición. Esta diferencia sugiere que edad y tabaquismo eran factores importantes que explicaban parcialmente la asociación observada en el análisis crudo.

https://doi.org/10.62486/978-9915-9928-3-9.Ch13

REFERENCIAS

Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al; STROBE Initiative. STROBE explanation and elaboration. Int J Surg. 2014;12(12):1500–24. Available at: https://doi.org/10.1016/j.ijsu.2014.07.014.
Critical Appraisal Skills Programme (CASP). CASP Cohort Study Checklist. Oxford: CASP; 2024 [cited 2025 Feb 8]. Available at: https://casp-uk.net/casp-tools-checklists/cohort-study-checklist/.
Guyatt G, Rennie D, Meade MO, Cook DJ, editors. Users' guides to the medical literature: A manual for evidence-based clinical practice. 3rd ed. New York: McGraw-Hill Education; 2015.
Lash TL, VanderWeele TJ, Haneuse S, Rothman KJ, editors. Modern epidemiology. 4th ed. Philadelphia: Wolters Kluwer; 2020. ISBN: 9781451193282.
Hernán MA, Robins JM. Causal inference: What if. Boca Raton: Chapman & Hall/CRC; 2020 [updated 2024 Jan 2; cited 2025 Feb 8]. Available at: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.
Zabert G, Patino CM, Ferreira JC. Prospective cohort studies and their contribution to public health and evidence-based medicine. J Bras Pneumol. 2025;51(4):e20250355. Available at: https://doi.org/10.36416/1806-3756/e20250355.
Wang X, Kattan MW. Cohort studies: Design, analysis, and reporting. Chest. 2020 Jul;158(1 Suppl):S72–8. Available from: https://doi.org/10.1016/j.chest.2020.03.014.
Dias S, Batalha MJ. Epidemiologic study designs: Cohort studies. Lusiadas Sci J. 2024;5(2). Available at: https://doi.org/10.48687/lsj.216.
Capili BA, Anastasi JK. Cohort studies. Am J Nurs. 2021 Dec 1;121(12):45–8. Available at: https://doi.org/10.1097/01.NAJ.0000803196.49507.08.
VanderWeele TJ, Ding P. Sensitivity analysis in observational research: Introducing the E-value. Ann Intern Med. 2017;167(4):268–74. Available at: https://doi.org/10.7326/M16-2607.
Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. J Am Stat Assoc. 1999;94(446):496–509. Availableat: https://doi.org/10.1080/01621459.1999.10474144.
Royston P, Parmar MKB. Restricted mean survival time: An alternative to the hazard ratio for the analysis of time-to-event data. BMC Med Res Methodol. 2013;13:152. Available at: https://doi.org/10.1186/1471-2288-13-152.
Pottegård A. Core concepts in pharmacoepidemiology: Fundamentals of the cohort and case–control study designs. PharmacoepidemiolDrugSaf. 2022;31(8):817–26. Availableat: https://doi.org/10.1002/pds.5482.
Perez-Guerrero EE, Guillen-Medina MR, Marquez-Sandoval F, Vera-Cruz JM, Gallegos-Arreola MP, et al. Methodological and statistical considerations for cross-sectional, case-control, and cohort studies. J Clin Med. 2024;13(14):4005. Available at: https://doi.org/10.3390/jcm13144005.
Xing Z, Zhang Y, Sun S, Cao W, Zhu J, Wang S, et al. Combined effect of ozone and household air pollution on COPD in young adults: A prospective cohort study. Thorax. 2024;79(5):449–57. Availableat: https://doi.org/10.1136/thorax-2023-220017.

Chapter 14 / Capítulo 14

Critical reading of scientific articles on health: A practical guide (Spanish Edition)
ISBN: 978-9915-9928-3-9
DOI: 10.62486/978-9915-9928-3-9.Ch14
Pages: 209-242

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Randomized Controlled Trials

Ensayos Clínicos Aleatorizados

Weymar Lehi Poma Luna
Ariel Sosa-Remón
Guiselle Carol Cabrera Morales
Helen Fernández Burgoa

1 Hospital Municipal Boliviano Holandes. La Paz, Bolivia.
2 Instituto Nacional de Oncología y Radiobiología. Habana, Cuba.
3 Hospital Oncológico, Caja Nacional de Salud. La Paz, Bolivia.
4 Ministerio de Salud y Deportes. Comité Regional de Integración Docente Asistencial e Investigación. Tarija Bolivia.

ABSTRACT

Introduction: Randomized controlled trials (RCTs) are the most methodologically rigorous experimental design for evaluating the efficacy and safety of healthcare interventions. Thanks to randomization, allocation concealment, and blinding, they minimize bias and establish causal relationships with high internal validity, making them the gold standard of Evidence-Based Medicine. Development: This chapter addresses the fundamental methodological principles of RCTs, including parallel, crossover, factorial, and cluster designs, as well as superiority, non-inferiority, and equivalence trials. Essential aspects such as participant selection criteria, randomization, blinding, outcome definition, follow-up, and statistical analysis using intention-to-treat, relative risk, absolute risk reduction, number needed to treat, and survival analysis are analyzed. Emphasis is placed on evaluating internal and external validity to ensure the reliability and applicability of the results. Conclusions: Randomized controlled trials (RCTs) represent the most robust tool for generating reliable scientific evidence to support clinical decisions and health policies. Their value depends not only on demonstrating therapeutic efficacy but also on methodological quality, transparency in execution, and the clinical relevance of the findings. Rigorous critical appraisal allows us to distinguish between statistically significant results and truly relevant benefits for patients.

Keywords: Randomized controlled trials; Randomization; Internal validity; Intention-to-treat analysis; Evidence-based medicine

RESUMEN

Introducción. Los ensayos clínicos aleatorizados (ECA) constituyen el diseño experimental de mayor rigor metodológico para evaluar la eficacia y seguridad de intervenciones sanitarias. Gracias a la aleatorización, el ocultamiento de la asignación y el enmascaramiento, permiten minimizar sesgos y establecer relaciones causales con elevada validez interna, siendo considerados el estándar de oro de la Medicina Basada en la Evidencia. Desarrollo. El capítulo aborda los principios metodológicos fundamentales de los ECA, incluyendo los diseños paralelos, cruzados, factoriales y por conglomerados, así como los ensayos de superioridad, no inferioridad y equivalencia. Se analizan aspectos esenciales como los criterios de selección de participantes, la aleatorización, el cegamiento, la definición de desenlaces, el seguimiento y el análisis estadístico mediante intención de tratar, riesgo relativo, reducción absoluta del riesgo, número necesario a tratar y análisis de supervivencia. De esta manera, se enfatiza la evaluación de la validez interna y externa para garantizar la confiabilidad y aplicabilidad de los resultados. Conclusiones. Los ECA representan la herramienta más robusta para generar evidencia científica confiable que sustente decisiones clínicas y políticas sanitarias. Su valor no depende únicamente de demostrar eficacia terapéutica, sino también de la calidad metodológica, la transparencia en la ejecución y la pertinencia clínica de los hallazgos. Una lectura crítica rigurosa permite distinguir entre resultados estadísticamente significativos y beneficios verdaderamente relevantes para los pacientes.

Palabras clave: Ensayos Clínicos Aleatorizados; Aleatorización; Validez Interna; Análisis por Intención de Tratar; Medicina Basada en la Evidencia

1. CASO CLÍNICO

La Sra. Julia, mujer de 60 años de edad, con diagnóstico de diabetes mellitus tipo 2 desde hace 12 años, acude a consulta en el Servicio de Medicina Interna para evaluación de su control metabólico. A pesar de recibir tratamiento farmacológico con metformina y sulfonilureas, presenta un control glucémico subóptimo, evidenciado por una hemoglobina glucosilada (HbA1c) de 8,6%. Como antecedentes relevantes, presenta hipertensión arterial adecuadamente controlada y una importante carga de riesgo cardiovascular familiar, dado que su hermano falleció por infarto agudo de miocardio a los 62 años. Durante la consulta, la paciente expresa interés por alternativas terapéuticas que, además de optimizar el control glucémico, contribuyan a reducir el riesgo de eventos cardiovasculares mayores. Esta inquietud resulta particularmente relevante considerando que la diabetes mellitus tipo 2 constituye un factor de riesgo independiente para enfermedad cardiovascular y que la reducción del riesgo cardiometabólico representa actualmente uno de los objetivos prioritarios del tratamiento integral de estos pacientes. En este contexto, el médico considera el posible uso de los inhibidores del cotransportador sodio-glucosa tipo 2 (iSGLT2), una clase farmacológica que ha demostrado beneficios clínicos más allá del control glucémico, incluyendo reducción de hospitalizaciones por insuficiencia cardíaca, enlentecimiento de la progresión de la enfermedad renal crónica y disminución de eventos cardiovasculares en poblaciones seleccionadas. Sin embargo, también reconoce que estos fármacos presentan limitaciones relacionadas con su costo, potenciales efectos adversos como infecciones genitourinarias y cetoacidosis diabética y la necesidad de identificar con precisión los subgrupos de pacientes que podrían obtener el mayor beneficio clínico en comparación con otras estrategias terapéuticas, como los agonistas del receptor del péptido similar al glucagón tipo 1 (GLP-1).

A partir de este escenario clínico surge la siguiente pregunta de investigación:

¿La adición de un inhibidor del cotransportador sodio-glucosa tipo 2 al tratamiento estándar es más eficaz y segura que el tratamiento convencional para reducir eventos cardiovasculares y mejorar los desenlaces clínicos en pacientes con diabetes mellitus tipo 2 y alto riesgo cardiovascular?

Responder esta interrogante requiere recurrir al diseño de investigación con mayor capacidad para establecer relaciones causales entre una intervención y sus desenlaces: los ensayos clínicos aleatorizados (ECA), considerados el estándar de referencia para evaluar la eficacia y seguridad de las intervenciones terapéuticas en salud.

2. OBJETIVOS

Los ECA son considerados el estándar de oro en investigación clínica porque permiten responder a una pregunta clave: ¿Funciona realmente una intervención y es segura en comparación con la práctica habitual o un placebo? Este estudio bien diseñado tiene varios objetivos:

- Evaluar la eficacia terapéutica: Busca determinar si la intervención produce el efecto deseado en la población de estudio.
- Evaluar la seguridad de la intervención: Además del beneficio, un ECA permite identificar los efectos adversos asociados al tratamiento.
- Establecer una relación causal: Gracias a la aleatorización y el enmascaramiento, los ECA son el diseño más confiable para decir que el efecto observado es consecuencia del tratamiento, y no de otros factores.
- Generar evidencia para guías y protocolos: Los resultados de los ECAs son la base para la medicina basada en la evidencia y permiten actualizar las guías clínicas internacionales. Es por ello que gracias a grandes ECAs, hoy sabemos que los iSGLT2 reducen hospitalizaciones por insuficiencia cardíaca y esta información ya se incorporó en guías de diabetes y cardiología.
- Optimizar la toma de decisiones clínicas: Un ECA no solo responde si un fármaco funciona, sino también qué magnitud de beneficio ofrece, cuántos pacientes hay que tratar (NNT) y si los beneficios superan los riesgos.

3. INTRODUCCIÓN

Los ECA representan el pilar metodológico de la Medicina Basada en la Evidencia (MBE). Desde mediados del siglo XX, este diseño se ha consolidado como el estándar de oro para

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

evaluar la eficacia y seguridad de intervenciones sanitarias, ya que, gracias a la aleatorización y el enmascaramiento, logra minimizar sesgos y establecer relaciones causales con un alto grado de validez interna.

Un ECA se caracteriza por asignar de manera aleatoria a los participantes a diferentes grupos de comparación (intervención, placebo o tratamiento estándar). Esta asignación al azar distribuye equitativamente los factores de confusión conocidos y desconocidos, lo que permite atribuir los resultados observados principalmente a la intervención evaluada. Por ello, cuando se realiza con calidad metodológica y ética, un ECA ofrece la evidencia más confiable para responder preguntas clínicas relevantes.En los últimos años, la importancia de los ECAs se ha reafirmado en múltiples contextos, desde los grandes ensayos de terapias cardiovasculares y oncológicas hasta los estudios pivotales que guiaron el uso de vacunas y tratamientos durante la pandemia de COVID-19. Sin embargo, también se han identificado desafíos y limitaciones que requieren atención:

- La necesidad de mejorar la calidad del reporte. Un metaanálisis reciente de más de 21,000 ensayos (1966–2024) mostró que, aunque la adherencia a la guía CONSORT ha aumentado, persisten deficiencias importantes en la descripción de la aleatorización, el ocultamiento de la asignación y los protocolos publicados.
- El impulso hacia los “large simple RCTs” o ensayos grandes y pragmáticos, apoyados en registros electrónicos de salud, que permiten evaluar de manera más eficiente y con costos reducidos intervenciones en poblaciones amplias.
- La búsqueda de mayor equidad y representatividad. La OMS publicó en 2024 una guía global que insta a incluir más países de ingresos bajos y medianos en los ensayos, para que la evidencia no se limite a contextos de alto recurso. Asimismo, la ICH E21, en desarrollo, busca revertir la exclusión histórica de mujeres embarazadas y lactantes, promoviendo una investigación más inclusiva.
- La necesidad de fortalecer la integridad científica. Algunos casos recientes de retractaciones han puesto de relieve la importancia de garantizar transparencia, ética e independencia en la conducción de los ECAs.

En este contexto, los ECAs no solo son fundamentales para establecer si un tratamiento funciona, sino también para determinar su magnitud de beneficio, su seguridad, su aplicabilidad en diferentes poblaciones y su relevancia clínica frente a costos y alternativas existentes.Para el médico, esto se traduce en un puente directo entre la investigación científica y la toma de decisiones en el consultorio o en la sala hospitalaria. Por ejemplo, gracias a los grandes ECAs multicéntricos sabemos hoy que los inhibidores del SGLT2 no solo reducen la glucosa, sino también las hospitalizaciones por insuficiencia cardíaca y la progresión de la enfermedad renal crónica, lo que ha transformado la práctica clínica y las guías internacionales de diabetes y cardiología.

Los ECA constituyen la herramienta más poderosa para evaluar intervenciones en salud. Su evolución reciente apunta hacia ensayos más simples, globales, inclusivos y transparentes, lo cual garantiza que la evidencia que generan sea más aplicable, confiable y justa para la diversidad de pacientes que atendemos en la práctica clínica.

4. METODOLOGÍA

Cuando un profesional de la salud se enfrenta a un artículo sobre un ensayo clínico aleatorizado, la sección de Metodología es el corazón del estudio. Allí se describe cómo fue diseñado, cómo se seleccionaron los participantes, de qué manera se aplicó la aleatorización, qué tipo de comparador se utilizó, cómo se definieron los desenlaces y qué estrategias se emplearon para garantizar la validez de los resultados. Un lector crítico debe detenerse con calma en este apartado, ya que la solidez del ensayo depende más del diseño que del hallazgo final.

4.1 TIPO DE DISEÑO

El tipo de diseño de un ensayo clínico determina la manera en que los participantes son distribuidos y cómo se comparan los resultados entre los grupos. Comprender esta característica es esencial para interpretar críticamente un artículo, ya que de ella dependen la validez de los hallazgos y la aplicabilidad clínica.

El diseño más común es el **ensayo paralelo**, en el cual cada participante es asignado a un grupo y permanece en él durante todo el estudio: uno recibe la intervención y el otro recibe

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

placebo o tratamiento estándar. Este modelo es fácil de interpretar y ofrece comparaciones claras entre grupos. En nuestro caso, el diseño más apropiado sería un ensayo paralelo donde un grupo reciba un iSGLT2 añadido al tratamiento habitual y el otro grupo reciba únicamente el tratamiento estándar. Así, al final del seguimiento se comparan los eventos cardiovasculares, la glucosa y los efectos adversos.

Existen también los **ensayos cruzados** (*crossover*), en los que los mismos participantes reciben las dos intervenciones en distintos momentos, separados por un "período de lavado" para evitar que los efectos de la primera intervención influyan en la segunda. Este diseño reduce la variabilidad, pues cada persona actúa como su propio control, pero no es adecuado cuando la enfermedad progresa con el tiempo o el efecto de la intervención es irreversible. En nuestro caso si la Sra. Julia participara en un crossover para evaluar un antihipertensivo, podría recibir primero el fármaco A, luego pasar un período de lavado y después recibir el fármaco B. No obstante, este diseño no sería apropiado para un iSGLT2, ya que sus efectos cardiovasculares y renales persisten en el tiempo y podrían alterar la comparación.

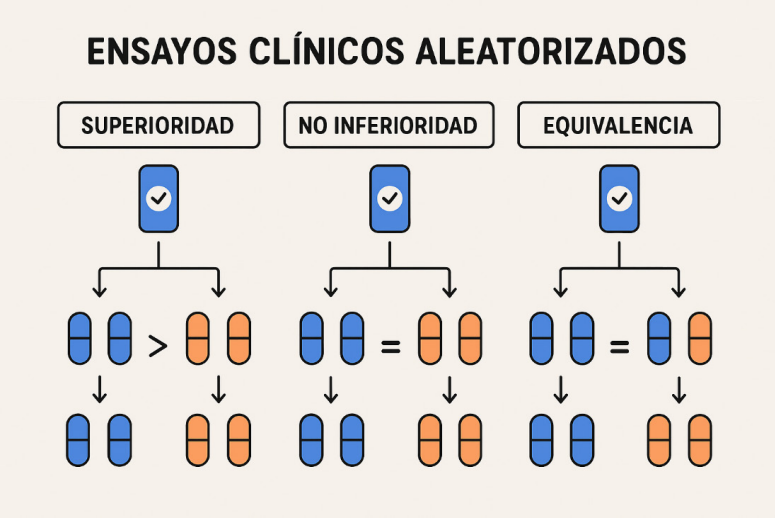
Otro diseño es la factorial, que permite estudiar más de una intervención en el mismo ensayo y en la misma población. Por ejemplo, un ensayo podría evaluar al mismo tiempo un iSGLT2 y una dieta hipocalórica, con cuatro grupos: solo fármaco, solo dieta, ambos o ninguno. Este diseño permite ahorrar recursos y responder dos preguntas en un solo estudio, pero puede complicar la interpretación si existe interacción entre las intervenciones.

Están los ensayos por conglomerados (clusterrandomizedtrials), en los que la aleatorización no se hace en individuos, sino en grupos completos, como hospitales, centros de salud o comunidades. Estos diseños son útiles cuando existe riesgo de "contaminación" entre los participantes o cuando la intervención se aplica a nivel colectivo. Por ejemplo, un ensayo que evalúe un programa de educación para mejorar el control de la diabetes podría aleatorizar centros de salud enteros en lugar de pacientes individuales.Para el lector crítico, identificar el tipo de diseño en la sección de metodología es el primer paso para valorar un ECA. Si el diseño está bien elegido y justificado, los resultados tendrán mayor validez y pertinencia clínica.

4.2 ENSAYOS DE SUPERIORIDAD, NO INFERIORIDAD Y EQUIVALENCIA

Un aspecto central en la metodología de un ensayo clínico es el marco de comparación que los investigadores adoptan para responder a su pregunta de investigación. En la lectura crítica, identificar si se trata de un ensayo de superioridad, no inferioridad o equivalencia es clave, porque cambia completamente la manera en que interpretamos los resultados.

Figura 1. Tipos de ensayos clínicos aleatorizados según su marco de diseño: superioridad, no inferioridad y equivalencia



Fuente: Elaboración propia

Ensayo de superioridad: Un ensayo de superioridad busca demostrar que la nueva intervención es más eficaz que el tratamiento de comparación. En este tipo de estudios, el resultado esperado es una diferencia significativa a favor del nuevo fármaco o procedimiento.

Ensayo de no inferioridad: En ocasiones, la meta no es demostrar que el nuevo tratamiento

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

sea mejor, sino que no sea clínicamente peor que el estándar. Esto es útil cuando la nueva opción ofrece otras ventajas: menos efectos adversos, administración más sencilla, menor costo o mejor adherencia. Aquí los investigadores establecen un margen (Δ), es decir, la diferencia máxima que aceptarían para considerar que el nuevo tratamiento no es peor que el comparador.

Ensayo de equivalencia: En los ensayos de equivalencia se busca demostrar que dos tratamientos tienen el mismo efecto clínico, dentro de un rango predefinido. Este marco suele aplicarse cuando lo que interesa es sustituir un medicamento por otro (por ejemplo, un genérico por el original) y confirmar que ambos son igualmente eficaces y seguros.

En el acápite de interpretación estadística se verá esto con más detalle

4.3 PARTICIPANTES Y CRITERIOS DE INCLUSIÓN/EXCLUSIÓN

Un elemento fundamental de la metodología en un ensayo clínico aleatorizado es la definición clara de quiénes fueron los participantes. Al leer críticamente un artículo, el primer paso es preguntarse: *¿los pacientes estudiados se parecen a los que yo atiendo en la práctica clínica?*. Para responder, los investigadores especifican dos aspectos clave: los criterios de inclusión y los criterios de exclusión.

Los criterios de inclusión definen las características necesarias para participar. Generalmente incluyen la edad, el diagnóstico confirmado de la enfermedad y, en algunos casos, parámetros de severidad o riesgo. Estos criterios aseguran que los participantes realmente pertenecen a la población objetivo del fármaco. Por ejemplo, en un ECA sobre iSGLT2, podrían incluir a adultos entre 40 y 75 años con diabetes mellitus tipo 2, HbA1c mayor a 7% y alto riesgo cardiovascular.

En cambio, los criterios de exclusión buscan proteger la seguridad de los pacientes y evitar factores que dificulten la interpretación de los resultados. Estos criterios limitan el riesgo de complicaciones graves y reducen variables que podrían sesgar los resultados. Por ejemplo, se podría excluir a pacientes con insuficiencia renal avanzada, mujeres embarazadas o aquellos con mala adherencia a controles médicos.

Para el lector crítico, la clave está en entender que cada criterio tiene un propósito, pero también un efecto sobre la validez externa del estudio, es decir, sobre su aplicabilidad. Si los criterios son demasiado estrictos, el estudio puede terminar incluyendo solo a un subgrupo muy seleccionado de pacientes, y sus resultados no necesariamente serán extrapolables a la práctica clínica habitual.

Aplicado al caso, un ensayo con criterios de inclusión que consideren diabetes tipo 2 mal controlada y alto riesgo cardiovascular sí la representaría. Pero si el ensayo excluyera a todos los pacientes mayores de 60 años o con hipertensión arterial, dejaría fuera a Julia y a muchos de nuestros pacientes habituales, limitando la utilidad clínica de los resultados.

4.4 ALEATORIZACIÓN Y OCULTAMIENTO DE LA ASIGNACIÓN

La **aleatorización** es el corazón de todo ensayo clínico aleatorizado. Su propósito es distribuir a los participantes de forma imparcial entre los grupos de intervención y control, de manera que cualquier diferencia observada en los resultados pueda atribuirse al tratamiento y no a otras características de los pacientes.

En la lectura crítica, lo primero que debe identificar el lector es cómo se realizó la aleatorización. Existen varios métodos de aleatorización que buscan garantizar la comparabilidad entre los grupos de estudio. La aleatorización simple asigna a los participantes de manera completamente aleatoria, similar a lanzar una moneda al aire o usar un generador de números aleatorios. Es fácil de implementar, pero en estudios con muestras pequeñas puede generar desequilibrios en el tamaño de los grupos o en características basales relevantes.

La aleatorización en bloques se utiliza para mantener un número similar de participantes en cada grupo a lo largo del estudio. Los pacientes se agrupan en bloques de tamaño fijo o variable (por ejemplo, 4 o 6), dentro de los cuales se distribuyen de forma equilibrada entre las intervenciones. Este método es especialmente útil en ensayos con reclutamiento progresivo, ya que evita diferencias importantes en el tamaño de los grupos en distintos momentos del estudio.

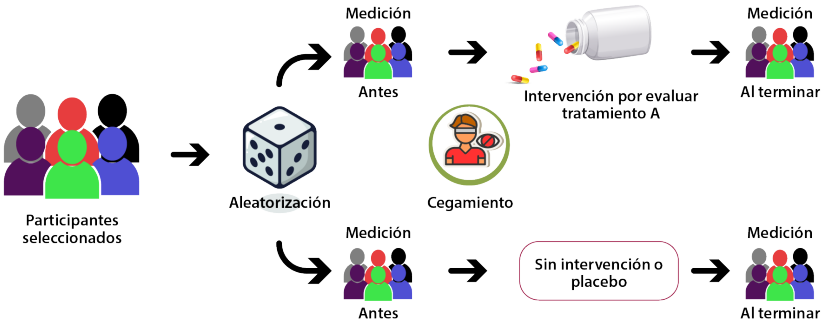
Por su parte, la aleatorización estratificada se emplea cuando existen variables clínicas relevantes que podrían influir en los resultados, como la edad, el sexo o la presencia de comorbilidades. En este caso, los participantes se clasifican previamente en subgrupos (estratos) según estas características, y luego se aleatorizan dentro de cada uno de ellos. De esta manera,

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

se asegura una distribución equilibrada de factores pronósticos clave entre los grupos de intervención.

En la práctica, es frecuente que los ensayos clínicos combinen estos métodos, por ejemplo, utilizando aleatorización estratificada con bloques, lo que permite mantener tanto el equilibrio en el tamaño de los grupos como en variables clínicas relevantes. Para el lector crítico, es fundamental verificar que el método de aleatorización esté claramente descrito y que sea adecuado para el diseño y tamaño del estudio.

Figura 2. Diseño clásico de un ensayo clínico aleatorizado



Fuente: Adaptado de Zurita-Cruz et al., 2018.

Ahora bien, la aleatorización por sí sola no es suficiente. Es igualmente importante el **ocultamiento de la asignación**, que consiste en impedir que los investigadores sepan de antemano a qué grupo se asignará un paciente. Este paso evita que, consciente o inconscientemente, los médicos incluyan a ciertos pacientes en un grupo con la esperanza de que se beneficien más o de que los resultados favorezcan al nuevo tratamiento. Un artículo de calidad debe explicar con claridad qué mecanismos se usaron para mantener oculta la secuencia de asignación, como sistemas computarizados centralizados o sobres sellados y opacos.

Volviendo a la Sra. Julia, imaginemos que participa en un ECA sobre iSGLT2. Si el investigador supiera antes de tiempo que ella será asignada al grupo control, podría —por buena intención— decidir no incluirla, pensando que su riesgo cardiovascular es alto y que debería recibir el nuevo medicamento. Esto introduciría un sesgo que invalidaría la comparación. Solo con una asignación estrictamente aleatoria y oculta podemos estar seguros de que los grupos son comparables y que las diferencias encontradas se deben al fármaco y no a la selección de pacientes.

Por ello, al leer un ensayo clínico, el médico debe detenerse a verificar dos puntos esenciales: ¿la aleatorización fue adecuada? ¿La asignación estuvo realmente oculta? Si la respuesta es afirmativa, entonces el estudio tiene un pilar sólido para sostener la validez de sus conclusiones.

4.5 ENMASCARAMIENTO

El enmascaramiento, también conocido como cegamiento, es otra piedra angular en los ECA. Consiste en mantener en secreto la asignación de los participantes a los grupos, con el fin de evitar que las expectativas o creencias de quienes participan en el estudio influyan en los resultados.

En la práctica, existen distintos niveles de enmascaramiento. El más simple es el **ciego simple**, en el que los pacientes no saben qué tratamiento reciben. Esto reduce el riesgo de que reporten mejoría solo por creer que están tomando un medicamento "nuevo". Más sólido es el **doble ciego**, donde ni los pacientes ni los investigadores que administran el tratamiento conocen la asignación. Así se evita que el médico, de manera consciente o inconsciente, trate a los pacientes de forma distinta según el grupo al que pertenecen. El nivel más riguroso es el **triple ciego**, en el cual incluso los analistas de datos desconocen la asignación hasta que el estudio finaliza.

¿Por qué es importante este detalle metodológico para el lector crítico? Porque la ausencia de cegamiento puede **generar sesgo de observación y efecto placebo.** Si un investigador sabe que su paciente está en el grupo experimental, puede evaluar con más optimismo sus síntomas o ser más riguroso en la búsqueda de efectos adversos en el grupo control. Del mismo modo, un paciente que sospecha estar en el grupo placebo puede reportar menos mejoría, aunque su evolución sea similar.

En nuestro caso, si la paciente se encuentra en un ensayo sobre iSGLT2, lo ideal sería que ella

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

no supiera si recibe el nuevo fármaco o un placebo y que su médico tratante tampoco lo supiera. De esa manera, cualquier diferencia en la reducción de eventos cardiovasculares o en la aparición de efectos adversos se debería realmente al medicamento y no a las percepciones o conductas de los involucrados.

Figura 3. Niveles de cegamiento en ensayos clínicos aleatorizados y su impacto en el riesgo de sesgo

Fuente: Elaboración propia

Para el lector crítico, debe preguntarse: ¿quién estuvo cegado en este ensayo? Si los pacientes, los médicos y los evaluadores de resultados estuvieron adecuadamente enmascarados, la confianza en los hallazgos aumenta considerablemente. Cuando el cegamiento no es posible —por ejemplo, en estudios quirúrgicos o en intervenciones conductuales—, el lector crítico debe verificar si se usaron desenlaces objetivos (como la mortalidad) o evaluadores independientes para minimizar el sesgo.

4.6 INTERVENCIÓN Y SEGUIMIENTO

En todo ensayo clínico aleatorizado, la **intervención** debe estar descrita con precisión, de modo que cualquier lector pueda entender exactamente qué tratamiento se administró, en qué dosis, por cuánto tiempo y bajo qué condiciones. Este nivel de detalle es esencial tanto para la reproducibilidad científica como para que el lector pueda valorar si la intervención es aplicable a su propia práctica clínica.

Un buen artículo especifica el nombre del fármaco o la estrategia terapéutica, la vía de administración, la frecuencia de las dosis y la duración del tratamiento. También debe indicar si hubo ajustes de dosis, protocolos de rescate o combinaciones con otros medicamentos de uso estándar. En nuestro caso, un ECA de iSGLT2 debería indicar, por ejemplo, que los participantes recibieron *dapagliflozina 10 mg por vía oral una vez al día*, añadida al tratamiento habitual con metformina y sulfonilurea, durante un período de tres años. Sin esta claridad, el lector crítico no podría interpretar con certeza los resultados.

El segundo aspecto es el seguimiento, es decir, la forma en que los investigadores acompañaron a los pacientes a lo largo del tiempo para medir tanto los beneficios como los riesgos del tratamiento. Un seguimiento adecuado incluye visitas periódicas para evaluar el desenlace principal (como la aparición de eventos cardiovasculares en nuestro ejemplo), la medición de variables intermedias (HbA1c, presión arterial, peso corporal), y el registro sistemático de eventos adversos. Cuanto más largo y completo sea el seguimiento, más confiables serán las conclusiones.

El lector crítico debe prestar especial atención a las **pérdidas de seguimiento**. Ningún ensayo logra que todos los pacientes permanezcan hasta el final. Si el ensayo ya está en marcha, y las pérdidas son tan masivas que comprometen el poder estadístico (1-B), lo correcto no es reemplazar sujetos "uno a uno", es mejor:

- Reportar honestamente las pérdidas en el diagrama de flujo (CONSORT).
- Discutir las limitaciones que esto genera en los resultados.
- Si el estudio es preliminar, podrías considerar ampliar el reclutamiento de forma global y transparente, pero nunca como una "sustitución" de individuos específicos, y siempre

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

previa aprobación del Comité de Ética e Investigación.

Es importante que el estudio reporte cuantos abandonaron, en que grupo y motivo. Una pérdida elevada o asimétrica puede sesgar los resultados. Si, por ejemplo, en un ensayo de iSGLT2 se pierde al 20% de los pacientes del grupo experimental por efectos adversos, y solo al 5% del grupo control, la validez del hallazgo se vería comprometida.

Finalmente, los artículos de calidad describen cómo evaluaron la **adherencia** al tratamiento. Esto puede hacerse con diarios de medicación, conteo de pastillas, medición de metabolitos en sangre o dispositivos electrónicos. En la práctica clínica sabemos que la adherencia es una de las principales barreras para obtener resultados; un ensayo que no la evalúe deja abierta la duda de si los beneficios observados se deben al fármaco o a que solo un grupo de pacientes cumplió estrictamente la indicación.

En conclusión, al leer la sección de intervención y seguimiento de un ensayo clínico, el lector critico debe preguntarse: ¿Está claramente descrita la intervención, incluyendo dosis, vía y duración?, ¿El, seguimiento fue suficiente para detectar beneficios y riesgos relevantes?, ¿Se reportaron las pérdidas y la adherencia de manera transparente?.

4.7 DESENLACES

Un ensayo clínico solo tiene sentido si responde a una pregunta concreta y medible. Esa respuesta se obtiene a través de los **desenlaces** o **outcomes**, que son los eventos o variables que los investigadores deciden medir para evaluar la eficacia y la seguridad de la intervención.

En la lectura crítica de un artículo, el primer paso es identificar el **desenlace primario**. Este es el más importante, porque el estudio fue diseñado y calculado en función de él. El desenlace primario debe ser clínicamente relevante, claramente definido y medible de manera objetiva. En un ECA sobre diabetes y riesgo cardiovascular, un desenlace primario apropiado podría ser la aparición de un evento cardiovascular mayor (infarto, accidente cerebrovascular o muerte cardiovascular). De esta manera, el estudio no solo mide el efecto sobre la glucosa, sino también sobre aquello que más preocupa al paciente y al médico: la reducción de complicaciones graves.

Además del desenlace primario, los ensayos suelen reportar **desenlaces secundarios**, que aportan información complementaria pero no deben interpretarse con el mismo peso. Estos desenlaces ayudan a entender mejor el perfil del tratamiento, aunque no fueron el foco principal del cálculo del tamaño muestral. En el caso, los desenlaces secundarios podrían incluir el cambio en la HbA1c, el peso corporal, la presión arterial o la función renal.

Otro aspecto importante son los **desenlaces de seguridad**, que deben estar descritos con tanto detalle como los de eficacia. Un buen ensayo clínico informa no solo los beneficios, sino también los riesgos: infecciones urinarias, deshidratación o hipotensión en el caso de los iSGLT2. Al leer un artículo, se debe preguntar: *¿se reportaron de manera completa los efectos adversos? ¿Se clasificaron por severidad y frecuencia?*

Un punto crítico en la lectura crítica es reconocer si los desenlaces son **duros (clínicos)** o **sustitutos (intermedios)**. Los desenlaces duros, como la mortalidad o los eventos cardiovasculares, tienen un impacto directo en la vida del paciente. Los desenlaces sustitutos, como la HbA1c, la presión arterial o el colesterol, predicen un beneficio potencial pero no siempre garantizan que se traduzca en una mejoría real en la salud. En la práctica clínica, ambos son útiles, pero los desenlaces duros tienen más peso para cambiar conductas y guías de práctica.

Finalmente, es importante que los desenlaces estén previamente definidos y registrados en un protocolo público (como ClinicalTrials.gov). Si un artículo cambia los desenlaces primarios después de iniciado el ensayo, o resalta solo los resultados positivos entre muchos desenlaces explorados, el lector debe ser cauteloso: puede tratarse de un sesgo de reporte.

En nuestro caso, un ensayo sólido debería tener como desenlace primario la reducción de eventos cardiovasculares mayores y, como secundarios, la mejoría en el control glucémico y la reducción de hospitalizaciones por insuficiencia cardíaca. Solo así podremos responder con certeza a la pregunta que motivó su consulta: *¿este medicamento no solo baja el azúcar, sino que además la protege de complicaciones graves?*

4.8 DURACIÓN DEL SEGUIMIENTO Y MANEJO DE PERDIDAS

La duración del seguimiento es uno de los aspectos que más influyen en la validez de los resultados. Un seguimiento demasiado corto puede mostrar una reducción transitoria en una

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

variable intermedia, pero no alcanzar a registrar los eventos clínicos relevantes ni los efectos adversos a largo plazo. Por el contrario, un seguimiento suficiente permite observar si el beneficio inicial se mantiene en el tiempo y si aparecen complicaciones inesperadas.

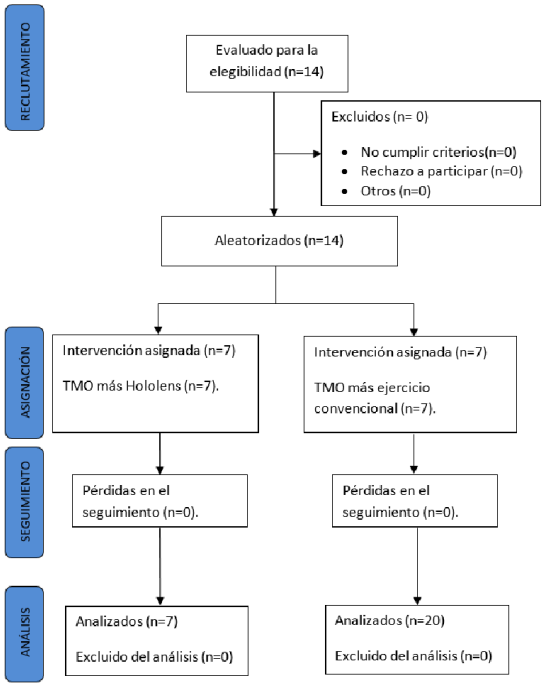
En la lectura crítica, el lector debe preguntarse: *¿el tiempo de seguimiento fue adecuado para detectar los desenlaces que se querían medir?* En un ensayo sobre fármacos para la diabetes, como en el caso de la Sra. Julia, un seguimiento de solo tres o seis meses podría mostrar una reducción en la HbA1c, pero no sería suficiente para evaluar si el medicamento reduce infartos, hospitalizaciones por insuficiencia cardíaca o la progresión de la enfermedad renal. Para este tipo de desenlaces cardiovasculares y renales, lo recomendable son seguimientos de al menos dos o tres años.

El otro aspecto crucial es el **manejo de las pérdidas de seguimiento**. En todo ensayo habrá pacientes que abandonen por efectos adversos, falta de motivación o traslado a otro centro de salud. Lo importante no es solo cuántos se pierden, sino también si la tasa de abandono es diferente entre los grupos. Una pérdida mayor en el grupo experimental, por ejemplo, podría ocultar problemas de seguridad o tolerancia.

Un buen artículo debe reportar con transparencia el número de pacientes que iniciaron el estudio, cuántos completaron el seguimiento y las razones de los abandonos. Lo ideal es que incluya un **diagrama de flujo CONSORT**, que muestre claramente el recorrido de los participantes desde la elegibilidad hasta el análisis final, como previamente habíamos mencionado.

En conclusión, al revisar la sección de metodología, el lector debe verificar que el tiempo de seguimiento sea suficiente para responder a la pregunta clínica y que las pérdidas hayan sido reportadas y analizadas con transparencia. Solo así podrá confiar en que los resultados reflejan lo que ocurriría en sus propios pacientes a largo plazo.

Figura 4. Diagrama de flujo de participantes según CONSORT en un ensayo clínico aleatorizado



Fuente: Elaboración propia

4.9 ANÁLISIS ESTADÍSTICO

Un ensayo clínico no solo se define por su diseño, su población y su seguimiento; también por la manera en que sus resultados serán analizados. Por eso, en la lectura crítica es fundamental revisar qué tipo de análisis estadístico fue planeado y si corresponde al diseño del estudio.

Existen dos enfoques principales: el análisis por intención de tratar (Intention-to-Treat, ITT) y el análisis por protocolo (Per Protocol, PP). El análisis por intención de tratarincluye a todos los pacientes en el grupo al que fueron asignados originalmente, aunque no hayan seguido el tratamiento hasta el final. Este enfoque preserva la aleatorización y refleja mejor lo que ocurre en

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

la práctica clínica real, donde no todos los pacientes cumplen al cien por ciento con la indicación. En cambio, el **análisis por protocolo** considera únicamente a aquellos que completaron el tratamiento según lo planeado, mostrando así el efecto "puro" de la intervención, aunque con mayor riesgo de sesgo porque excluye a quienes abandonaron o cambiaron de tratamiento.

Hoy en día, las guías metodológicas internacionales recomiendan que el análisis principal de un ECA sea por intención de tratar, ya que garantiza mayor validez interna y es más representativo de la práctica clínica. No obstante, muchos artículos también reportan el análisis por protocolo como complemento. El lector crítico debe comprobar si el estudio justifica adecuadamente qué población analizó y si existen diferencias importantes entre ambos enfoques.Además del tipo de análisis, es importante que el artículo especifique cómo se manejarán los **datos faltantes**. Estrategias como la imputación de valores o los análisis de sensibilidad permiten evaluar si los resultados se mantienen robustos pese a las pérdidas. Si el estudio omite este detalle, la interpretación de sus hallazgos puede ser poco confiable.

Al revisar la sección de metodología de un ECA, el lector debe fijarse en que el artículo describa: Qué población será analizada (ITT, PP o ambas). Cómo se tratarán los datos faltantes. Si el análisis planeado es coherente con los objetivos y el diseño del estudio.

Tabla 1. Elementos metodológicos esenciales en los Ensayos Clínicos Aleatorizados: descripción estandarizada y aplicación al caso

Elemento metodológico	Cómo se describe en un artículo científico	Aplicación al caso
Tipo de diseño	Se especifica si es paralelo, cruzado, factorial o por conglomerados, y si el objetivo es de superioridad, no inferioridad o equivalencia.	Ensayo paralelo, multicéntrico, de superioridad: iSGLT2 añadido al tratamiento estándar vs placebo.
Participantes	Se detallan los criterios de inclusión (diagnóstico, edad, factores de riesgo) y de exclusión (condiciones que pueden alterar seguridad o validez).	Adultos de 40–75 años con DM2, HbA1c ≥ 7%, con riesgo cardiovascular. Se excluyen pacientes con insuficiencia renal avanzada o embarazo.
Aleatorización y ocultamiento	Se describe el método de asignación (simple, bloques, estratificada), así como el sistema de ocultamiento (centralizado, sobres opacos, software).	Aleatorización en bloques y estratificada por edad y presencia de enfermedad cardiovascular; asignación centralizada en línea para evitar sesgo de selección.
Enmascaramiento (cegamiento)	Se especifica quién desconocía la asignación: pacientes, médicos, evaluadores o analistas (simple, doble o triple ciego).	Doble ciego: ni la Sra. Julia ni sus médicos saben si recibe iSGLT2 o placebo; los eventos cardiovasculares son evaluados por comité independiente.
Intervención	Se define el fármaco o tratamiento, dosis, vía, frecuencia, duración y protocolos de rescate.	Dapagliflozina 10 mg/día por vía oral, añadida al tratamiento estándar, durante 3 años. Ajustes según tolerancia.
Seguimiento	Se describe la frecuencia de visitas, qué variables se registraron (clínicas, bioquímicas, de seguridad) y cómo se evaluó la adherencia.	Consultas cada 3 meses, medición de HbA1c, presión arterial, función renal; adherencia evaluada con conteo de comprimidos y registros electrónicos.
Desenlaces	Se define un desenlace primario y los secundarios, junto con desenlaces de seguridad.	Primario: eventos cardiovasculares mayores (infarto, ACV, muerte cardiovascular). Secundarios: HbA1c, peso, presión arterial, progresión de ERC. Seguridad: infecciones urinarias y episodios de hipotensión.
Duración del seguimiento	Se justifica si el tiempo fue suficiente para evaluar beneficios y riesgos relevantes.	Tres años, adecuado para detectar diferencias en desenlaces cardiovasculares.
Pérdidas de seguimiento	Se reporta cuántos pacientes se retiraron y por qué, idealmente con diagrama de flujo CONSORT.	10% de abandonos en ambos grupos; causas reportadas: efectos adversos leves, falta de adherencia, traslado a otros centros.
Tipo de análisis estadístico planeado	Se aclara si el análisis será por intención de tratar (ITT), por protocolo (PP) o ambos, y cómo se tratarán los datos faltantes.	Análisis principal por ITT (todos los asignados se analizan en su grupo original). Secundario por PP para verificar consistencia. Datos faltantes manejados con imputación múltiple.

Nota: La tabla resume los componentes metodológicos que deben estar presentes en la publicación de un ensayo clínico aleatorizado según las guías internacionales (CONSORT 2010 y sus extensiones, SPIRIT 2013/2025, ICH E9[R1]). La primera columna identifica el elemento metodológico, la segunda describe cómo se debe reportar en un artículo científico y la tercera ejemplifica su aplicación en el caso ficticio de iSGLT2.

5. VALIDEZ

En la lectura critica, se debe preguntar si los resultados son **válidos**. En otras palabras,

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

¿podemos confiar en lo que este estudio encontró y aplicarlo a nuestros pacientes? La validez de un ensayo clínico se analiza en dos dimensiones: **interna** y **externa**.

5.1 VALIDEZ INTERNA

La validez interna se refiere a la confianza de que las diferencias observadas entre los grupos se deben realmente a la intervención y no a errores sistemáticos o sesgos. Es la base científica del estudio: si un ensayo carece de validez interna, sus conclusiones pierden valor. Para evaluar esta validez, el lector crítico debe revisar varios aspectos:

- Aleatorización adecuada: asegura que los grupos sean comparables desde el inicio.
- Ocultamiento de la asignación: evita el sesgo de selección.
- Cegamiento: reduce la influencia de las expectativas de pacientes e investigadores.
- Seguimiento completo y balanceado: minimiza el sesgo por pérdidas.
- Análisis coherente con el diseño: un ITT preserva la aleatorización y refleja mejor la práctica clínica.

5.2 VALIDEZ EXTERNA

La validez externa, también llamada generalización o aplicabilidad, se refiere a qué tanto los resultados de un ensayo pueden extrapolarse a la práctica clínica cotidiana. Un estudio puede tener una validez interna impecable pero ser poco útil si la población estudiada no se parece a la que vemos a diario en nuestros hospitales y consultorios. Para analizar esta validez, el lector debe preguntarse:

- ¿Los criterios de inclusión y exclusión fueron tan estrictos que dejaron fuera a pacientes comunes (ancianos, embarazadas, personas con múltiples comorbilidades)?
- ¿Las condiciones del estudio se asemejan a la práctica real o son demasiado controladas?
- ¿El seguimiento fue lo bastante largo para captar los desenlaces clínicos relevantes?

5.3 EL EQUILIBRIO ENTRE VALIDEZ INTERNA Y EXTERNA

Un buen ensayo clínico debe aspirar a ambas: alta validez interna y adecuada validez externa. A veces, sin embargo, existe una tensión entre ambas. Los estudios muy estrictos logran gran control y reducen sesgos, pero sacrifican representatividad. Los ensayos pragmáticos, en cambio, buscan parecerse a la práctica real, pero corren mayor riesgo de variabilidad y pérdidas de seguimiento.

La lectura crítica consiste en identificar este equilibrio y decidir si los hallazgos del ensayo son confiables y aplicables al paciente que tenemos frente a nosotros.

Para el lector critico

- La validez interna responde a la pregunta: *¿puedo confiar en que los resultados son ciertos para esta población estudiada?*
- La validez externa responde a la pregunta: *¿puedo aplicar estos resultados a mis pacientes en la práctica clínica?*

6. INTERPRETACIÓN ESTADÍSTICA

Leer un ensayo clínico aleatorizado no significa solo verificar que fue bien diseñado: también es imprescindible entender sus resultados. La sección de *Resultados* en un artículo suele presentar tablas, gráficos y medidas estadísticas que pueden parecer complejas. Sin embargo, con un enfoque sistemático, cualquier lector puede aprender a interpretar correctamente los hallazgos y aplicarlos a su práctica clínica.

Para la correcta interpretación, se utilizarán estos datos:

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

Tabla 2. Resultados principales de un ensayo clínico aleatorizado (Control vs. iSGLT2) en pacientes con diabetes tipo 2 y alto riesgo cardiovascular (N=2000, seguimiento 36 meses)

Grupo	N	MACE (n, %)	Hosp. IC (n, %)	Mortalidad (n, %)	Inf. genital (n, %)	CAD (n, %)	Δ HbA1c (media ± DE)	Δ Peso (kg, media ± DE)	Δ PAS (mmHg, media ± DE)
Control	1000	136 (13.6%)	125 (12.5%)	71 (7.1%)	32 (3.2%)	2 (0.2%)	−0.59 ± 0.96	−0.37 ± 2.48	−1.1 ± 7.8
iSGLT2	1000	81 (8.1%)	68 (6.8%)	44 (4.4%)	64 (6.4%)	4 (0.4%)	−0.87 ± 1.00	−2.94 ± 2.99	−2.7 ± 8.0

Fuente: Elaboración propia

6.1 PRINCIPIOS GENERALES

El primer principio que un lector crítico debe recordar es que un resultado numérico nunca debe interpretarse de manera aislada. Una diferencia estadísticamente significativa no siempre significa que sea clínicamente relevante, y un resultado no significativo no necesariamente implica que el tratamiento no funcione.

Por ello, la lectura crítica exige responder dos preguntas distintas:

- ¿El resultado es estadísticamente significativo? (es decir, ¿podemos descartar que sea producto del azar?).
- ¿El resultado es clínicamente importante? (es decir, ¿el beneficio es suficiente para cambiar la práctica en beneficio del paciente?).

Un segundo principio es distinguir entre las dos herramientas estadísticas más frecuentes: el valor p y el intervalo de confianza (IC). El valor p indica la probabilidad de obtener los resultados observados si en realidad no existiera diferencia entre los grupos. Aunque tradicionalmente se considera "significativo" un $p < 0,05$, este umbral es arbitrario y no refleja la magnitud del efecto.

El intervalo de confianza, en cambio, ofrece más información: indica un rango dentro del cual se encuentra el valor real del efecto con un 95% de seguridad estadística. Un IC estrecho refleja mayor precisión; un IC amplio indica mayor incertidumbre. Además, el IC permite valorar tanto la significancia estadística como la relevancia clínica, porque muestra el rango posible de beneficios o riesgos.

6.2 MEDIDAS DE EFECTO DE VARIABLES DICOTÓMICAS (CRUDAS)

RIESGO RELATIVO

El Riesgo Relativo (RR) compara la probabilidad de que ocurra un evento en el grupo intervención frente al grupo control y se interpreta de esta forma:

- Si RR = 1, no hay diferencia.
- Si RR < 1, la intervención protege (reduce el riesgo).
- Si RR > 1, la intervención aumenta el riesgo.

Tabla 3. Riesgo relativo de eventos cardiovasculares mayores (MACE) a 36 meses en pacientes con diabetes tipo 2 y alto riesgo cardiovascular (N=2000)

Grupo	Eventos / N	Riesgo (%)	
Control	136 / 1000	13,6	
iSGLT2	81 / 1000	8,1	
Medida de efecto	Valor puntual	IC95%	p-valor
Riesgo Relativo (RR)	0,6	0,46 – 0,77	<0,001

Fuente:Elaboración propia

Para su interpretación crítica:

- Riesgo Relativo (RR): El RR de 0,60 significa que los pacientes tratados con iSGLT2 tuvieron un 40% menos de riesgo relativo de presentar un evento cardiovascular mayor en comparación con los pacientes del grupo control.
- Intervalo de confianza (IC95%): El IC va de 0,46 a 0,77, lo que significa que, con un 95% de confianza estadística, el efecto verdadero del fármaco se encuentra en ese rango. Como

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

el IC no cruza el valor nulo (RR=1), la diferencia es estadísticamente significativa. Además, el rango es clínicamente relevante: incluso en el peor escenario (RR=0,77), el iSGLT2 aún ofrece un beneficio del 23%.

- Valor p: El p < 0,001 confirma que la probabilidad de que esta diferencia se deba al azar es extremadamente baja.

En la práctica clínica, esto significa que si la Sra. Julia recibe iSGLT2, tiene entre un 23% y un 54% menos riesgo de sufrir un infarto, ACV o muerte cardiovascular que si continúa solo con el tratamiento convencional.

REDUCCIÓN ABSOLUTA DE RIESGO (RAR)

La RAR mide la diferencia en el riesgo de presentar un evento entre el grupo control y el grupo intervención. A diferencia del Riesgo Relativo, que expresa una proporción, la RAR muestra la diferencia concreta en porcentaje de pacientes afectados, lo cual facilita la traducción a la práctica clínica.

Tabla 4. Reducción absoluta del riesgo (RAR) de eventos cardiovasculares mayores (MACE) a 36 meses en pacientes con diabetes tipo 2 y alto riesgo cardiovascular (N=2000)

Grupo	Eventos / N	Riesgo (%)	
Control	136 / 1000	13,6	
iSGLT2	81 / 1000	8,1	
Medida de efecto	Valor puntual	IC95%	p-valor
RAR	5,50%	2,8% – 8,2%	<0,001

Fuente:Elaboración propia

Para su Interpretación crítica

- RAR = 5,5%: Esto significa que, por cada 100 pacientes tratados con iSGLT2 durante 3 años, se evitan 5,5 eventos cardiovasculares mayores que hubieran ocurrido con el tratamiento estándar.
- IC95% (2,8% – 8,2%): Estamos 95% seguros de que el beneficio real se encuentra dentro de ese rango. En el peor escenario, el iSGLT2 prevendría 2 a 3 eventos por cada 100 pacientes; en el mejor, hasta 8.
- p < 0,001: Confirma que la diferencia es estadísticamente significativa y muy poco probable que se deba al azar.

Si la Sra. Julia y 99 pacientes similares fueran tratados con iSGLT2 en lugar del tratamiento convencional, entre 3 y 8 de ellos evitarían un infarto, ACV o muerte cardiovascular en tres años gracias al nuevo fármaco.

REDUCCIÓN DE RIESGO RELATIVO (RRR)

La RRR expresa el beneficio de la intervención en términos relativos: indica qué proporción del riesgo en el grupo control fue reducida gracias al tratamiento. Es una medida muy frecuente en artículos porque suele mostrar efectos más llamativos que la RAR. Sin embargo, debe interpretarse con cautela: un porcentaje alto puede impresionar, pero sin la RAR y el NNT pierde contexto clínico.

Tabla 5. Reducción relativa del riesgo (RRR) de eventos cardiovasculares mayores (MACE) a 36 meses en pacientes con diabetes tipo 2 y alto riesgo cardiovascular (N=2000)

Grupo	Eventos / N	Riesgo (%)	
Control	136 / 1000	13,6	
iSGLT2	81 / 1000	8,1	
Medida de efecto	Valor puntual	IC95%	p-valor
RRR	40%	23% – 54%	<0,001

Fuente: Elaboración propia

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

Para su interpretación crítica

- RRR = 40%: El tratamiento con iSGLT2 reduce en un 40% el riesgo relativo de presentar un evento cardiovascular mayor comparado con el tratamiento convencional.
- IC95% (23% – 54%): Con 95% de certeza, el beneficio real está entre un 23% y un 54% de reducción relativa del riesgo.
- p < 0,001: La diferencia es estadísticamente significativa.

En el caso de la Sra. Julia, la RRR muestra de manera impactante el beneficio: "el iSGLT2 reduce un 40% el riesgo de infarto, ACV o muerte cardiovascular". Sin embargo, el médico crítico no debe quedarse solo con este número. Para tomar decisiones, es necesario complementarlo con la **RAR** y el **NNT**, que muestran cuántos pacientes se benefician en términos absolutos.

NÚMERO NECESARIO A TRATAR (NNT)

El NNT indica a cuántos pacientes hay que tratar con la intervención, durante un período específico, para evitar un evento adicional respecto al control. Importante que el NNT siempre debe ir acompañado del horizonte temporal (p. ej., a 36 meses) y, cuando sea posible, de su IC95%, que se obtiene invirtiendo los límites del IC de la RAR.

Tabla 6. Número necesario a tratar (NNT) para prevenir un evento cardiovascular mayor (MACE) a 36 meses en pacientes con diabetes tipo 2 y alto riesgo cardiovascular (N=2000)

Grupo	Eventos / N	Riesgo (%)		
Control	136 / 1000	13,6		
iSGLT2	81 / 1000	8,1		
Medida de efecto	Valor puntual	IC95% del NNT*	Horizonte	p-valor**
NNT	18	13 – 36	36 meses	<0,001

Fuente: Elaboración propia

Para su interpretación crítica

- Debemos tratar a 18 pacientes como la Sra. Julia con iSGLT2 durante 36 meses para prevenir 1 evento cardiovascular mayor adicional respecto al tratamiento estándar.
- Precisión: el IC95% 13–36 indica que, con la incertidumbre estadística del estudio, el "verdadero" NNT podría estar entre 13 (más favorable) y 36 (más conservador).
- Decisión clínica: el NNT traduce el beneficio a un lenguaje clínico tangible. Junto con el NNH (para infecciones genitales ≈ 31 en nuestro ejemplo) permite ponderar beneficio vs. riesgo y conversar con la paciente de forma clara.

Tip: el NNT debe estar reportado siempre con el tiempo ("a 36 meses") y no redondear al número más cercano, sino hacia arriba (regla estándar) para mantener una interpretación conservadora.

NÚMERO NECESARIO A DAÑAR (NNH)

El NNH indica a cuántos pacientes hay que tratar con la intervención, durante un periodo específico, para causar 1 evento adverso adicional respecto al control. Igual que con el NNT, el NNH siempre debe reportarse con su horizonte temporal (p. ej., a 36 meses) y, cuando sea posible, con IC95%, obtenidos invirtiendo los límites del IC de la diferencia absoluta de riesgos.

Tabla 7. Aumento absoluto del riesgo (AAR) y número necesario para dañar (NNH) en infecciones genitales a 36 meses en pacientes con diabetes tipo 2 y alto riesgo cardiovascular (N=2000)

Grupo	Eventos / N	Riesgo (%)		
Control	32 / 1000	3,2		
iSGLT2	64 / 1000	6,4		
Medida de efecto	Valor puntual	IC95%	Horizonte	p-valor
AAR	3,20%	1,3% – 5,1%	36 meses	<0,001
NNH	31	20 – 75*	36 meses	—

Fuente: Elaboración propia

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

Para su interpretación crítica

- Si tratamos 31 pacientes como la Sra. Julia con iSGLT2 durante 36 meses, 1 paciente adicional presentará infección genital en comparación con el tratamiento estándar.
- Precisión: el IC95% 20–75 indica la incertidumbre: en el escenario menos favorable, 1 evento adicional por cada 20 tratados; en el más favorable, por cada 75.

Decisión clínica: el NNH debe evaluarse junto con el NNT. En nuestro ejemplo: NNT (MACE) ≈ 18 a 36 meses, NNH (infección genital) ≈ 31 a 36 meses. Esto sugiere que, en promedio, el beneficio cardiovascular ocurre con mayor frecuencia que el daño infeccioso, aunque ambos deben discutirse con la paciente y prevenirse (educación, higiene, vigilancia de síntomas)

Traducir la estadística a la práctica clínica es esencial para la toma de decisiones compartidas. Cuando leas un ensayo clínico, utiliza el Número Necesario a Tratar (NNT) para visualizar el impacto real.

6.3 MEDIDAS DE EFECTO DE VARIABLES DICOTÓMICAS (AJUSTADAS)

Cuando el interés está en el tiempo al evento (no solo si ocurrió, sino *cuándo*), se utiliza el análisis de supervivencia: Se construyen curvas de Kaplan-Meier para visualizar la probabilidad acumulada de permanecer libre de evento, se aplica un modelo de regresión de Cox que ajusta por covariables (edad, sexo, factores de riesgo), el resultado es el Hazard Ratio (HR), que veremos continuación:

Concepto de Kaplan-Meier: Cuando un ensayo clínico estudia un desenlace dicotómico (por ejemplo, muerte, infarto, hospitalización), no siempre basta con saber cuántos pacientes tuvieron el evento al final del seguimiento. También importa cuándo ocurrió: ¿los eventos se acumularon rápidamente en los primeros meses o de forma más tardía?

Para responderlo, se utilizan las curvas de Kaplan-Meier, que muestran la probabilidad acumulada de estar libre del evento a lo largo del tiempo (figura 5).

- Cada caída en la curva representa que un paciente (o un grupo de pacientes) presentó el evento.
- La distancia entre las curvas indica la diferencia entre los grupos a lo largo del seguimiento.

Concepto de Hazard Ratio (HR): En los ensayos clínicos con desenlaces de tiempo hasta evento (muerte, infarto, hospitalización), no basta con saber cuántos pacientes tuvieron el evento. Es igualmente importante entender cuándo ocurrió.El HR es la medida estadística más usada en este contexto, puede entenderse como el riesgo instantáneo de que ocurra un evento en un momento dado, en pacientes que aún no lo han presentado y se lo calcula con la regresión de Cox que permite ajustar con covariables relevantes

Para su interpretación del HR:

- HR = 1: no hay diferencia.
- HR < 1: la intervención reduce el riesgo (ej. HR=0,70 → un 30% menos riesgo instantáneo)
- HR > 1: la intervención aumenta el riesgo.

Concepto de regresión de Cox: El HR se obtiene a través de un modelo de regresión de Cox, que permite analizar el tiempo hasta el evento, comparar grupos (intervención vs control) y ajustar por covariables (edad, sexo, comorbilidades, etc.), aumentando la validez del resultado.

Para entender todo esto presentaremos la siguiente tabla:

Tabla 8. Eventos cardiovasculares mayores (MACE) a 36 meses: análisis crudo y ajustado mediante regresión de Cox en pacientes con diabetes tipo 2 y alto riesgo cardiovascular (N=2000)

Grupo	Eventos / N	Riesgo acumulado (%)	RR (IC95%)	HR ajustado (IC95%)*	p-valor
Control	136 / 1000	13,6	Ref.	Ref.	
iSGLT2	81 / 1000	8,1	0,60 (0,46–0,77)	0,72 (0,58–0,89)	0,002

Fuente: Elaboración propia con apoyo de Inteligencia artificial

Para su interpretación crítica de la tabla 8:

- El RR (0,60) muestra la diferencia acumulada al final del seguimiento: iSGLT2 reduce los eventos de 13,6% a 8,1%.
- El HR (0,72) muestra la reducción del riesgo en cualquier momento del seguimiento, tomando en cuenta cuándo ocurrió cada evento y ajustando por covariables.
- Ambos resultados van en la misma dirección, pero el HR ofrece una visión más dinámica y ajustada.

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

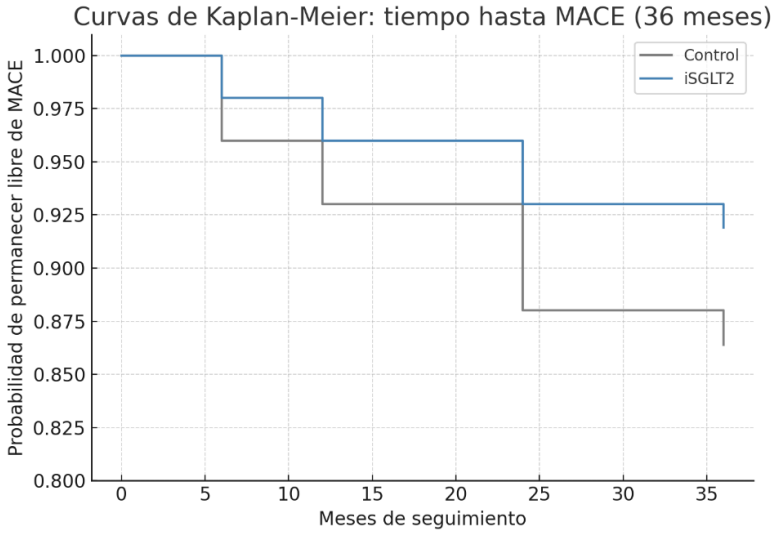
El lector crítico debe revisar siempre: Las curvas de Kaplan-Meier: ¿se separan temprano y se mantienen separadas?, El HR ajustado: ¿qué variables se incluyeron en el ajuste y están clínicamente justificadas?

El RR menciona *qué pasó al final del estudio*; el HR menciona *cómo se fueron acumulando los eventos a lo largo del tiempo*.

Para la interpretación critica de la figura 5:

- Se observa cómo las curvas comienzan a separarse desde los primeros meses y se mantienen diferentes hasta el final.
- La curva del grupo iSGLT2 se mantiene más alta → mayor probabilidad de permanecer libre de eventos.
- La diferencia acumulada concuerda con lo que vimos en la tabla (13,6% vs 8,1% de eventos).

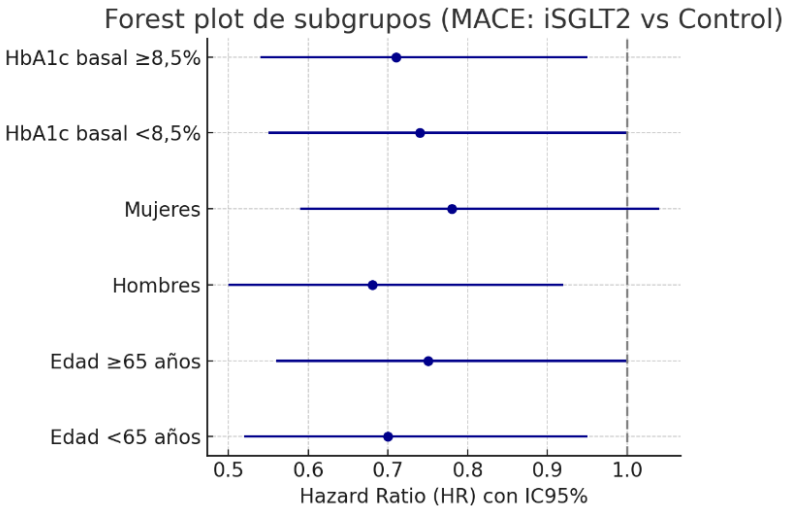
Figura 5. Curvas de Kaplan-Meier: tiempo hasta un evento cardiovascular mayor (MACE) en pacientes con diabetes tipo 2 y alto riesgo cardiovascular (seguimiento 36 meses, N=2000)



Nota: El análisis de supervivencia se realizó mediante el método de Kaplan-Meier; comparación entre grupos con prueba de log-rank.

Asi mismo en algunos artículos prefieren presentar estos resultados por medio de una grafica de Forest plot como se muestra a continuación:

Figura 6. Forest plot de subgrupos: hazard ratio (HR) para eventos cardiovasculares mayores (MACE) a 36 meses en pacientes tratados con iSGLT2 vs. control



Fuente: Elaboración propia

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

En este tipo de gráficos, los cuadrados representan el efecto de cada subgrupo, mientras que el diamante (o el punto central en este caso) representa el resultado global del estudio.

Regla de oro: Si el diamante o sus brazos horizontales (intervalos de confianza) no tocan ni cruzan la línea vertical de la unidad (valor nulo = 1), el resultado es estadísticamente significativo. En este capítulo, al estar a la izquierda de la línea, demuestra que el iSGLT2 es superior al control en todos los análisis

Para su interpretación:

- Muestra que el beneficio de iSGLT2 (HR <1) es consistente en todos los subgrupos (edad, sexo, HbA1c basal).
- Algunos IC95% se acercan a 1, lo que refleja mayor incertidumbre en subgrupos más pequeños.
- En la práctica, estos gráficos se usan para reforzar la robustez y generalización de los resultados.

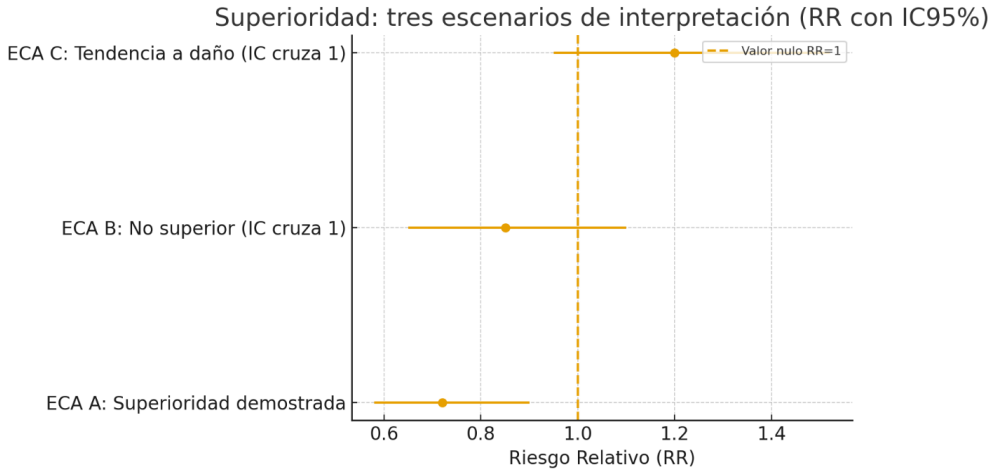
En un artículo, debajo del gráfico suele haber una nota: "Los análisis de subgrupos son exploratorios y deben interpretarse con cautela."

ENSAYO DE SUPERIORIDAD

Un ensayo de superioridad se diseña para demostrar que una intervención es mejor que el tratamiento comparador (placebo o tratamiento estándar). La hipótesis nula es que no hay diferencia entre grupos, y el objetivo del ensayo es rechazarla. Este tipo de ensayo es el más frecuente en medicina, sobre todo cuando se evalúan nuevos fármacos que se espera superen al placebo o a la terapia tradicional, utiliza el RR y suele hacerse por ITT, son los ensayos que cambian guias cuando muestran que una terapia supera a la que ya existe.

Así mismo en su interpretación se pueden encontrar tres casos

Figura 7. Ensayos clínicos de superioridad: tres escenarios de interpretación según riesgo relativo (RR) con intervalo de confianza al 95%



Nota: RR = riesgo relativo; IC95% = intervalo de confianza al 95%.

Su interpretación es de la siguiente forma:

- ECA A: Superioridad demostrada (RR=0,72; IC95% 0,58–0,90): el IC no cruza 1 → hay evidencia de beneficio.
- ECA B: No superior (RR=0,85; IC95% 0,65–1,10): el IC cruza 1 → no se puede afirmar superioridad (resultado "negativo/inconcluso").
- ECA C: Tendencia a daño (RR=1,20; IC95% 0,95–1,52): el punto sugiere perjuicio, pero el IC cruza 1 → estadísticamente no concluyente.

En nuestro ejemplo la pregunta de investigación seria: ¿El iSGLT2 añadido al tratamiento estándar reduce más los eventos cardiovasculares mayores (MACE) que el tratamiento convencional en pacientes con diabetes tipo 2 y alto riesgo cardiovascular?

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

Tabla 9. Ensayo de superioridad: incidencia de eventos cardiovasculares mayores (MACE) a 36 meses en pacientes con diabetes tipo 2 y alto riesgo cardiovascular (N=2000)

Grupo	Eventos / N	Riesgo (%)	
Control	136 / 1000	13,6	
iSGLT2	81 / 1000	8,1	
Medida de efecto	Valor puntual	IC95%	p-valor
Riesgo Relativo (RR)	0,6	0,46 – 0,77	<0,001
Reducción absoluta (RAR)	5,50%	2,8% – 8,2%	<0,001
NNT	18	13 – 36	—

Fuente: Elaboración propia

Su interpretación crítica

- El RR de 0,60 indica que el iSGLT2 reduce en un 40% el riesgo relativo de MACE frente al control.
- El IC95% (0,46–0,77) no cruza el 1, lo que confirma que el beneficio es estadísticamente significativo. Incluso en el peor escenario (RR=0,77), el tratamiento sigue mostrando un beneficio clínico relevante.
- La RAR de 5,5% significa que, por cada 100 pacientes tratados, se previenen entre 3 y 8 eventos cardiovasculares.
- El NNT de 18 traduce el hallazgo a la práctica: se necesita tratar a 18 pacientes como la Sra. Julia durante 3 años para evitar un evento adicional.
- El p < 0,001 respalda la significancia estadística, pero el verdadero valor interpretativo lo aporta el IC, que muestra la magnitud y la precisión del efecto.

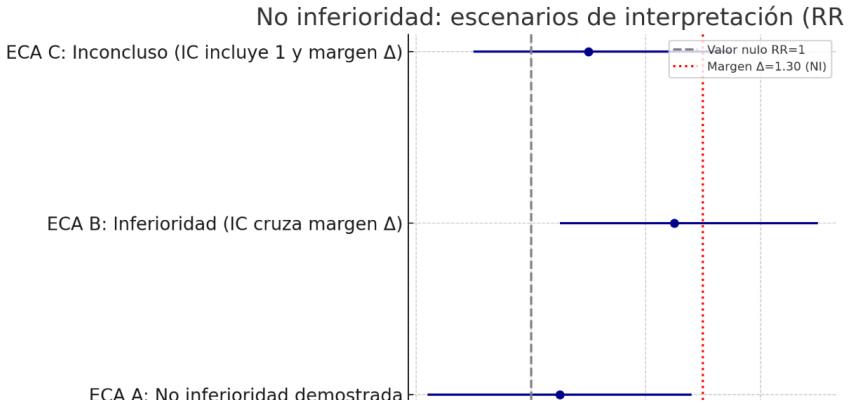
Este ensayo de superioridad demuestra que el iSGLT2 es más eficaz que el tratamiento estándar para prevenir eventos cardiovasculares en pacientes con diabetes tipo 2 y alto riesgo, lo cual lo convierte en una opción sólida para su manejo.

ENSAYO DE NO INFERIORIDAD (NI)

Un ensayo de no inferioridad (NI) busca demostrar que una nueva intervención no es clínicamente peor que el tratamiento estándar, dentro de un margen predefinido que debe estar clínicamente justificada (Δ). Este tipo de ensayo es útil cuando el nuevo tratamiento puede ofrecer ventajas adicionales (menos efectos adversos, menor costo, vía de administración más sencilla, mejor adherencia), aunque no necesariamente supere en eficacia al comparador.

En vez de demostrar "ser mejor", se intenta probar que el tratamiento es "suficientemente bueno", es decir, que no pierde eficacia más allá de lo clínicamente aceptable. Su hipótesis nula nos dice que la intervención es inferior al estándar, su IC no debe cruzar el margen y debe reportarse tanto en ITT como PP

Figura 8. Ensayos clínicos de no inferioridad: escenarios de interpretación según riesgo relativo (RR) con intervalo de confianza al 95% y margen preespecificado (Δ)



Nota: RR = riesgo relativo; IC95% = intervalo de confianza al 95%; Δ = margen de no inferioridad.

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

Para su interpretación:

- ECA A: No inferioridad demostrada (RR=1,05; IC95% 0,82–1,28): el IC está por debajo del margen Δ=1,30 → el nuevo tratamiento no es clínicamente peor.
- ECA B: Inferioridad (RR=1,25; IC95% 1,05–1,50): el IC cruza el margen Δ=1,30 → el tratamiento es peor que el estándar.
- ECA C: Resultado inconcluso (RR=1,10; IC95% 0,90–1,35): el IC incluye tanto el valor nulo (1) como el margen Δ=1,30 → no puede concluirse ni no inferioridad ni superioridad.

En nuestro ejemplo: ¿El iSGLT2 es no inferior a un agonista de GLP-1 en la prevención de eventos cardiovasculares en pacientes con diabetes tipo 2 y alto riesgo cardiovascular?

Se define un margen de no inferioridad Δ = 1,3 para el Riesgo Relativo (RR). Esto significa que se acepta hasta un 30% más de riesgo con iSGLT2 respecto al GLP-1, siempre que ofrezca otras ventajas (por ejemplo, menor costo y mayor facilidad de administración).

Tabla 10. Ensayo de no inferioridad: incidencia de eventos cardiovasculares mayores (MACE) a 36 meses en pacientes tratados con iSGLT2 frente a GLP-1 (N=2000)

Grupo	Eventos / N	Riesgo (%)		
GLP-1	110 / 1000	11		
iSGLT2	115 / 1000	11,5		
Medida de efecto	Valor puntual	IC95%	Margen Δ	Conclusión
Riesgo Relativo (RR)	1,05	0,82 – 1,28	1,3	No inferioridad demostrada

Fuente: Elaboración propia

Para su interpretación crítica:

- El RR de 1,05 indica que la incidencia de MACE fue similar en ambos grupos, con solo un 5% más de riesgo en el grupo iSGLT2.
- El IC95% (0,82–1,28) se mantiene dentro del margen de no inferioridad (Δ=1,30), lo que significa que no podemos descartar una ligera pérdida de eficacia, pero esta se encuentra dentro del rango clínicamente aceptable.
- Como resultado, el iSGLT2 puede considerarse no inferior al GLP-1 para prevención de eventos cardiovasculares.
- En este contexto, la decisión clínica podría inclinarse hacia el iSGLT2 si sus ventajas adicionales (menor costo, facilidad de administración, efectos renales positivos) resultan relevantes para la paciente.
- Si bien los agonistas de GLP-1 son una opción probada, un ensayo de no inferioridad podría justificar el uso de iSGLT2 en su lugar, especialmente si se busca un tratamiento oral, accesible y con beneficios adicionales, sin que ello signifique un sacrificio clínicamente importante en eficacia cardiovascular.

ENSAYO DE EQUIVALENCIA

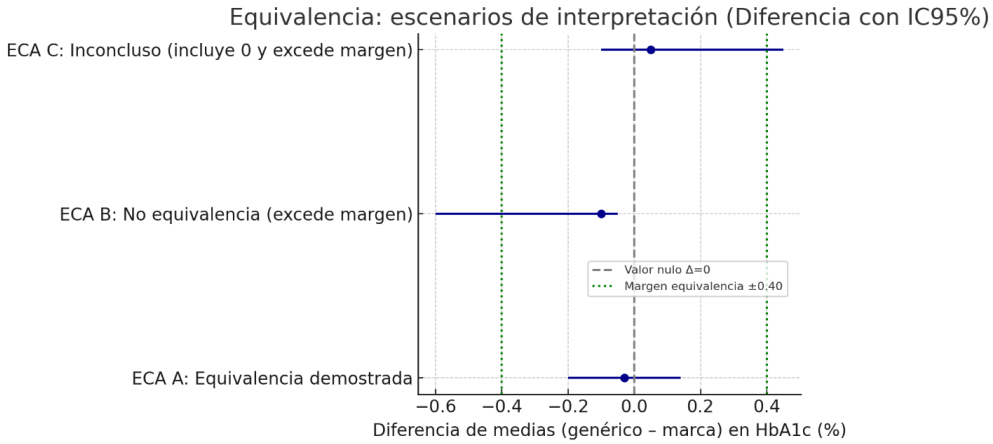
Un ensayo de equivalencia busca demostrar que dos intervenciones producen efectos clínicos iguales dentro de un rango predefinido de tolerancia. A diferencia de los ensayos de no inferioridad, que aceptan que un tratamiento puede ser “no peor” que otro, los ensayos de equivalencia intentan probar que no hay diferencias significativas en ninguna dirección.

Este tipo de diseño se usa con frecuencia para comparar medicamentos genéricos o biosimilares con sus versiones de referencia, con el objetivo de demostrar que ofrecen la misma eficacia y seguridad. Su hipótesis nula nos dice que la diferencia entre ambos medicamentos es mayor al margen en alguna dirección, su margen de equivalencia (±Δ) debe establecerse antes y justificar clínicamente qué diferencia se considera irrelevante, su IC debe encontrarse dentro del margen

En nuestro ejemplo su pregunta de investigación: ¿Un iSGLT2 genérico es equivalente a la versión de marca en la reducción de HbA1c en pacientes con diabetes tipo 2 y alto riesgo cardiovascular? Se define un margen de equivalencia de ±0,4% para la HbA1c. Esto significa que diferencias menores a ±0,4% no se consideran clínicamente relevantes.

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

Figura 9. Ensayos clínicos de equivalencia: escenarios de interpretación según diferencia de medias en HbA1c con intervalo de confianza al 95% y márgenes preespecificados (±Δ)



Nota: IC95% = intervalo de confianza al 95%; Δ = margen de equivalencia; la diferencia de medias se expresa como cambio en HbA1c (%).

Para su interpretación:

- ECA A: Equivalencia demostrada: Diferencia = –0,03% (IC95% –0,20% a +0,14%), el IC queda completamente dentro de ±0,40 → equivalentes
- ECA B: No equivalencia (excede margen): Diferencia = –0,10% (IC95% –0,60% a –0,05%). El IC rebasa el margen inferior (–0,40) → no se puede concluir equivalencia (es más "potente" que el margen permite).
- ECA C: Inconcluso (incluye 0 y excede margen): Diferencia = +0,05% (IC95% –0,10% a +0,45%)., El IC incluye 0 (no hay diferencia clara) pero supera el margen superior (+0,40) → no concluye equivalencia.

Tabla 11. Ensayo de equivalencia: cambio en HbA1c a 36 meses en pacientes tratados con iSGLT2 marca versus genérico

Grupo	Δ HbA1c media (±DE)			
iSGLT2 marca	–0,87 ± 1,00			
iSGLT2 genérico	–0,90 ± 1,05			
Medida de efecto	Diferencia media (genérico – marca)	IC95%	Margen de equivalencia	Conclusión
Cambio en HbA1c	–0,03%	–0,20% a +0,14%	±0,40%	Equivalentes

Fuente: Elaboración propia

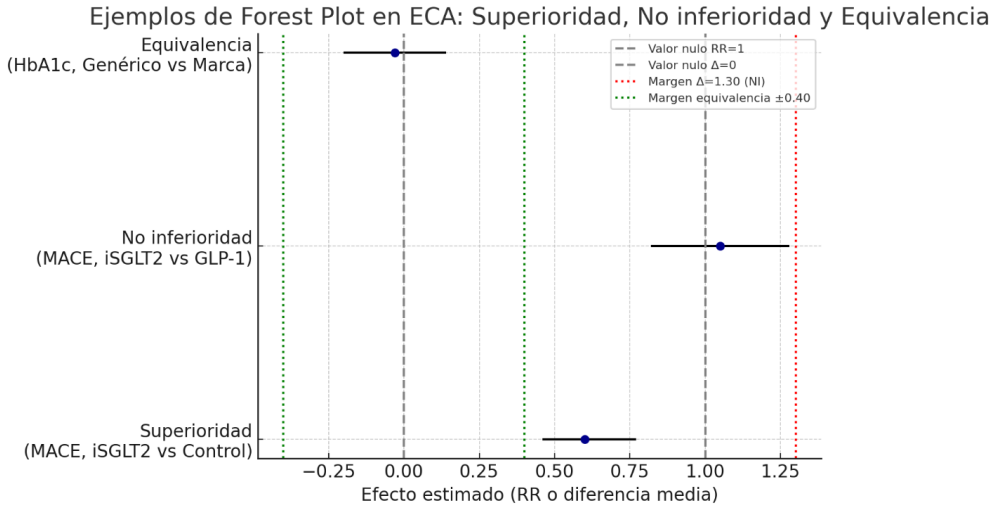
Para su interpretación crítica

- La diferencia media en reducción de ambos en HbA1c fue –0,03%, es decir, prácticamente nula.
- El IC95% (–0,20% a +0,14%) se encuentra completamente dentro del margen de equivalencia (±0,40%), lo que respalda la conclusión de equivalencia.
- Esto significa que ambos tratamientos, marca y genérico, ofrecen la misma eficacia para reducir HbA1c en pacientes con diabetes tipo 2 y riesgo cardiovascular.
- En un ensayo de equivalencia, si el genérico de iSGLT2 muestra eficacia y seguridad similares al de marca, su médico podría recetarle con confianza la versión más accesible, sabiendo que obtendrá el mismo beneficio clínico.

Muchas veces todos estos márgenes se presentan en un Forest plot y queda de la siguiente forma:

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

Figura 10. Ejemplos de forestplot en ensayos clínicos aleatorizados: superioridad, no inferioridad y equivalencia



Nota: RR = riesgo relativo; IC95% = intervalo de confianza al 95%; Δ = margen de no inferioridad; margen de equivalencia = ±0,40%

Para la interpretación del forestplot comparativo con los tres tipos de ensayos clínicos:

- Superioridad (RR=0,60; IC95% 0,46–0,77): el IC no cruza 1 → demuestra que iSGLT2 es superior al control.
- No inferioridad (RR=1,05; IC95% 0,82–1,28): el IC queda dentro del margen Δ=1,30 → concluye no inferioridad frente a GLP-1.
- Equivalencia (Δ HbA1c –0,03; IC95% –0,20 a +0,14): el IC está completamente dentro de ±0,40 → demuestra equivalencia entre genérico y marca

Perlas clínicas y metodológicas para interpretar ECA según su marco

EN ENSAYOS DE SUPERIORIDAD

- Lo que importa es que el IC95% no cruce 1 (para RR) o 0 (para diferencias).
- Si el IC incluye el valor nulo → el ensayo no puede concluir superioridad, aunque el valor puntual sugiera beneficio.
- El p-valor acompaña, pero lo esencial es el IC: muestra tanto la magnitud del efecto como su precisión.
- Pregunta clave para el lector: *¿el efecto es estadísticamente significativo y clínicamente relevante?*

EN ENSAYOS DE NO INFERIORIDAD (NI)

- Aquí lo central es el margen Δ de no inferioridad, que debe estar justificado clínicamente.
- Un IC que incluye 1 no invalida la NI, siempre que se mantenga dentro del Δ.
- Los ensayos de NI deben reportar tanto análisis por intención de tratar (ITT) como por protocolo (PP): si coinciden, la conclusión es robusta.
- Pregunta clave para el lector: *¿el IC se mantiene dentro del margen predefinido de NI? ¿Ese margen es razonable?*

EN ENSAYOS DE EQUIVALENCIA

- Requieren un margen bilateral (±Δ) definido antes de iniciar el ensayo.
- Para concluir equivalencia, el IC95% debe quedar completamente dentro de ambos márgenes.
- Si el IC toca o sobrepasa cualquiera de los límites, no se puede concluir equivalencia, aunque el valor puntual esté cerca de 0.
- Pregunta clave para el lector: *¿el IC cabe íntegramente dentro de los márgenes de equivalencia?*

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

ERRORES FRECUENTES

- Concluir superioridad en un ensayo de NI porque el IC muestra un valor favorable (¡no es lo que estaba diseñado para probar!).
- Confiar solo en el valor p sin revisar el intervalo de confianza ni los márgenes.
- No verificar si el margen Δ fue definido antes y con criterio clínico (un margen amplio puede "maquillar" la inferioridad).
- Olvidar que la relevancia clínica es tan importante como la estadística: un resultado significativo puede no ser relevante en la práctica.

6.4 MEDIDAS DE EFECTO DE VARIABLES CONTINUAS

En muchos ensayos clínicos los desenlaces no se expresan como "evento sí/no", sino como valores numéricos continuos, en estos casos, la medida de efecto más frecuente es la diferencia de medias entre grupos, calculada al final del estudio o como cambio desde el valor basal, son ejemplos: niveles de HbA1c, presión arterial, colesterol LDL, peso corporal, eGFR, etc. Sus características son:

Hipótesis nula (H0): no hay diferencia entre la media de ambos grupos (diferencia = 0).

Hipótesis alternativa (H1): existe diferencia entre las medias (diferencia ≠ 0, o en un sentido específico si el ensayo es unilateral).

Intervalo de confianza (IC95%): si el IC incluye el 0, no hay significancia estadística.

Importancia clínica: incluso cuando hay diferencia significativa, el lector crítico debe preguntarse si esa magnitud de cambio es realmente útil para los pacientes.

DIFERENCIA DE MEDIAS

La diferencia de medias es la medida de efecto más común cuando los desenlaces son variables cuantitativas (p. ej., HbA1c, peso, presión arterial).

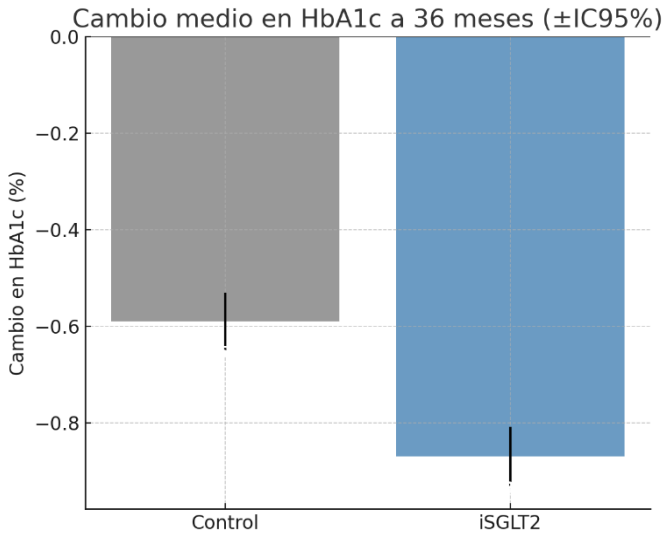
Eso se presenta de la siguiente forma:

Tabla 12. Cambio en HbA1c a 36 meses en pacientes con diabetes tipo 2 y alto riesgo cardiovascular (Control vs. iSGLT2)

Grupo	N	Δ HbA1c media (±DE)	Diferencia media (IC95%)	p-valor
Control	1000	−0,59 ± 0,96		
iSGLT2	1000	−0,87 ± 1,00	−0,28 (−0,39 a −0,17)	<0,001

Fuente: Elaboración propia

Figura 11. Cambio medio en HbA1c (%) a 36 meses en pacientes con diabetes tipo 2 y alto riesgo cardiovascular (Control vs. iSGLT2)



Nota: HbA1c = hemoglobina glicosilada; los valores representan la media de cambio con intervalos de confianza al 95% (IC95%).

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

Para su interpretación crítica

- La diferencia de –0,28% indica que el iSGLT2 logró una reducción de HbA1c ligeramente mayor que el control.
- El IC95% (–0,39 a –0,17) no incluye 0, lo que confirma que la diferencia es estadísticamente significativa.
- Sin embargo, la magnitud del efecto es modesta: una reducción de ~0,3% en HbA1c rara vez cambia la práctica por sí sola.
- El lector crítico debe distinguir entre significancia estadística y relevancia clínica: este beneficio glucémico puede no justificar el uso del fármaco, pero cobra sentido si se combina con sus efectos sobre eventos cardiovasculares y renales.

DIFERENCIA DE MEDIAS AJUSTADAS

En muchos ensayos clínicos, la comparación de medias entre grupos no se hace de manera "cruda" (simple resta de medias), sino ajustada por variables basales u otros factores relevantes. Esto se logra mediante técnicas estadísticas como:

- ANCOVA (Análisis de covarianza): ajusta la diferencia final tomando en cuenta los valores basales de la variable (p. ej., HbA1c inicial).
- Modelos lineales mixtos: permiten incorporar medidas repetidas en el tiempo (longitudinales), ajustando por la evolución de cada paciente.

Esto se realiza porque reduce el efecto de posibles desequilibrios basales (aunque haya aleatorización, siempre puede haber pequeñas diferencias), además que incrementa la precisión de la estimación y permite interpretar la diferencia como "el efecto neto del tratamiento", aislando otras influencias.

Para entenderlo mejor imaginemos que en nuestro ensayo ficticio, al inicio los pacientes del grupo iSGLT2 tenían una HbA1c ligeramente mayor (8,7% vs 8,5%). Si solo comparamos la reducción final (–0,87% vs –0,59%), la diferencia es –0,28%, pero, al usar un modelo ANCOVA que ajusta por HbA1c basal, edad y sexo, la diferencia ajustada puede ser más precisa:

Tabla 13. Cambio en HbA1c a 36 meses: comparación cruda y ajustada en pacientes con diabetes tipo 2 y alto riesgo cardiovascular (Control vs. iSGLT2)

Grupo	Δ HbA1c media (±DE)	Diferencia cruda (IC95%)	Diferencia ajustada (IC95%)*	p-valor
Control	–0,59 ± 0,96			
iSGLT2	–0,87 ± 1,00	–0,28 (–0,39 a –0,17)	–0,25 (–0,36 a –0,14)	<0,001

Nota: HbA1c = hemoglobina glicosilada; DE = desviación estándar; IC95% = intervalo de confianza al 95%.

Para su interpretación crítica:

- El ajuste no cambia radicalmente la conclusión (sigue siendo significativo), pero la estimación es más precisa.
- Para el lector crítico, es clave verificar qué variables se usaron en el ajuste: Si son pertinentes (ej., HbA1c basal, edad), el análisis gana validez, si son excesivas o irrelevantes, el ajuste puede introducir sesgo.
- La diferencia ajustada suele ser la que se reporta en el artículo principal.
- Cuando lea un ECA con desenlaces continuos, se debe revisar siempre si la diferencia reportada es cruda o ajustada. Los ensayos de mayor calidad suelen usar ANCOVA o modelos mixtos para dar resultados más sólidos.

MODELOS MIXTOS PARA MEDIDAS REPETIDAS (MMRM)

En muchos ensayos, los pacientes son evaluados en múltiples momentos de seguimiento (por ejemplo, HbA1c a los 6, 12, 24 y 36 meses). Comparar solo la media al final puede desaprovechar información, además de que algunos pacientes pueden tener datos faltantes en ciertos puntos.

Para manejar esto, se utilizan los MMRM:

- Incorporan todas las mediciones de cada paciente.
- Permiten estimar el efecto del tratamiento a lo largo del tiempo, no solo al final.
- Tienen en cuenta la correlación intrapaciente (las mediciones de un mismo individuo están relacionadas).

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

- Son más robustos frente a datos faltantes (si faltan algunas visitas, aún pueden incluirse las demás del mismo paciente).

Para entenderlo utilizaremos el siguiente ejemplo:

Tabla 14. Cambio en HbA1c a lo largo del estudio (36 meses): análisis crudo y ajustado en pacientes con diabetes tipo 2 y alto riesgo cardiovascular (Control vs. iSGLT2)

Grupo	N	Δ HbA1c media (±DE)	Diferencia cruda (IC95%)	Diferencia ajustada (IC95%)*	p-valor
Control	1000	–0,59 ± 0,96			
iSGLT2	1000	–0,87 ± 1,00	–0,28 (–0,39 a –0,17)	–0,25 (–0,36 a –0,14)	<0,001

Fuente: Elaboración propia con apoyo de Inteligencia artificial

Para su interpretación crítica

- El análisis crudo muestra una diferencia de –0,28% a favor de iSGLT2, estadísticamente significativa (no cuenta el 0 entre los intervalos)
- El análisis ajustado corrige por posibles diferencias iniciales (HbA1c basal ligeramente distinta, edad, sexo), y la estimación final es –0,25%, también significativa.
- Aunque las cifras varían un poco, la conclusión es la misma: el iSGLT2 logra una reducción mayor de HbA1c.

El lector crítico debe fijarse en qué variables fueron incluidas en el ajuste: Si son clínicamente relevantes, aumentan la validez del resultado, si son excesivas o arbitrarias, pueden introducir sesgo.

6.5 ANÁLISIS DE SUPERVIVENCIA

En los ensayos clínicos aleatorizados, cuando el desenlace principal es un evento (infarto, muerte, hospitalización), no basta con conocer qué proporción de pacientes lo presentó al final del estudio. También es necesario saber en qué momento ocurrió y cómo fue la acumulación de eventos en el tiempo.

Para responder estas preguntas, se utilizan técnicas de análisis de supervivencia, los mismos que ya se mencionaron en base a HR, curvas de Kaplan-Meier y regresión de Cox

Para aclarar:

- La curva de Kaplan-Meier muestra cómo se van acumulando los eventos en el tiempo.
- El log-rank test compara estadísticamente ambas curvas. ($p<0,05$ las curvas son diferentes)
- El HR (Cox) resume el efecto del tratamiento durante todo el seguimiento y puede ajustarse por covariables.

6.6 ANÁLISIS DE SUBGRUPOS

En los ensayos clínicos, una pregunta frecuente es si el efecto del tratamiento es igual para todos los pacientes o si varía según ciertas características. Para responder esto, los investigadores realizan análisis de subgrupos, dividiendo a los participantes en categorías como:

- Edad (ej.: <65 vs ≥65 años)
- Sexo (hombres vs mujeres).
- Severidad de la enfermedad al inicio (HbA1c baja vs alta, estadio de insuficiencia cardiaca, etc.).
- Comorbilidades (con vs sin enfermedad renal).

El objetivo es explorar si el efecto del tratamiento es consistente en toda la población del estudio.

Riesgos de sobreinterpretación: Aunque son atractivos, los análisis de subgrupos presentan limitaciones importantes:

- Menor poder estadístico: al dividir la muestra, cada subgrupo tiene menos pacientes y aumenta la probabilidad de resultados falsos.
- Multiplicidad de comparaciones: cuantos más subgrupos se analicen, mayor es la probabilidad de encontrar diferencias "significativas" solo por azar.
- Hallazgos espurios: pueden aparecer diferencias que no tienen base biológica ni clínica. Un ejemplo clásico fue un ensayo de aspirina en el que, en análisis secundarios, parecía que el efecto solo existía en ciertos signos zodiacales: un resultado absurdo, pero "estadísticamente significativo" por azar.
- Influencia en la práctica clínica: algunos lectores cometen el error de recomendar un

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

tratamiento solo a un subgrupo específico cuando el efecto global del ensayo lo respalda para toda la población.

Cómo usarlos con cautela:

- Los análisis de subgrupos deben estar preespecificados en el protocolo del estudio, no inventados después de analizar los datos.
- Lo más importante en un ensayo es siempre el resultado global. Los subgrupos sirven para explorar consistencia, no para generar conclusiones definitivas.
- Tienen valor si: existe un fundamento fisiopatológico o clínico que justifique la diferencia, los resultados son consistentes en la mayoría de los subgrupos, el ensayo fue diseñado con suficiente poder para analizar el subgrupo (lo cual rara vez ocurre).

6.7 ANÁLISIS DE ROBUSTEZ

Cuando un ensayo clínico muestra resultados estadísticamente significativos, el lector crítico no debe conformarse con el valor p ni con el intervalo de confianza. Una pregunta fundamental es si esos resultados son **robustos**, es decir, si se mantienen consistentes cuando se aplican distintos métodos de análisis, escenarios alternativos o ajustes estadísticos.

El análisis de robustez busca responder a esta inquietud. En la práctica, los investigadores suelen contrastar los hallazgos principales con diferentes aproximaciones para demostrar que no dependen de un único modelo o supuesto. Por ejemplo, uno de los contrastes más frecuentes es comparar los resultados obtenidos bajo el principio de intención de tratar (ITT) con los del análisis por protocolo (PP).

- **En el análisis ITT** se incluyen todos los pacientes según el grupo al que fueron aleatorizados, aunque no hayan cumplido el tratamiento o se hayan perdido durante el seguimiento. Esto preserva la fuerza de la aleatorización y es considerado el análisis principal.
- **El análisis PP**, en cambio, restringe la muestra a los participantes que siguieron el tratamiento según lo planificado, lo que en teoría refleja mejor la eficacia del fármaco, pero introduce más riesgos de sesgo. Cuando ambos análisis coinciden en la dirección y magnitud del efecto, el lector puede tener mayor seguridad en que el hallazgo es confiable.

Otra forma habitual de evaluar la robustez es a través de **análisis de sensibilidad**. Estos se utilizan especialmente para explorar el impacto de las pérdidas de seguimiento o de los datos faltantes. El investigador puede probar distintos escenarios, por ejemplo, asumir que todos los pacientes perdidos en el grupo intervención presentaron el evento, o que ninguno lo presentó. Si el resultado principal se mantiene significativo en ambos extremos, puede considerarse más sólido.

También es frecuente repetir los análisis utilizando **modelos estadísticos alternativos**. Un ensayo puede presentar un HR ajustado por edad y sexo, pero luego mostrar que al añadir otras covariables como HbA1c basal, hipertensión o dislipidemia, los resultados se mantienen en la misma dirección. Esta consistencia refuerza la confianza en que el efecto del tratamiento no depende de un único modelo.

En conjunto, el análisis de robustez no busca generar resultados nuevos, sino **poner a prueba la estabilidad de las conclusiones**. Para el lector crítico, representa una herramienta esencial: un ensayo cuyo beneficio se mantiene bajo diferentes enfoques transmite mayor seguridad que aquel cuyos resultados cambian de forma sustancial al modificar las condiciones de análisis.

Para poder ser entendido se presentan los siguientes datos:

Tabla 15. Análisis de robustez para eventos cardiovasculares mayores (MACE) a 36 meses en pacientes con diabetes tipo 2 y alto riesgo cardiovascular

Tipo de análisis	HR (IC95%)	p-valor
Intención de tratar	0,72 (0,58–0,89)	0,002
Por protocolo	0,70 (0,55–0,88)	0,001
Sensibilidad*	0,75 (0,61–0,92)	0,004

Fuente: Elaboración propia

Para su interpretación, la tabla muestra que, independientemente del método de análisis, los resultados favorecen al iSGLT2. El HR se mantiene entre 0,70 y 0,75, lo que indica una reducción robusta del riesgo de eventos cardiovasculares mayores. El análisis de sensibilidad, diseñado para tensionar los datos al máximo, confirma que el efecto no depende de un único escenario, sino

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

que se mantiene clínicamente relevante en diferentes supuestos.

Un excelente lector critico debe conocer los siguientes puntos:

- La fortaleza de un ensayo no se mide solo por el valor p. La robustez implica que los resultados se mantengan estables bajo diferentes formas de análisis.
- ITT primero, PP después. El análisis por ITT es el estándar porque conserva la aleatorización. El PP puede mostrar un efecto más pronunciado, pero si ambos coinciden en dirección y magnitud, la confianza en el hallazgo aumenta.
- Los datos perdidos importan. Un buen ensayo explora escenarios distintos para las pérdidas de seguimiento mediante análisis de sensibilidad. Si el beneficio persiste incluso en el "peor escenario posible", el resultado es sólido.
- Los modelos de ajuste deben ser coherentes. Añadir o quitar covariables relevantes (edad, sexo, comorbilidades) no debería cambiar radicalmente la estimación del efecto. Si ocurre, el lector debe sospechar fragilidad en los resultados.
- La consistencia es la clave. Cuando las conclusiones se mantienen a través de diferentes métodos (RR, HR, ITT, PP, sensibilidad), el hallazgo gana fuerza y se convierte en evidencia más confiable.

6.8 ERRORES FRECUENTES EN LA INTERPRETACIÓN

La lectura crítica de ensayos clínicos no consiste únicamente en identificar resultados estadísticamente significativos. Muchos lectores caen en interpretaciones equivocadas que pueden llevar a decisiones clínicas poco fundamentadas. A continuación, se resumen los errores más comunes y cómo evitarlos.

Confundir RRR con RAR: Uno de los errores más extendidos es quedarse solo con la Reducción Relativa del Riesgo (RRR). Esta medida suele producir cifras llamativas: "el nuevo fármaco reduce el riesgo un 40%". Sin embargo, si el riesgo absoluto es bajo, la ganancia real para el paciente puede ser mínima. La Reducción Absoluta del Riesgo (RAR) y el Número Necesario a Tratar (NNT) traducen el beneficio a un lenguaje clínico comprensible: "por cada 18 pacientes tratados se evita un evento". El lector crítico debe priorizar siempre estas medidas absolutas.

Enfocarse solo en el p-valor: El p-valor responde a la pregunta de si la diferencia observada podría explicarse por azar, pero no informa sobre la magnitud ni sobre la importancia clínica del efecto. Un $p<0,05$ no convierte automáticamente un hallazgo en relevante, ni un $p>0,05$ descarta utilidad clínica. El intervalo de confianza (IC95%) aporta mucha más información, ya que muestra el rango posible de efecto y su precisión.

Ignorar pérdidas de seguimiento: Las pérdidas durante el seguimiento pueden distorsionar gravemente los resultados. Si un porcentaje importante de pacientes se pierde y no se sabe si presentaron el evento, el efecto puede estar sobreestimado o infraestimado. El lector crítico debe revisar siempre si el porcentaje de pérdidas (si es alto, la validez interna está comprometida), cómo se manejaron estadísticamente (análisis por intención de tratar, análisis de sensibilidad).

No diferenciar significancia estadística de relevancia clínica: Un resultado puede ser estadísticamente significativo, pero clínicamente irrelevante. Por ejemplo, una reducción de 1 mmHg en la presión arterial puede alcanzar $p<0,05$ en un ensayo grande, pero no cambia la práctica clínica. A la inversa, un resultado no significativo podría seguir siendo clínicamente valioso si muestra una tendencia consistente en estudios previos o en metaanálisis. El lector debe recordar que significancia estadística ≠ relevancia clínica.

Tabla 16. Errores frecuentes en la interpretación de ensayos clínicos aleatorizados y cómo evitarlos

Error frecuente	Cómo evitarlo
Confundir RRR con RAR	Revisar siempre la Reducción Absoluta del Riesgo (RAR) y el NNT para entender el beneficio real.
Enfocarse solo en el p-valor	Priorizar el intervalo de confianza (IC95%), que muestra magnitud y precisión del efecto.
Ignorar pérdidas de seguimiento	Verificar el porcentaje de pérdidas y cómo fueron manejadas (ITT, sensibilidad).
No diferenciar significancia de relevancia clínica	Preguntar si la magnitud del efecto es clínicamente importante, más allá de la significancia estadística.

Fuente: Elaboración propia con apoyo de Inteligencia artificial

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

7. SESGOS Y LIMITACIONES

Aunque los ensayos clínicos aleatorizados son considerados el "estándar de oro" para evaluar la eficacia de las intervenciones, no están exentos de limitaciones. La lectura crítica exige que el médico pueda reconocer los sesgos más frecuentes y entender cómo afectan la validez de los resultados.

7.1 SESGOS

Sesgo de selección: Ocurre cuando los grupos no son comparables desde el inicio. La aleatorización busca evitarlo, pero si la secuencia de asignación no se genera o no se oculta adecuadamente, los investigadores pueden influir de manera consciente o inconsciente en qué pacientes ingresan a cada brazo.

Sesgo de desempeño: Aparece cuando los pacientes o investigadores saben qué tratamiento reciben y esto modifica su conducta. El cegamiento (simple, doble o triple) reduce este riesgo. Por ejemplo, un paciente que sabe que está en el grupo "activo" puede mejorar su adherencia, mientras que uno en el grupo "placebo" podría desanimarse.

Sesgo de detección: Se produce cuando los evaluadores conocen la asignación de los pacientes y su juicio clínico o interpretación de pruebas se ve influenciada. El cegamiento de quienes evalúan los desenlaces es clave para minimizarlo.

Sesgo de desgaste (attritionbias): Resulta de las pérdidas durante el seguimiento. Si muchos pacientes abandonan y estos no se distribuyen de forma equilibrada entre los grupos, el efecto observado puede estar distorsionado. El análisis por intención de tratar ayuda a mitigar este sesgo.

Sesgo de reporte: Ocurre cuando los investigadores seleccionan qué desenlaces publicar (generalmente los favorables), omitiendo los negativos o neutros. La transparencia y el registro previo del protocolo en plataformas como ClinicalTrials.gov son esenciales para evitarlo.

7.2 LIMITACIONES

Generalización restringida: los pacientes incluidos suelen ser seleccionados con criterios estrictos. Por ello, los resultados pueden no reflejar lo que ocurre en la práctica clínica real con poblaciones más heterogéneas.

Duración insuficiente: algunos ensayos no siguen a los pacientes el tiempo suficiente para detectar desenlaces relevantes (ejemplo: mortalidad a largo plazo).

Uso de desenlaces sustitutos: en lugar de medir directamente la mortalidad o la calidad de vida, se utilizan variables intermedias (p. ej., niveles de laboratorio), que no siempre predicen de manera confiable el beneficio clínico.

Conflictos de interés: los estudios financiados por la industria pueden tener sesgos de diseño, análisis o reporte a favor de la intervención.

Costo y complejidad: los ECA requieren grandes recursos, lo que limita su viabilidad para todas las preguntas de investigación.

El lector crítico debe recordar que ningún ensayo es perfecto. La clave está en reconocer qué sesgos pudieron intervenir, cómo los autores intentaron controlarlos y qué impacto real tienen en la validez interna y en la aplicabilidad externa de los resultados; todo esto permitirá tener un análisis crítico para decidir si es aplicable o no a la práctica medica que se enfrenta.

8. CHECKLIST

Este checklist está pensado para que, al leer un artículo, se pueda marcar si cada componente está adecuadamente reportado (Sí/No/Parcial). Se basa en las guías CONSORT 2010 (25 ítems) y su explicación–elaboración; lo complementa con SPIRIT 2013 (para cotejar el protocolo), ICH-E9(R1) (estimands y sensibilidad) y requisitos de registro OMS/ICTRP (transparencia).

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

Tabla 17. Checklist internacional para lectura crítica de ECA

Sección	Subsección	Criterios / Contenido (sin modificar el texto)
A. Título y resumen	—	• Identifica explícitamente el estudio como ensayo aleatorizado.
		• Resumen estructurado: diseño, población, intervención, comparador, desenlaces, resultados principales con medidas de efecto e intervalos de confianza, y conclusiones.
B. Introducción	—	• Fundamento: justificación científica clara.
		• Objetivos/hipótesis: precisos y con el marco declarado (superioridad, no inferioridad, equivalencia).
C. Métodos	1) Diseño del ensayo	• Tipo: paralelo, factorial, crossover, por conglomerados (clúster), adaptativo, pragmático, multi-brazo; razón de asignación.
		• Si es no inferioridad/equivalencia: margen (Δ) definido y justificado clínicamente, estrategia analítica preespecificada (IC una/dos colas, poblaciones ITT/PP).
		• Si es adaptativo: reglas de adaptación, control del error tipo I, análisis interinos y plan estadístico previo.
		• Si es factorial: definición de efectos principales e interacciones y cómo se analizarán.
		• Si es crossover: periodos, efecto de arrastre y periodo de lavado.
		• Si es clúster: número y tamaño de clústeres, ICC en cálculo muestral y análisis.
	2) Participantes y entorno	• Criterios de inclusión/exclusión claros; escenarios (hospital/APS/país/centros); período de reclutamiento.
	3) Intervenciones	• Detalle suficiente para replicar (dosis, pauta, vía, duración); cointervenciones permitidas/prohibidas; adherencia. Idealmente, seguir una guía de descripción estandarizada (p. ej., TIDieR).
	4) Desenlaces	• Primarios y secundarios preespecificados (definición operativa, momento de medición, dirección clínica del efecto).
		• Si hay sustitutos: declarar validez y relevancia.
	5) Tamaño muestral y monitoreo	• Supuestos usados (tamaño de efecto/varianza/tasa de eventos/Δ si aplica), pérdidas esperadas, planes de interinos y DSMB (comité de monitoreo de datos).
	6) Aleatorización y ocultamiento	• Generación de secuencia (tipo y restricciones).
		• Ocultamiento de la asignación (mecanismo).
		• Responsabilidades: quién generó/asignó/inscribió a los participantes.
	7) Cegamiento	• Quién estuvo ciego (pacientes, personal, evaluadores) y cómo se garantizó; similitud de comparadores.
	8) Plan estadístico (SAP) y estimand	• Métodos para el desenlace primario y secundarios; poblaciones de análisis (ITT/PP).
		• Manejo de datos faltantes y análisis de sensibilidad; control de multiplicidad; subgrupos preespecificados.
		• Estimand definido: población objetivo, variable, eventos intercurrentes y medida-resumen; análisis principal y sensibilidades alineadas con ese estimand.
	9) Ética, registro y protocolo	• Aprobación ética; registro prospectivo con número visible; acceso a protocolo/versión y a enmiendas.
D. Resultados	—	• Flujograma (diagrama CONSORT): aleatorizados, tratados, analizados, pérdidas/exclusiones (con causa).
		• Cronograma de reclutamiento/seguimiento; razones de detención si hubo.
		• Características basales por grupo (tabla 1) con comparabilidad.
		• N analizado por desenlace y población (ITT/PP).
		• Resultados del desenlace primario con medidas de efecto (absolutas y relativas) y IC; claridad del horizonte temporal.
		• Para no inferioridad/equivalencia: figura con IC y margen Δ (límites visibles).
		• Análisis adicionales: subgrupos/ajustes; distinguir prespecificados de exploratorios.
		• Daños/seguridad: definiciones, gravedad/seriedad, denominadores por grupo, poblaciones de seguridad y método de captura de eventos.
E. Discusión		• Limitaciones: sesgos potenciales, imprecisión, multiplicidad; generalizabilidad (aplicabilidad).
		• Balance beneficios–daños y contexto de evidencia (comparación con estudios previos y guías).

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

Sección	Subsección	Criterios / Contenido (sin modificar el texto)
F. Otra información		· Financiación y rol del patrocinador (diseño, análisis, escritura, publicación).
		· Conflictos de interés y acceso a datos/código si corresponde.
Extensiones útiles para diseños especiales (marcar cuando apliquen)		· No inferioridad/Equivalencia: margen Δ, justificación clínica, coherencia ITT/PP, gráfica con Δ; interpretación centrada en márgenes.
		· Clúster: ICC y su IC, número/tamaño de clústeres, flujo de clústeres e individuos, ajuste por clustering.
		· Pragmáticos: flexibilidad de intervención y comparador, cointervenciones, adherencia, representatividad del entorno.
		· Adaptativos: reglas de adaptación, simulaciones, control de error, enmascaramiento operacional, independencia del DSMB.
		· Multi-brazo: control de multiplicidad y jerarquía de comparaciones.
		· Factoriales: definición y análisis de interacción.
		· Crossover: periodos, efectos de secuencia/arrastre, periodo de lavado.
		· Sustitutos: validez del marcador proxy y coherencia con desenlaces duros.
		· Resultados reportados por pacientes (PRO/QoL): instrumentos validados, mínimos cambios clínicamente importantes.
Doble verificación con protocolo y registro		Antes de cerrar la lectura crítica:
		1. Contrasta el artículo con su protocolo (objetivos, desenlaces, SAP, manejo de intercurrentes, adherencia).
		2. Verifica el registro prospectivo (fechas, población, desenlaces declarados) y su consistencia con lo publicado.
		3. Comprueba que el estimand esté definido y que existan análisis de sensibilidad coherentes.

Fuente: Elaboración propia

9. UTILIDAD DE LOS ECA

Los ECA son considerados el estándar de oro para evaluar la eficacia y seguridad de las intervenciones en salud. Su utilidad radica en que, cuando están bien diseñados y ejecutados, ofrecen la mejor evidencia posible para guiar la práctica clínica. Sin embargo, también tienen limitaciones que el lector crítico debe conocer para aplicarlos con criterio.

9.1 UTILIDAD

Las aportaciones principales son:

- **Causalidad confiable**: Gracias a la aleatorización, los ECA reducen el sesgo de selección y equilibran variables conocidas y desconocidas entre grupos. Esto permite atribuir con mayor certeza las diferencias de resultados a la intervención estudiada.
- **Medidas directas de eficacia y seguridad**: Un ECA bien diseñado permite estimar el beneficio absoluto y relativo de un tratamiento, así como los riesgos de eventos adversos. Esto se traduce en indicadores clínicos como el NNT y el NNH, que ayudan a decidir si una intervención merece ser utilizada en la práctica.
- **Base para guías clínicas**: La mayoría de las recomendaciones terapéuticas de sociedades científicas y guías internacionales se apoyan en ECA de alta calidad. Sin ellos, la medicina se basaría en observaciones menos confiables o en experiencia personal.
- **Protección del paciente**: Los ECA incluyen mecanismos éticos (consentimiento informado, comités de ética, monitoreo de seguridad) que buscan equilibrar la innovación con la seguridad del participante.

9.2 LIMITACIONES A CONSIDERAR

A pesar de su relevancia, los ECA no son perfectos:

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

- Suelen incluir poblaciones muy seleccionadas, lo que limita la generalización a la práctica diaria.
- Tienen costos elevados y duración prolongada, por lo que muchas preguntas clínicas importantes quedan sin respuesta.
- Algunos ensayos priorizan desenlaces sustitutos (biomarcadores, cifras de laboratorio) en lugar de desenlaces clínicamente relevantes como mortalidad o calidad de vida.
- Existe el riesgo de conflictos de interés cuando son financiados por la industria.

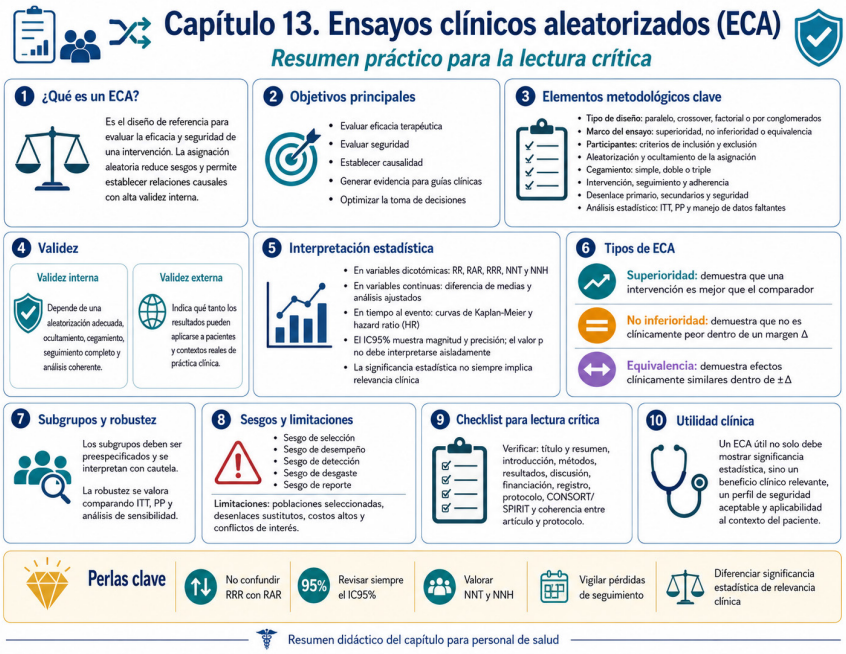
9.3 APLICABILIDAD EN LA PRÁCTICA CLÍNICA

El lector crítico debe recordar que la verdadera utilidad de un ECA no se mide solo en términos de significancia estadística, sino en su capacidad de responder:

- ¿Este tratamiento ofrece un beneficio clínico relevante para mis pacientes?
- ¿El riesgo de efectos adversos es aceptable?
- ¿El contexto de mi práctica permite aplicar lo estudiado (disponibilidad, costos, adherencia, sistema de salud)?

10. UTILIDAD

La aplicación de los hallazgos del ensayo clínico en la Sra. Julia trasciende el control de la hemoglobina glicosilada. La utilidad clínica de los iSGLT2 en este escenario se traduce en una nefroprotección activa y una reducción significativa del riesgo cardiovascular. Al trasladar la evidencia a la práctica, el médico no solo trata una cifra laboratorial, sino que interviene en el pronóstico vital de la paciente, disminuyendo la probabilidad de progresión a enfermedad renal crónica terminal y hospitalizaciones por insuficiencia cardíaca. Esta sección permite al clínico decidir, con base en la magnitud del beneficio (RRA y NNT), si la intervención justifica el costo y los posibles efectos adversos en su contexto local.



11. PREGUNTAS DE AUTOEVALUACIÓN

PREGUNTAS DE SELECCIÓN MÚLTIPLE

1. ¿Cuál es la principal ventaja de la aleatorización en un ECA? a) Reducir el tamaño muestral necesario b) Equilibrar variables conocidas y desconocidas entre grupos c) Eliminar la necesidad de cegamiento d) Evitar el reporte de pérdidas

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

2. En un ensayo de no inferioridad, ¿qué condición debe cumplirse para concluir no inferioridad? a) Que el IC no cruce el valor nulo (1 o 0) b) Que el IC se mantenga por debajo del margen Δ preespecificado c) Que el valor p sea <0,05 d) Que las pérdidas de seguimiento sean menores al 5%

3. ¿Qué medida expresa el número de pacientes que deben ser tratados para evitar un evento adverso? a) RRR b) RAR c) NNT d) NNH

4. En un análisis de supervivencia, ¿qué representa el Hazard Ratio (HR)? a) El riesgo acumulado al final del estudio b) El riesgo instantáneo de evento en cualquier momento del seguimiento c) El porcentaje de pacientes que completaron el protocolo d) El número necesario para causar un evento adverso

5. ¿Cuál de los siguientes es un sesgo que puede ocurrir en los ECA? a) Sesgo de selección b) Sesgo de desempeño c) Sesgo de detección d) Todas las anteriores

PREGUNTAS ABIERTAS

6. Explique con sus propias palabras la diferencia entre un ensayo de superioridad y uno de no inferioridad.

7. ¿Por qué se considera el análisis por intención de tratar (ITT) como el estándar en los ECA?

8. ¿Cómo se interpreta clínicamente el NNT y qué limitaciones tiene como medida de efecto?

9. ¿Qué información aporta una curva de Kaplan-Meier que no puede obtenerse a partir de un RR acumulado al final del seguimiento?

10. Mencione tres errores frecuentes en la interpretación de un ECA y explique cómo evitarlos en la práctica clínica.

RESPUESTAS:

1. b) Equilibrar variables conocidas y desconocidas entre grupos. La aleatorización distribuye de forma aleatoria tanto los factores medibles como los no medibles entre los grupos, reduciendo el sesgo de selección y permitiendo que las diferencias en los resultados se atribuyan con mayor validez a la intervención.

2. b) Que el IC se mantenga por debajo del margen Δ preespecificado. En los ensayos de no inferioridad, el objetivo no es demostrar diferencia, sino que la intervención no sea peor que el comparador más allá de un margen clínicamente aceptable (Δ). Para ello, todo el intervalo de confianza debe permanecer dentro de ese margen.

3. c) NNT. El número necesario a tratar (NNT) indica cuántos pacientes deben recibir una intervención para prevenir un evento adicional. Es una medida directa de aplicabilidad clínica, derivada de la reducción absoluta del riesgo.

4. b) El riesgo instantáneo de evento en cualquier momento del seguimiento. El hazard ratio (HR) representa la razón de tasas de ocurrencia del evento a lo largo del tiempo, considerando el momento en que ocurren los eventos, no solo su frecuencia acumulada.

5. d) Todas las anteriores. Los ensayos clínicos pueden presentar diferentes tipos de sesgo: de selección (si la aleatorización falla), de desempeño (si no hay cegamiento adecuado) y de detección (si la evaluación del desenlace no es objetiva).

6. Diferencia entre ensayo de superioridad y no inferioridad. Un ensayo de superioridad busca demostrar que una intervención es mejor que otra en términos de un desenlace clínico. En cambio, un ensayo de no inferioridad pretende demostrar que una intervención no es significativamente peor que el tratamiento estándar dentro de un margen clínicamente aceptable. Mientras el primero busca diferencia, el segundo busca equivalencia práctica en eficacia.

7. Importancia del análisis por intención de tratar (ITT). El análisis por intención de tratar incluye a todos los pacientes según el grupo al que fueron originalmente asignados, independientemente de si completaron o no la intervención. Esto preserva los beneficios de la aleatorización, evita sesgos y refleja mejor la efectividad en condiciones reales de práctica clínica.

8. Interpretación clínica del NNT y sus limitaciones. El NNT indica cuántos pacientes deben tratarse para prevenir un evento adicional, lo que permite dimensionar el beneficio de una intervención en términos prácticos. Sin embargo, tiene limitaciones: depende del riesgo basal de la población, puede variar según el tiempo de seguimiento y no refleja directamente los posibles efectos adversos.

9. Aporte de la curva de Kaplan-Meier. La curva de Kaplan-Meier permite visualizar la

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

probabilidad de permanecer libre de un evento a lo largo del tiempo, mostrando cuándo ocurren los eventos y cómo se separan los grupos durante el seguimiento. Esto no puede obtenerse únicamente con un riesgo relativo acumulado al final del estudio.

10. Errores frecuentes en la interpretación de un ECA y cómo evitarlos. Un error frecuente es confundir la reducción relativa del riesgo con la absoluta, lo que puede sobreestimar el beneficio; esto se evita revisando siempre el RAR y el NNT. Otro error es enfocarse únicamente en el valor p, ignorando el intervalo de confianza, por lo que se debe priorizar la magnitud y precisión del efecto. Finalmente, ignorar las pérdidas de seguimiento puede sesgar los resultados, por lo que es fundamental evaluar su magnitud y cómo fueron manejadas (por ejemplo, mediante análisis ITT o de sensibilidad).

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

REFERENCIAS

CONSORT Group. CONSORT 2025 statement: Updated guideline for reporting randomised trials. BMJ. 2025. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC11996237/.
Chan A-W, Mayo-Wilson E, Li T, et al. SPIRIT 2025 statement: Harmonized protocol guidance for randomized trials. JAMA. 2025. Availableat: https://doi.org/10.1001/jama.2025.8411.
Butcher NJ, Monsour A, Mew EJ, et al. Guidelines for reporting outcomes in trial reports: The CONSORT-Outcomes 2022 extension. JAMA. 2022;328(22):2252–64. Availablefrom: https://jamanetwork.com/journals/jama/fullarticle/2799401/.
Junqueira DR, Zorzela L, Golder S, et al. CONSORT Harms 2022 statement, explanation, and elaboration: Updated guideline for the reporting of harms in randomized trials. BMJ. 2023;381:e073725. Available at: https://doi.org/10.1136/bmj-2023-073725.
Kahan BC, Elbourne D, Hall SS, et al. Reporting of factorial randomized trials: Extension of the CONSORT 2010 statement. BMJ. 2025;388:e081123. Available at: https://www.equator-network.org/reporting-guidelines/reporting-of-factorial-randomized-trials-extension-of-the-consort-2010-statement/.
Sterne JAC, Savović J, Page MJ, et al. RoB 2: A revised tool for assessing risk of bias in randomised trials. BMJ. 2019;366:l4898. Availableat: https://doi.org/10.1136/bmj.l4898.
Higgins JPT, Thomas J, Chandler J, et al., editors. Cochrane handbook for systematic reviews of interventions. Version 6.5 (updated 2024). Chapter 8, Assessing risk of bias in a randomized trial (RoB 2). Available at: https://www.cochrane.org/es/authors/handbooks-and-manuals/handbook/current/chapter-08.
International Council for Harmonisation (ICH). E9(R1) addendum on estimands and sensitivity analysis in clinical trials (Step 5). 2020. Available from: https://www.ema.europa.eu/en/ich-e9-statistical-principles-clinical-trials-scientific-guideline.
Mauri L, D'Agostino RB Sr. Challenges in the design and interpretation of noninferiority trials. N Engl J Med. 2017;377(14):1357–67. Available at: https://doi.org/10.1056/NEJMra1510063.
Schandelmaier S, Briel M, Varadhan R, et al. Development of the instrument to assess the credibility of effect modification analyses (ICEMAN) in randomized controlled trials and meta-analyses. CMAJ. 2020;192(32):E901–6. Availableat: https://doi.org/10.1503/cmaj.200077.
Kahan BC, Hindley J, Edwards M, et al. The estimands framework: A primer on the ICH E9(R1) addendum. BMJ. 2024;384:e076316. Availableat: https://doi.org/10.1136/bmj-2023-076316.
Dumas E, Stensrud MJ. How hazard ratios can mislead and why it matters in practice. Eur J Epidemiol. 2025;40(6):603–9. Availableat: https://doi.org/10.1007/s10654-025-01250-9.

https://doi.org/10.62486/978-9915-9928-3-9.Ch14

Chapter 15 / Capítulo 15

Critical reading of scientific articles on health: A practical guide (Spanish Edition)
ISBN: 978-9915-9928-3-9
DOI: 10.62486/978-9915-9928-3-9.Ch15
Pages: 243-262

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Quasi-Experimental Studies

Estudios cuasiexperimentales

Jhossmar Cristians Auza-Santivañez
Pamela Gutierrez Villegas
Aaron Eduardo Carvajal-Tapia

1 Hospital del Gran Chaco "Fray Quebracho" de III nivel. Instituto Académico Científico Quispe-Cornejo. Tarija, Bolivia
2 Hospital Obrero Número 30. La Paz, Bolivia.
3 Hospital de la Mujer "Dr. Percy Boland Rodríguez", Unidad de Cuidados Intensivos Neonatales. Santa Cruz, Bolivia.

ABSTRACT

Introduction. Quasi-experimental studies constitute an essential methodological alternative for evaluating clinical interventions when randomization is not feasible for ethical, logistical, or administrative reasons. Their application allows for the generation of evidence in real-world healthcare settings, although with greater susceptibility to selection bias and confounding. Development. This chapter describes the main quasi-experimental designs, their strengths, limitations, and strategies for controlling threats to internal validity, including multivariable adjustments, propensity score matching, and sensitivity analysis. It also addresses the critical interpretation of results, the evaluation of methodological quality, and the use of international tools for assessing the risk of bias and clinical applicability. Conclusions. Quasi-experimental studies represent a valuable source of evidence for decision-making when randomized controlled trials are not feasible. Their proper interpretation requires a critical analysis of the design, potential biases, and adjustment methods employed, promoting a reflective and contextualized application of their findings in clinical practice and public health.

Keywords: Quasi-experimental Studies

RESUMEN

Introducción. Los estudios cuasiexperimentales constituyen una alternativa metodológica esencial para evaluar intervenciones clínicas cuando la aleatorización no es factible por razones éticas, logísticas o administrativas. Su aplicación permite generar evidencia en escenarios asistenciales reales, aunque con mayor susceptibilidad a sesgos de selección y confusión. Desarrollo. El capítulo describe los principales diseños cuasiexperimentales, sus fortalezas, limitaciones y estrategias para controlar amenazas a la validez interna, incluyendo ajustes multivariables, puntuación de propensión y análisis de sensibilidad. Asimismo, aborda la interpretación crítica de los resultados, la evaluación de la calidad metodológica y el uso de herramientas internacionales para valorar el riesgo de sesgo y la aplicabilidad clínica. Conclusiones: Los estudios cuasiexperimentales representan una fuente valiosa de evidencia para la toma de decisiones cuando los ensayos clínicos aleatorizados no son viables. Su adecuada interpretación exige un análisis crítico del diseño, los sesgos potenciales y los métodos de ajuste empleados, favoreciendo una aplicación reflexiva y contextualizada de sus hallazgos en la práctica clínica y la salud pública.

Palabras clave. Estudios cuasiexperimentales

1. CASO CLÍNICO

En un Hospital de tercer nivel ubicado a 3600 metros sobre el nivel del mar, se implementó un Programa Intensivo de Rehabilitación Neuromotora (PIRN) dirigido a pacientes con accidente cerebrovascular (ACV) isquémico reciente, como complemento a la estrategia convencional de rehabilitación. Debido a restricciones operativas relacionadas con la disponibilidad de recursos humanos especializados y capacidad física instalada, la asignación de los pacientes a cada modalidad terapéutica no se realiza mediante un proceso de aleatorización, sino en función de la disponibilidad diaria de cupos y del orden de ingreso al servicio. Esta circunstancia imposibilita la realización de un ensayo clínico aleatorizado, pero configura un escenario idóneo para la aplicación de un diseño cuasiexperimental con grupos no equivalentes, representativo de las condiciones habituales de la práctica clínica. Los participantes son incorporados al programa durante las dos primeras semanas posteriores al evento cerebrovascular y reciben protocolos terapéuticos estandarizados. El grupo de intervención es sometido al PIRN, consistente en tres horas diarias de terapia intensiva durante un período de ocho semanas, mientras que el grupo comparador recibe rehabilitación convencional con una duración de una hora diaria. Ambos grupos continúan recibiendo el mismo tratamiento médico estándar, de acuerdo con las recomendaciones clínicas vigentes para el manejo del ACV isquémico.

Los desenlaces de interés incluyen variables clínicas y funcionales relevantes. Entre los desenlaces principales se encuentran la recuperación motora, evaluada mediante la escala de Fugl-Meyer, y la independencia funcional, medida a través del Índice de Barthel. Como desenlaces secundarios se consideran el grado global de discapacidad, valorado mediante la Escala Modificada de Rankin (mRS), las pruebas de movilidad funcional y la frecuencia de reingresos hospitalarios. Estas variables son evaluadas de manera sistemática en tres momentos predefinidos: al inicio del estudio, a los tres meses y a los seis meses de seguimiento. Aunque la ausencia de asignación aleatoria incrementa el riesgo de sesgo de selección y de confusión, los investigadores pueden minimizar estas amenazas mediante la comparación detallada de las características basales de los grupos y la aplicación de métodos estadísticos de ajuste, tales como modelos multivariables o técnicas basadas en puntuaciones de propensión. Estas estrategias permiten obtener estimaciones más precisas y válidas del efecto de la intervención. Este escenario ilustra cómo los estudios cuasiexperimentales pueden generar evidencia clínicamente relevante y metodológicamente sólida cuando las limitaciones éticas, logísticas o administrativas impiden la realización de un ensayo clínico aleatorizado, aportando información útil para la toma de decisiones en contextos asistenciales reales.

2. OBJETIVOS

Evaluar la efectividad y seguridad de una intervención en escenarios clínicos reales, cuando la aleatorización no es posible o no es ética. A diferencia de los ensayos clínicos controlados, estos diseños permiten generar evidencia en condiciones que reflejan mejor la práctica cotidiana.

De manera más específica, los estudios cuasiexperimentales buscan:

- Estimar el efecto de la intervención sobre desenlaces clínicos relevantes, ajustando las diferencias iniciales entre los grupos para reducir sesgos.
- Comparar la evolución de los pacientes en el tiempo, identificando si los cambios observados pueden atribuirse a la intervención.
- Explorar la variabilidad de los resultados en subgrupos clave, como edad, severidad de la enfermedad o momento de inicio del tratamiento.
- Describir la seguridad y la adherencia, registrando eventos adversos y el grado en que los pacientes completan la intervención.
- Generar evidencia preliminar que sirva como base para ensayos clínicos aleatorizados futuros o para orientar decisiones en políticas de salud.

En conjunto, estos objetivos muestran cómo los ensayos cuasiexperimentales, aunque más vulnerables a sesgos que los aleatorizados, son herramientas valiosas para responder preguntas relevantes en contextos reales de práctica clínica y salud pública.

3. INTRODUCCIÓN

Los estudios cuasiexperimentales constituyen una alternativa metodológica fundamental dentro de la investigación clínica y la evaluación de intervenciones sanitarias. Estos diseños

https://doi.org/10.62486/978-9915-9928-3-9.Ch15

ocupan una posición intermedia entre los estudios observacionales y los ensayos clínicos aleatorizados, ya que incorporan una intervención activa por parte del investigador, pero carecen de asignación aleatoria de los participantes a los grupos de estudio (1,2). Aunque presentan una menor capacidad para controlar factores de confusión en comparación con los ensayos clínicos aleatorizados, ofrecen importantes ventajas cuando la aleatorización no es posible, no resulta ética o es inviable debido a limitaciones logísticas, administrativas o de recursos (1,3). En la práctica clínica, la salud pública y la gestión de servicios sanitarios son frecuentes los escenarios en los que deben adoptarse decisiones basadas en evidencia sin disponer de ensayos clínicos aleatorizados. La evaluación de programas de rehabilitación, estrategias de vacunación, intervenciones educativas, modificaciones organizacionales o políticas de control del tabaquismo constituye ejemplos representativos de situaciones en las que los estudios cuasiexperimentales permiten valorar el impacto de una intervención en condiciones reales de implementación (3,4). En este contexto, estos diseños proporcionan evidencia relevante para orientar la toma de decisiones clínicas, epidemiológicas y de salud pública.La principal fortaleza de los estudios cuasiexperimentales radica en la combinación de dos características esenciales: la introducción deliberada de una intervención sobre una población determinada y la ausencia de asignación aleatoria de los participantes a los grupos de comparación (1,2). Esta última característica la diferencia de los ensayos clínicos tradicionales y constituye la principal fuente de vulnerabilidad metodológica, debido al incremento del riesgo de sesgo de selección, confusión y otras amenazas a la validez interna (5,6).Sin embargo, precisamente por desarrollarse en entornos asistenciales reales y bajo las condiciones habituales de atención sanitaria, estos estudios suelen presentar una elevada validez externa y una mayor aplicabilidad de sus resultados. Por tanto, permiten evaluar la efectividad de intervenciones complejas en poblaciones heterogéneas y en escenarios que reflejan de manera más fiel la realidad cotidiana de los sistemas de salud (7,8).

Desde la perspectiva de la Medicina Basada en la Evidencia, la adecuada interpretación de los estudios cuasiexperimentales exige una evaluación crítica rigurosa de su diseño, ejecución, análisis e interpretación. El lector debe ser capaz de identificar posibles fuentes de sesgo, valorar las estrategias empleadas para controlar factores de confusión y juzgar la validez, precisión y relevancia clínica de los resultados obtenidos (1,5).El objetivo de estecapítulo es proporcionar al lector los fundamentos conceptuales y metodológicos necesarios para reconocer, analizar e interpretar críticamente los estudios cuasiexperimentales. De esta manera, se pretende desarrollar competencias que permitan evaluar la calidad de la evidencia generada por estos diseños, comprender sus fortalezas y limitaciones, y determinar la aplicabilidad de sus hallazgos en la práctica clínica, la salud pública y la toma de decisiones sanitarias.

4. METODOLOGÍA

Los estudios cuasiexperimentales se caracterizan porque el investigador interviene activamente (introduce un tratamiento, programa o política), pero la asignación de los participantes a los grupos no es aleatoria. Esta característica la diferencia de los ensayos clínicos aleatorizados y explica por qué son más vulnerables a sesgos, en especial el de selección.

Existen diferentes diseños cuasiexperimentales, cada uno con utilidades y limitaciones particulares.

4.1 ESTUDIO CON GRUPOS NO EQUIVALENTES

En este diseño se comparan dos o más grupos de pacientes, pero la asignación no es aleatoria, sino que depende de factores externos. Eso significa que los grupos pueden empezar siendo diferentes en aspectos que influyen directamente en los resultados, lo que se conoce como sesgo de selección.Ejemplo aplicado al caso clínico: algunos pacientes ingresan al Programa Intensivo de Rehabilitación Neuromotora (PIRN) y otros a la rehabilitación convencional, no por decisión clínica o por azar, sino por la disponibilidad de cupos. Esto implica que, al inicio, un grupo puede incluir más pacientes jóvenes, con menor gravedad del ACV o con más apoyo familiar, mientras que el otro concentre pacientes más graves o con más comorbilidades.

Si estas diferencias no se reconocen ni se corrigen, el estudio puede atribuir al programa intensivo un beneficio que en realidad se debe a que sus pacientes estaban “mejor de entrada”.

Las claves para la lectura crítica son:

- Verificar si los autores describen con detalle las características basales de ambos grupos (edad,

https://doi.org/10.62486/978-9915-9928-3-9.Ch15

sexo, gravedad del ACV, comorbilidades, función previa).

- Evaluar si se usaron ajustes estadísticos (como modelos de regresión o técnicas de emparejamiento por puntuación de propensión) para equilibrar las diferencias iniciales.
- Preguntarse si, aun con los ajustes, pueden existir variables no medidas (p. ej., motivación del paciente, calidad del apoyo familiar) que sigan influyendo en los resultados.

En síntesis, este diseño puede ofrecer información valiosa, pero la validez de sus conclusiones depende de la transparencia en la comparación inicial y de la solidez de los métodos utilizados para reducir las diferencias entre los grupos.

4.2 ESTUDIOS ANTES Y DESPUÉS (PRE-POST)

En este diseño se observa a un mismo grupo de pacientes en dos momentos: antes de recibir la intervención y después de completarla. El análisis se centra en identificar si hubo cambios en los desenlaces tras la exposición al programa o tratamiento.Ejemplo aplicado al caso clínico: los pacientes que ingresan al Programa Intensivo de Rehabilitación Neuromotora (PIRN) se evalúan al inicio y luego nuevamente a los seis meses. Si la mayoríamuestra mejoría en la independencia funcional, podría pensarse que el programa es efectivo.

Sin embargo, la principal debilidad de este diseño es que no cuenta con un grupo comparador. La mejoría observada podría deberse no solo al PIRN, sino también a otros factores: la recuperación natural del cerebro tras un ACV, el tratamiento médico estándar, el apoyo familiar o incluso cambios espontáneos en la motivación de los pacientes. Por lo tanto, aunque los resultados sean llamativos, no siempre se puede atribuir el efecto exclusivamente a la intervención.

Las claves de lectura crítica son:

- Verificar si el estudio reconoce esta limitación y discute la posibilidad de factores externos que expliquen el cambio.
- Preguntar si se incluyó un grupo comparador (aunque sea retrospectivo o histórico) para contextualizar los resultados.
- Analizar si se utilizaron varios puntos de medición en el tiempo, en lugar de solo dos, lo cual fortalece la interpretación al mostrar tendencias.

En resumen, los estudios antes y después son útiles como primera aproximación, porque permiten detectar cambios asociados a una intervención de forma rápida y económica. No obstante, sus hallazgos deben interpretarse con cautela, ya que la ausencia de un grupo control debilita la atribución causal directa.

4.3 ESTUDIO DE SERIES TEMPORALES INTERRUMPIDAS

Este diseño analiza la evolución de una variable en múltiples puntos de tiempo antes y después de aplicar una intervención. Su fortaleza es que permite diferenciar los cambios espontáneos de la tendencia de aquellos que aparecen de forma abrupta tras la introducción de la medida.

Ejemplo aplicado al caso clínico: el hospital decide implementar el Programa Intensivo de Rehabilitación Neuromotora (PIRN) en enero de 2024. Para evaluar su impacto, los investigadores revisan de manera mensual, durante todo un año, el porcentaje de pacientes con buena recuperación funcional. Si la tendencia se mantenía estable en los meses previos y se observa un cambio brusco después de la introducción del programa, se puede atribuir con mayor confianza la mejoría al PIRN.

Su principal riesgo radica en que no todos los cambios observados en la tendencia se deben a la intervención. Otros factores externos, como la contratación de más fisioterapeutas, la actualización de protocolos médicos o una campaña comunitaria, pueden coincidir con la implementación y explicar la mejoría.

Las claves de lectura crítica son:

- Verificar si se recolectaron suficientes puntos de tiempo antes y después de la intervención; contar solo con uno o dos registros debilita la evidencia.
- Revisar si los autores describieron posibles factores externos que pudieron influir en la tendencia.
- Observar si se analizaron las series con métodos estadísticos apropiados (p. ej., regresión segmentada) que permitan separar cambios abruptos de variaciones graduales.

En síntesis, las series temporales interrumpidas ofrecen un nivel de evidencia mayor que

https://doi.org/10.62486/978-9915-9928-3-9.Ch15

los estudios antes y después, porque incorporan la dimensión del tiempo y permiten observar patrones más sólidos. No obstante, su interpretación requiere un análisis cuidadoso de los posibles factores concurrentes.

4.4 DISEÑO CON CONTROLES HISTÓRICOS

En este tipo de estudio, los pacientes que reciben la nueva intervención se comparan con un grupo de referencia que fue tratado en el pasado, antes de que la intervención estuviera disponible. Es decir, el control no es concurrente, sino histórico.

Ejemplo aplicado al caso clínico: los pacientes que participaron en el Programa Intensivo de Rehabilitación Neuromotora (PIRN) en 2024 se comparan con pacientes tratados únicamente con rehabilitación convencional en 2022, antes de que existiera el PIRN en el hospital.

Su principal riesgo es que los grupos pueden diferir no solo por la intervención, sino también por los cambios en la práctica clínica a lo largo del tiempo. Por ejemplo, en 2024 puede haber más acceso a fármacos, nuevas guías de manejo, mayor capacitación del personal o incluso diferencias en la población que llega al hospital. Todo esto puede influir en los resultados y sesgar la comparación.

Las claves de lectura crítica son:

- Verificar si los autores describieron con detalle las diferencias contextuales entre el periodo histórico y el actual.
- Revisar si se aplicaron ajustes estadísticos para reducir las diferencias entre las cohortes de distintos años.
- Preguntarse si los cambios observados pueden explicarse por mejoras generales del sistema de salud y no necesariamente por la intervención evaluada.

En síntesis, los diseños con controles históricos son una alternativa útil cuando no se dispone de un grupo control concurrente, pero sus resultados deben interpretarse con precaución, porque los cambios propios del tiempo pueden distorsionar la atribución causal.

4.5. DISEÑO CON TÉCNICAS ESTADÍSTICAS DE EMPAREJAMIENTO (PROPENSITY SCORE)

Una de las principales debilidades de los estudios cuasiexperimentales es que los grupos suelen empezar siendo distintos. Para reducir esa diferencia se utiliza el propensity score o “puntuación de propensión”. Esta técnica estima la probabilidad de que un paciente reciba la intervención en función de sus características basales (edad, sexo, comorbilidades, severidad de la enfermedad, entre otras). Luego, se pueden emparejar pacientes con puntajes similares de ambos grupos o aplicar ponderaciones estadísticas que equilibran las diferencias.Ejemplo aplicado al caso clínico: supongamos que los pacientes del PIRN eran, en promedio, más jóvenes y con menor severidad del ACV que los de la rehabilitación convencional. Para evitar que estas diferencias distorsionen los resultados, los investigadores calculan la probabilidad de entrar al PIRN según edad, severidad y comorbilidades. Posteriormente, comparan solo a los pacientes que tienen una probabilidad similar de haber sido asignados a cada programa.

El riesgoradica en que, aunque el método es útil, solo puede corregir las diferencias en variables que fueron medidas. Factores no registrados como la motivación del paciente, la calidad del apoyo familiar o la distancia al hospital pueden seguir introduciendo sesgo.

Las claves de lectura crítica son:

- Revisar qué variables se incluyeron en el cálculo de la puntuación y si son clínicamente relevantes.
- Observar si los autores muestran un balance adecuado entre los grupos después del emparejamiento (tablas o gráficos de distribución de covariables).
- Preguntarse si persisten variables importantes no medidas que podrían afectar los resultados.

El uso de técnicas de emparejamiento mediante propensity score hace que los estudios cuasiexperimentales sean más comparables y robustos, aunque nunca alcanzan la solidez de un ensayo aleatorizado.

4.6 DISEÑOS DE REGRESIÓN CONTINUA

Este diseño se utiliza cuando la asignación a la intervención depende de un punto de corte en una variable continua. Permite comparar pacientes muy similares, situados justo por encima

https://doi.org/10.62486/978-9915-9928-3-9.Ch15

o por debajo del umbral, lo que genera un escenario cercano a la aleatorización.Ejemplo aplicado al caso clínico: el hospital establece que los pacientes con un puntaje de NIHSS ≥10 ingresen automáticamente al Programa Intensivo de Rehabilitación Neuromotora (PIRN), mientras que aquellos con puntajes <10 reciben rehabilitación convencional. En este caso, los pacientes con puntajes de 9 y 10 son prácticamente comparables en severidad, pero fueron asignados a distintos programas por la regla del punto de corte.

El riesgo de este diseño radica en que solo es aplicable cuando existe un criterio de elegibilidad claro y estricto, y sus conclusiones se limitan a los pacientes cercanos al umbral, lo que restringe la generalización.

Las claves de lectura crítica son:

- Confirmar que el umbral de asignación fue respetado y aplicado de forma consistente.
- Evaluar si los pacientes cercanos al punto de corte realmente son comparables.
- Preguntarse si los resultados pueden extrapolarse más allá de los pacientes próximos al umbral.

4.7 DISEÑOS DE ASIGNACIÓN POR ALTERNANCIA O CUOTAS

En lugar de un proceso aleatorio, los pacientes se asignan siguiendo una regla preestablecida o secuencial, como días de la semana, turnos o cuotas de ingreso.

Ejemplo aplicado al caso clínico: los pacientes que ingresan lunes, miércoles y viernes se derivan a PIRN, mientras que los de martes y jueves a rehabilitación convencional. Aunque parece un método ordenado, no es realmente aleatorio, y existe el riesgo de que algunos pacientes elijan acudir en determinado día, introduciendo sesgo.

Su riesgo radica en que, si la asignación sigue un patrón predecible, pueden generarse diferencias sistemáticas entre los grupos, por ejemplo, que pacientes más motivados se presenten en días específicos.

Las claves de lectura crítica son:

- Verificar si la asignación se realizó estrictamente según la regla y sin manipulación por parte de los investigadores o pacientes.
- Analizar si existe algún patrón oculto que pueda explicar diferencias entre los grupos.
- Considerar que este método es más frágil frente a sesgos que la aleatorización verdadera.

Tabla 1. Diseños cuasiexperimentales más utilizados: características, riesgos y claves de lectura crítica

Diseño	Ejemplo aplicado al caso clínico (PIRN vs rehabilitación convencional)	Riesgo principal	Claves de lectura crítica
Grupos no equivalentes	Comparar pacientes asignados a PIRN frente a rehabilitación convencional según cupos disponibles.	Diferencias iniciales entre grupos (edad, severidad, comorbilidades).	Revisar características basales; comprobar si se aplicaron ajustes estadísticos adecuados.
Antes y después (pre–post)	Evaluar independencia funcional en un mismo grupo de pacientes antes y seis meses después de PIRN.	Otros factores (recuperación natural, tratamiento médico estándar) pueden explicar el cambio.	Preguntar si hubo grupo comparador; valorar si se usaron varios puntos de medición.
Series temporales interrumpidas	Medir mensualmente la recuperación funcional antes y después de implementar PIRN en el hospital.	Factores externos coincidentes (contratación de personal, cambios de protocolo).	Verificar número suficiente de mediciones; comprobar si se aplicaron análisis de tendencia apropiados.
Controles históricos	Comparar pacientes tratados con PIRN en 2024 frente a pacientes con rehabilitación convencional en 2022.	Cambios en la práctica clínica y en la población a lo largo del tiempo.	Evaluar si los autores describen diferencias contextuales; revisar si hubo ajustes estadísticos.
Propensity Score (emparejamiento/ ponderación)	Emparejar pacientes de PIRN y de rehabilitación convencional con edad, severidad y comorbilidades similares.	Solo ajusta variables medidas; factores no registrados pueden sesgar.	Revisar variables incluidas en el score; comprobar balance entre grupos tras el emparejamiento.
Regresión discontinua	Asignar PIRN a pacientes con NIHSS ≥10 y rehabilitación convencional a quienes tienen <10.	Conclusiones limitadas a los pacientes cercanos al punto de corte; riesgo de mala aplicación del umbral.	Verificar si el umbral se aplicó de manera estricta; comprobar que los pacientes a ambos lados del corte son comparables; reflexionar sobre la generalización de los resultados.

https://doi.org/10.62486/978-9915-9928-3-9.Ch15

Diseño	Ejemplo aplicado al caso clínico (PIRN vs rehabilitación convencional)	Riesgo principal	Claves de lectura crítica
Asignación por alternancia o cuotas	Derivar a PIRN a pacientes que ingresan lunes/miércoles/viernes y a convencional los martes/jueves.	Riesgo de sesgo si los pacientes eligen acudir en determinados días o si el patrón de ingreso no es aleatorio.	Revisar si la regla se aplicó sin manipulación; analizar posibles patrones de ingreso; considerar que es un método más frágil que la aleatorización verdadera.

Fuente: Elaboración propia.

5. EVALUACIÓN DE LA VALIDEZ CLÍNICA Y RELEVANCIA

Uno de los mayores desafíos al interpretar un estudio cuasiexperimental es determinar qué tan confiables son sus resultados y si realmente pueden aplicarse a la práctica clínica. A diferencia de los ensayos aleatorizados, estos diseños requieren un análisis más cuidadoso para detectar posibles sesgos y valorar su impacto en las conclusiones.La validez clínica y la relevancia de un estudio cuasiexperimental pueden evaluarse en cinco aspectos fundamentales:

Selección de los grupos. La población suele seleccionarse por “conveniencia” e “idoneidad” por lo tanto, es esencial verificar si los grupos comparados eran similares al inicio del estudio. Si no lo eran, se debe analizar si los autores aplicaron estrategias para equilibrar esas diferencias (ajustes estadísticos, emparejamiento por puntuación de propensión, estratificación).Ejemplo del caso clínico: pacientes asignados a PIRN podrían haber sido, en promedio, más jóvenes y menos graves que los de rehabilitación convencional. La lectura crítica debe comprobar si los investigadores describieron bien estas diferencias y cómo las controlaron.Medición de las variables. Los desenlaces deben ser claros, objetivos y medidos de forma estandarizada en todos los grupos. Siempre es deseable que los evaluadores estén cegados a la intervención recibida.*Ejemplo:* la escala Fugl-Meyer o el Índice de Barthel (instrumentos estandarizados) deben aplicarse con el mismo protocolo, sin que el evaluador sepa a qué programa asistió el paciente.Control de factores de confusión. Un buen estudio cuasiexperimental no se limita a comparar resultados crudos, sino que ajusta por variables que podrían explicar las diferencias.Ejemplo: ajustar por edad, tiempo desde el ACV, severidad inicial y comorbilidades, porque todos estos factores influyen en la recuperación, independientemente del tipo de rehabilitación.

Duración y seguimiento. El tiempo de observación debe ser suficiente para detectar efectos clínicamente relevantes, y las pérdidas de seguimiento deben analizarse con transparencia. Una pérdida elevada o desigual entre grupos puede sesgar los resultados.Ejemplo: si en PIRN más pacientes abandonan porque no pueden asistir tres horas al día, los resultados podrían parecer mejores simplemente porque solo permanecieron los más motivados.

Generalización y aplicabilidad. El lector debe preguntarse si los hallazgos se pueden extrapolar a su propio contexto clínico. Un programa exitoso en un hospital de referencia con gran infraestructura puede no tener el mismo impacto en un centro de menor complejidad.

6. INTERPRETACIÓN ESTADÍSTICA

La estadística en un estudio cuasiexperimental no es un fin en sí misma, sino una herramienta que ayuda al lector a distinguir si los cambios observados se deben realmente a la intervención o a factores ajenos. Para la lectura crítica, lo más importante no es memorizar las pruebas utilizadas, sino saber qué significan los resultados y cómo interpretarlos clínicamente.

6.1 COMPARACIONES INICIALES

El primer paso es analizar si los grupos eran comparables antes de la intervención. Esto se suele presentar en la tabla de características basales, donde se describen variables como edad, sexo, gravedad de la enfermedad o presencia de comorbilidades. Si desde el inicio un grupo está compuesto por pacientes más jóvenes, con menos comorbilidades o con una enfermedad menos severa, es muy probable que ese grupo muestre mejores resultados, aun cuando ambos reciban exactamente el mismo tratamiento.

Para entenderlo mejor se presenta la siguiente tabla del estudio:

https://doi.org/10.62486/978-9915-9928-3-9.Ch15

Tabla 2. Características basales de los pacientes incluidos en el estudio cuasiexperimental sobre rehabilitación post-ACV isquémico en el Hospital de Clínicas de La Paz (n=100).

Variable	PIRN (n=50)	Convencional (n=50)	Diferencia (IC95%)	p-valor
Edad (años, media ± DE)	62 ± 10	68 ± 9	–6 (–10 a –2)	0,01
Sexo masculino (%)	60	58	+2 (–18 a +22)	0,84
NIHSS basal (media ± DE)	9,2 ± 2,1	11,0 ± 2,3	–1,8 (–2,9 a –0,7)	0,002
Tiempo a inicio de rehab. (días)	6 ± 2	8 ± 3	–2 (–3,4 a –0,6)	0,03
Hipertensión (%)	70	74	–4 (–22 a +14)	0,65
Diabetes mellitus (%)	30	28	+2 (–15 a +19)	0,82
Fibrilación auricular (%)	18	22	–4 (–18 a +10)	0,6
mRS previo al ACV (0–2, %)	96	94	+2 (–7 a +11)	0,72

La columna de diferencias incluye el intervalo de confianza al 95% (IC95%) para cada comparación. PIRN: ProgramaIntensivo de RehabilitaciónNeuromotora; ACV: accidente cerebrovascular; NIHSS: National Institutes of Health Stroke Scale; mRS: modified Rankin Scale. El valor de p se calculó con la prueba t de Student para variables continuas y la prueba χ^2 para variables categóricas. Un $p < 0,05$ se considera significativo.

La tabla 2 de características basales es el primer filtro que el lector debe revisar en cualquier estudio cuasiexperimental. Allí se observa si los grupos de comparación (en este caso, PIRN vs rehabilitación convencional) eran semejantes al inicio, hay tres diferencias importantes:

- **Edad:** el grupo PIRN tiene una media de 62 años, frente a 68 en el grupo convencional. La diferencia es de –6 años (IC95%: –10 a –2), estadísticamente significativa (*p* = 0,01). Esta diferencia no es trivial: los pacientes más jóvenes suelen recuperarse mejor tras un ACV, lo que podría favorecer artificialmente al PIRN.
- **Severidad inicial del ACV (NIHSS):** el grupo PIRN presenta menor gravedad inicial (9,2 puntos vs 11,0). La diferencia de –1,8 puntos (IC95%: –2,9 a –0,7) también es significativa (*p* = 0,002). Esta es quizá la diferencia más relevante, porque la severidad inicial es uno de los predictores más fuertes de la recuperación neurológica.
- **Tiempo hasta el inicio de la rehabilitación:** los pacientes de PIRN comenzaron en promedio dos días antes que los del grupo convencional (6 vs 8 días, *p* = 0,03). El inicio precoz de la rehabilitación es un factor reconocido que mejora los desenlaces, lo que puede introducir un sesgo a favor de PIRN.

En contraste, otras variables como sexo, hipertensión, diabetes, fibrilación auricular y función previa (mRS antes del ACV) muestran distribuciones similares, sin diferencias significativas. Esto significa que, al menos en esas dimensiones, los grupos son comparables.

Por ello, el lector debe acostumbrarse a revisar con cuidado la tabla basal y preguntarse: ¿los grupos eran realmente comparables? ¿El estudio aplicó algún tipo de ajuste para equilibrar esas diferencias? Esta reflexión es esencial para determinar si los resultados reflejan el efecto de la intervención o si, en realidad, lo que se observa es consecuencia de un sesgo de selección.

6.2 CAMBIOS DENTRO DE CADA GRUPO

Un error frecuente al interpretar estudios cuasiexperimentales es entusiasmarse únicamente con la mejoría observada en el grupo que recibe la intervención. Que un grupo de pacientes mejore no significa necesariamente que el tratamiento sea responsable de ese cambio. La comparación entre grupos es la que otorga validez al hallazgo. Observar mejoría dentro de un solo grupo puede ser interesante, pero no basta para atribuir causalidad.

Para poder entender mejor ello, se presenta la siguiente tabla:

Tabla 3. Resultados clínicos a 6 meses en pacientes con accidente cerebrovascular isquémico tratados con Programa Intensivo de Rehabilitación Neuromotora (PIRN) o rehabilitación convencional (n=100).

Desenlace	PIRN (n=50)	Convencional (n=50)	Diferencia ajustada (IC95%)	p-valor
Fugl-Meyer (cambio, media ± DE)	+15 ± 5	+8 ± 6	+6,5 (3,2 a 9,8)	<0,001
Barthel (cambio, media ± DE)	+40 ± 10	+28 ± 12	+9,0 (4,0 a 14,0)	0,002
mRS 0–2 (%)	60	40	RR 1,5 (1,1 a 2,1)	0,01
Reingresos hospitalarios (%)	10	18	–8 (–20 a +4)	0,18

https://doi.org/10.62486/978-9915-9928-3-9.Ch15

Desenlace	PIRN (n=50)	Convencional (n=50)	Diferencia ajustada (IC95%)	p-valor
Adherencia (≥80% sesiones, %)	70	78	–8 (–24 a +8)	0,3

La columna de diferencias muestra estimaciones ajustadas por edad, severidad inicial (NIHSS) y tiempo al inicio de rehabilitación. IC95%: intervalo de confianza al 95%; RR: riesgo relativo. Un p < 0,05 se considera significativo.

Para su interpretación crítica:

- Escalas motoras y funcionales (Fugl-Meyer y Barthel): los pacientes de PIRN mejoraron más que los de rehabilitación convencional. La diferencia ajustada en Fugl-Meyer fue de 6,5 puntos (IC95%: 3,2 a 9,8), lo que indica no solo significancia estadística (p < 0,001), sino también una diferencia clínicamente relevante.
- Discapacidad global (mRS): el 60% de los pacientes en PIRN alcanzó independencia funcional (mRS 0–2), frente al 40% en el grupo convencional. El riesgo relativo de 1,5 (IC95%: 1,1 a 2,1) significa que los pacientes en PIRN tuvieron un 50% más de probabilidad de lograr un buen desenlace.
- Reingresos hospitalarios: fueron menores en PIRN (10% vs 18%), pero la diferencia no fue significativa (IC95% incluye 0, p = 0,18). Esto debe interpretarse con cautela: podría ser un hallazgo por azar o un efecto real que el estudio no tuvo poder suficiente para detectar.
- Adherencia: la adherencia fue algo menor en PIRN (70% vs 78%), aunque sin diferencias significativas. Esto sugiere que la intensidad del programa puede haber dificultado la asistencia completa, un aspecto relevante para valorar la aplicabilidad en la práctica clínica.

6.3 COMPARACIÓN ENTRE GRUPOS

El corazón de todo estudio cuasiexperimental está en la comparación entre los grupos. No basta con saber que los pacientes de un programa mejoran, sino determinar si esa mejoría es mayor que la observada en el grupo de control.

Primero, el lector debe preguntarse si esa diferencia es estadísticamente significativa, es decir, si el resultado probablemente no se deba al azar.

Segundo, y aún más importante, se debe valorar si la diferencia es clínicamente relevante.

Para entender ello, se usará la tabla 3 para interpretar la comparación entre grupos donde los pacientes del PIRN mejoraron en promedio 15 puntos en la escala Fugl-Meyer, mientras que los de rehabilitación convencional mejoraron 8 puntos. La diferencia bruta es de 7 puntos a favor del PIRN. A primera vista, este dato suena convincente. Sin embargo, el lector crítico no debe quedarse solo con el número. Lo relevante es analizar dos aspectos:

- La precisión estadística del hallazgo. Esta se refleja en el intervalo de confianza (IC). Si la diferencia fue de 7 puntos con un IC95%: 3 a 11, significa que, incluso en el escenario más conservador, el PIRN sigue ofreciendo al menos 3 puntos adicionales de recuperación frente al tratamiento convencional. (En diferencias de medias, el IC no debe cruzar el 0: si lo hace, podría indicar que la intervención no tiene efecto real, en medidas de asociación como riesgos relativos u odds ratio, el IC no debe cruzar el 1, porque esto indicaría ausencia de efecto).
- La relevancia clínica.Más allá de la estadística, el lector debe preguntarse si el efecto observado tiene un impacto en la vida del paciente. En rehabilitación post-ACV, una ganancia de 7 puntos en Fugl-Meyer puede significar la diferencia entre un paciente que recupera la movilidad funcional del brazo y otro que permanece con limitaciones significativas.

En síntesis, la comparación entre grupos no se limita a “quién mejoró más”, sino que requiere responder dos preguntas clave:¿La diferencia observada es estadísticamente confiable (IC y *p*valor) ?, ¿Esa diferencia es lo suficientemente grande como para tener impacto en la práctica clínica y en la calidad de vida del paciente?

6.4 AJUSTES POR DIFERENCIAS BASALES

Las diferencias entre grupos al inicio pueden sesgar los resultados. Por eso, los análisis ajustados —que corrigen por variables como edad, severidad de la enfermedad o tiempo hasta la intervención— son más confiables que las comparaciones crudas.

Los investigadores aplican métodos estadísticos para corregir esas diferencias basales. La idea es “nivelar el terreno” y que la comparación sea lo más justa posible. Existen varias estrategias:

- **Regresión multivariada:** es la más común. Se introducen en el modelo estadístico las

https://doi.org/10.62486/978-9915-9928-3-9.Ch15

variables que diferían entre grupos (por ejemplo, edad, severidad inicial, tiempo al inicio de la rehabilitación). Así se calcula el efecto de la intervención corrigiendo por esas variables.

- **Propensity Score (puntuación de propensión):** se estima la probabilidad de que cada paciente reciba la intervención según sus características basales. Luego se emparejan pacientes con puntuaciones similares o se aplica un peso estadístico. Esto ayuda a equilibrar los grupos.
- **Estratificación:** se divide la muestra en subgrupos homogéneos (por ejemplo, pacientes menores de 65 y mayores de 65 años) y se compara dentro de cada estrato.

En nuestro caso se realiza el ajuste con un modelo de regresión multivariada, que incluyó la edad, el puntaje NIHSS basal y el tiempo hasta el inicio de la rehabilitación contando con la siguiente tabla:

Tabla 4. Resultados principales (cambio en Fugl-Meyer a 6 meses) antes y después del ajuste por variables basales

Análisis	PIRN (n=50)	Convencional (n=50)	Diferencia (IC95%)	p-valor
Crudo	+15 ± 5	+8 ± 6	+7,0 (3,5 a 10,5)	<0,001
Ajustado*	+14,5 ± 5	+8,5 ± 6	+6,0 (2,8 a 9,2)	<0,001

*Ajustado por edad, severidad del ACV (NIHSS) y tiempo al inicio de rehabilitación.

Para su interpretación crítica:

- Análisis crudo: la diferencia inicial es de 7 puntos (IC95%: 3,5 a 10,5). Aunque significativa, esta estimación no considera que los pacientes de PIRN eran más jóvenes y menos graves.
- Análisis ajustado: al corregir por edad, severidad y tiempo al inicio, la diferencia se reduce a 6 puntos (IC95%: 2,8 a 9,2), pero sigue siendo estadística y clínicamente significativa. Esto aumenta la confianza en que el efecto observado se debe al programa y no solo a un sesgo de selección.
- Reflexión clave para el lector: si el efecto desaparece después del ajuste, probablemente estaba explicado por diferencias iniciales entre los grupos. En cambio, si el efecto se mantiene (aunque sea un poco menor), es más confiable atribuirlo a la intervención.

6.5 MEDIDAS DE EFECTO MAS COMUNES EN ESTUDIOS CUASIEXPERIMENTALES

Los resultados suelen expresarse mediante tres medidas de efecto: diferencias de medias, riesgos relativos (RR) y odds ratio (OR). La Tabla 2 del caso clínico ofrece ejemplos claros de cómo se reportan y cómo deben interpretarse.

Diferencia de medias: Es la distancia entre los promedios de un desenlace en dos grupos. Se usa cuando la variable es continua (p. ej., puntuaciones de escalas clínicas).

En la escala Fugl-Meyer, los pacientes del PIRN mejoraron 15 puntos, frente a 8 puntos en el grupo convencional. La diferencia ajustada fue de 6,5 puntos (IC95%: 3,2 a 9,8; p<0,001).

- Interpretación: los pacientes en PIRN mejoraron, en promedio, 6,5 puntos más que los de rehabilitación convencional.
- Clave crítica: como el IC no cruza el 0, la diferencia es estadísticamente significativa. Además, una ganancia de más de 5 puntos en Fugl-Meyer se considera clínicamente importante, lo que refuerza la relevancia del hallazgo.

Riesgo relativo (RR): Es la razón de probabilidades absolutas de que ocurra un desenlace en un grupo respecto al otro. Se usa con variables dicotómicas (sí/no).

Al analizar el grado de independencia global (mRS 0–2), el 60% de los pacientes en PIRN alcanzó este desenlace, frente al 40% en el grupo convencional. El RR fue 1,5 (IC95%: 1,1 a 2,1; p=0,01).

- Interpretación: los pacientes en PIRN tuvieron 1,5 veces más probabilidad de lograr independencia funcional que los de rehabilitación convencional.
- Clave crítica: como el IC no cruza el 1, el resultado es significativo. Además, un aumento relativo del 50% en la independencia funcional tiene gran impacto clínico.

Odds ratio (OR): Aunque en esta tabla no se muestra un OR explícito, algunos artículos lo reportan en lugar del RR. La lógica es similar:

- Ejemplo hipotético: un OR de 2,0 para independencia funcional significa que las “odds” (oportunidades) de lograr ese desenlace se duplican con PIRN frente a rehabilitación convencional.
- Clave crítica: igual que con el RR, un OR es significativo si su IC no cruza el 1.

https://doi.org/10.62486/978-9915-9928-3-9.Ch15

Tabla 5. Medidas de efecto más comunes en estudios cuasiexperimentales. Ejemplo aplicado al caso

Medida de efecto	Definición básica	clínico (PIRN vs rehabilitación convencional)	Clave de interpretación
Diferencia de medias (DM)	Distancia entre los promedios de un desenlace en dos grupos. Se aplica a variables continuas.	Fugl-Meyer: PIRN mejoró 15 puntos, convencional 8. La DM = 7 puntos a favor de PIRN.	El IC no debe incluir 0. Una diferencia distinta de 0 indica un efecto real.
Riesgo relativo (RR)	Razón entre las probabilidades absolutas de un desenlace en dos grupos. Se aplica a variables dicotómicas (sí/no).	Independencia funcional (mRS 0–2): PIRN 60%, convencional 40%. El RR = 1,5.	El IC no debe incluir 1. Un RR > 1 indica mayor probabilidad de desenlace con la intervención.
Odds ratio (OR)	Razón de las "odds" (oportunidades) de un desenlace en un grupo respecto al otro. Similar al RR, pero usa odds en lugar de riesgos absolutos.	Hipotético: OR de independencia funcional con PIRN = 2,0, lo que significa que las oportunidades de lograrla se duplican frente al grupo convencional.	El IC no debe incluir 1. Un OR > 1 indica mayor oportunidad de desenlace con la intervención.

Fuente: Elaboración propia.

6.6 ANÁLISIS DE SENSIBILIDAD

El análisis de sensibilidad es un paso clave en los estudios cuasiexperimentales, porque permite comprobar si los resultados son robustos, es decir, si se mantienen estables cuando se modifican ciertas condiciones del análisis. En otras palabras, es una forma de preguntarse: ¿seguiríamos llegando a la misma conclusión si analizáramos los datos de otra manera?

- El análisis de sensibilidad puede hacerse de distintas maneras:
- Excluyendo pacientes perdidos o con datos incompletos, para ver si los resultados cambian.
- Probando distintos métodos de ajuste (por ejemplo, distintos modelos de regresión o diferentes calipers en el emparejamiento por puntuación de propensión).
- Analizando subgrupos específicos (por ejemplo, solo pacientes con ACV moderado o solo los mayores de 65 años).

Para su interpretación: si los resultados cambian drásticamente en los análisis de sensibilidad, significa que las conclusiones del estudio son frágiles y deben interpretarse con cautela. En cambio, si los resultados se mantienen similares bajo distintos escenarios, el lector puede tener mayor confianza en la solidez de la evidencia.

Para comprenderlo se presenta la siguiente tabla:

Tabla 6. Análisis de sensibilidad para el desenlace primario (cambio en Fugl-Meyer a 6 meses) en pacientes con ACV isquémico (n=100).

Análisis	PIRN (n=50)	Convencional (n=50)	Diferencia (IC95%)	p-valor
Crudo	+15 ± 5	+8 ± 6	+7,0 (3,5 a 10,5)	<0,001
Ajustado*	+14,5 ± 5	+8,5 ± 6	+6,0 (2,8 a 9,2)	<0,001
Excluyendo abandonos	+15 ± 5	+9 ± 6	+6,0 (2,5 a 9,5)	0,001
Propensity score (emparejamiento 1:1)	+14,8 ± 5	+9,1 ± 6	+5,7 (2,3 a 9,1)	0,002
Propensity score (ponderación por IPTW)	+14,6 ± 5	+8,8 ± 6	+5,8 (2,6 a 9,0)	0,001

* Ajustado por edad, severidad inicial (NIHSS) y tiempo al inicio de rehabilitación.

Para su interpretación crítica de la tabla

- Consistencia de los resultados: en todos los escenarios (crudo, ajustado, excluyendo abandonos, distintos métodos de propensity score), la diferencia entre PIRN y convencional se mantiene entre 5,7 y 7 puntos, siempre significativa.
- Robustez: como los hallazgos no cambian sustancialmente al modificar la estrategia de análisis, se puede tener mayor confianza en que el beneficio del PIRN es real y no producto de un artefacto estadístico.
- Reflexión crítica para el lector: si en algún escenario los resultados hubieran perdido significancia o se hubieran invertido, la conclusión sería frágil. El lector debe preguntarse: ¿el efecto se mantiene estable en todos los análisis de sensibilidad?

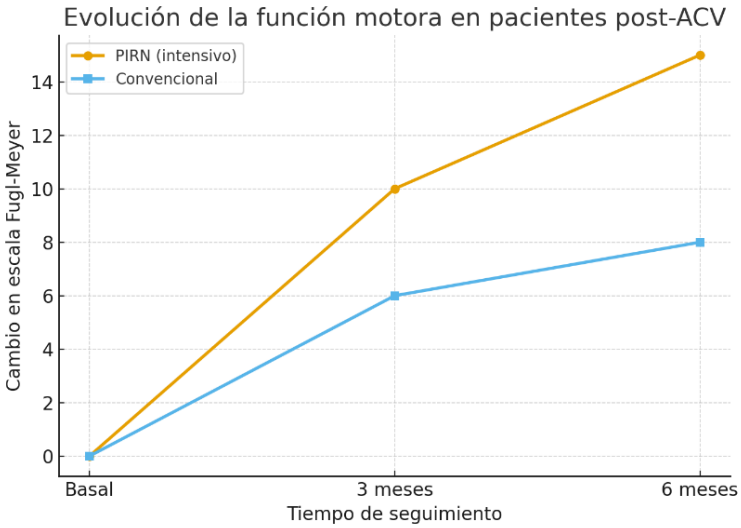
https://doi.org/10.62486/978-9915-9928-3-9.Ch15

6.7 REPRESENTACIÓN GRÁFICA DE RESULTADOS

Gráfico de evolución en el tiempo (líneas o barras): Los gráficos permiten visualizar la evolución y la magnitud de las diferencias entre grupos de forma más clara que las tablas, donde:

- Los gráficos de líneas se utilizan para mostrar cambios a lo largo del tiempo, permitiendo observar tendencias y la pendiente de mejoría.
- Los gráficos de barras se emplean para comparar proporciones o valores finales entre grupos, facilitando la interpretación de diferencias puntuales.

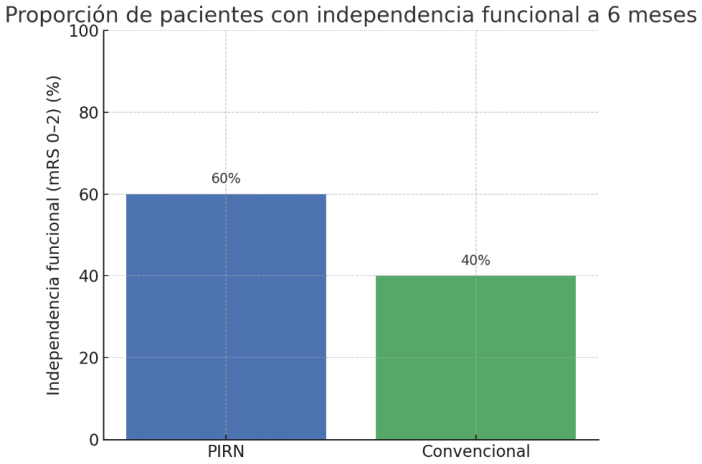
Figura 1: Evolución del cambio en la escala Fugl-Meyer en pacientes post-ACV, medido al inicio, a los 3 y a los 6 meses.



Para su interpretación: ambos grupos mejoran con el tiempo, pero la curva del PIRN es más pronunciada y se separa progresivamente de la convencional. Esto sugiere que el beneficio adicional puede atribuirse al programa intensivo, siempre que la diferencia sea estadísticamente significativa.

Gráfico de barras comparativas: Este tipo de gráfico es muy usado en estudios cuasiexperimentales porque permite comparar proporciones de desenlaces entre dos o más grupos de forma inmediata y visual. Cada barra representa la frecuencia o porcentaje de un desenlace (por ejemplo, el porcentaje de pacientes que logran independencia funcional).Para la Interpretación de este tipo de graficas lo primero es recordar que la diferencia visual entre las barras no garantiza por sí misma un efecto real, es indispensable comprobar si esa diferencia es estadísticamente significativa, además, el lector debe valorar si la diferencia es clínicamente relevante y otro aspecto importante es considerar el tamaño muestral, para entender mejor la interpretación grafica se presenta lo siguiente:

Figura 2: Proporción de pacientes con independencia funcional (mRS 0–2) a los 6 meses.



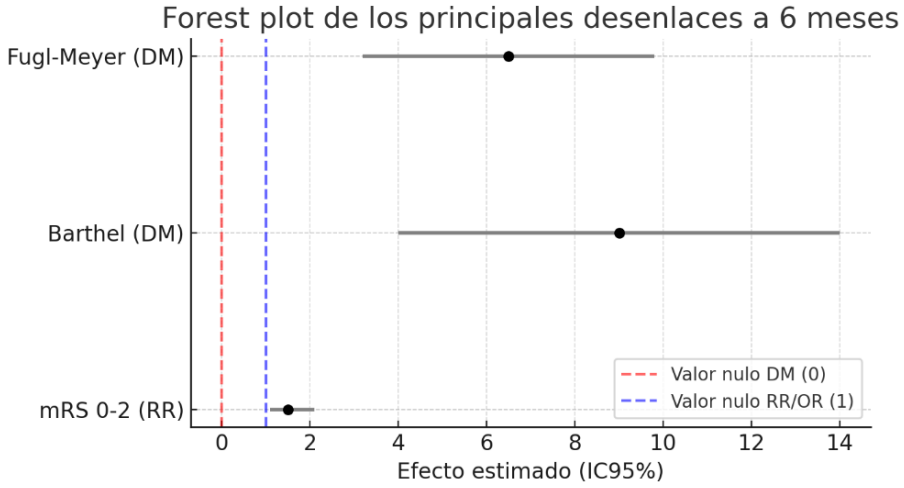
https://doi.org/10.62486/978-9915-9928-3-9.Ch15

Para su interpretación: el 60% de los pacientes en PIRN alcanzó independencia funcional frente al 40% en la rehabilitación convencional. La diferencia se visualiza de inmediato en la altura de las barras. Lo importante no es solo que una barra sea más alta, sino comprobar si esa diferencia es estadísticamente significativa y si tiene un impacto clínico relevante en la recuperación de los pacientes.

Gráfico de Forest Plot: Es una forma visual de resumir los resultados principales de un estudio. Cada desenlace se representa con un punto central (el efecto estimado) y una línea horizontal (el intervalo de confianza al 95%), los principios generales de interpretación son los mismos:

- Si el desenlace es una diferencia de medias, el valor nulo es 0.
- Si el desenlace es un riesgo relativo (RR) u odds ratio (OR), el valor nulo es 1.
- La regla de oro es que, si la línea del intervalo cruza el valor nulo, el resultado no es estadísticamente significativo.

Figura 3: Principales desenlaces a 6 meses en pacientes post-ACV (PIRN vs rehabilitación convencional).



Para su interpretación crítica:

- Fugl-Meyer (DM 6,5 puntos, IC95%: 3,2 a 9,8): el intervalo no cruza el 0, lo que indica una diferencia real a favor de PIRN.
- Barthel (DM 9 puntos, IC95%: 4,0 a 14,0): el efecto también es estadística y clínicamente significativo.
- Independencia funcional mRS 0–2 (RR 1,5, IC95%: 1,1 a 2,1): como el IC no cruza el 1, los pacientes en PIRN tienen 1,5 veces más probabilidad de lograr independencia funcional.

6.8 OTROS PARÁMETROS A CONSIDERAR

En la práctica de lectura crítica hay otros aspectos adicionales que vale la pena mencionar brevemente para cerrar esta sección:

Tamaño muestral y potencia estadística: Un resultado no significativo ($p > 0,05$) no siempre significa ausencia de efecto: puede deberse a que el estudio tuvo muy pocos pacientes para detectar una diferencia real. Se debe revisar si los autores calcularon el tamaño muestral y si discuten la posibilidad de error tipo II (falso negativo).

- Ejemplo: si solo 20 pacientes completaron PIRN y 20 la rehabilitación convencional, incluso diferencias grandes pueden no alcanzar significancia.

Magnitud del efecto vs. Significancia: Un resultado puede ser estadísticamente significativo, pero clínicamente irrelevante. El lector debe siempre balancear significancia estadística con relevancia clínica.

- Ejemplo: una diferencia de 1 punto en Fugl-Meyer puede ser significativa si el estudio es muy grande, pero no cambia la vida del paciente.

Pérdidas de seguimiento y análisis por protocolo/intención de tratar: En estudios cuasiexperimentales, las pérdidas pueden ser diferentes entre grupos y distorsionar los resultados. Se debe revisar si los autores usaron análisis "por intención de tratar" (todos los pacientes

https://doi.org/10.62486/978-9915-9928-3-9.Ch15

asignados) o solo "por protocolo" (solo los que completaron).

- Si muchos pacientes abandonan PIRN porque es más exigente, los resultados pueden reflejar solo a los más motivados.

Subgrupos y heterogeneidad de efectos: Algunos estudios reportan resultados diferentes en ciertos subgrupos (ej., mayores vs menores de 65 años). El lector debe ser cauto: si se hacen demasiados subgrupos, puede aparecer significancia por azar.Estos análisis ayudan a entender mejor para quién funciona la intervención.

Tabla 7. Guía de interpretación estadística en estudios cuasiexperimentales

Aspecto a revisar	Qué observar	Ejemplo aplicado al caso PIRN	Claves de interpretación crítica
1. Comparación inicial de grupos	Tabla de características basales	PIRN más jóvenes, menos graves y con inicio más precoz	Si las diferencias no se ajustan, los resultados estarán sesgados.
2. Cambios dentro de cada grupo	Evolución pre–post en cada rama	PIRN mejora 15 puntos en Fugl-Meyer	La mejoría puede deberse a recuperación natural. Se necesita comparar entre grupos.
3. Comparación entre grupos	Diferencia de mejoría entre intervención y control	PIRN +15 vs Convencional +8 → diferencia de 7	Lo relevante es la diferencia y si el IC no cruza el valor nulo.
4. Ajustes por diferencias basales	Análisis crudo vs ajustado	Diferencia ajustada = 6 puntos	Si el efecto se mantiene tras el ajuste, es más confiable.
5. Medidas de efecto	DM, RR, OR según el desenlace	DM = 6,5 puntos; RR = 1,5	En DM el IC no debe cruzar 0; en RR/OR no debe cruzar 1.
6. Intervalos de confianza y p-valor	Precisión y significancia	IC95%: 3,2 a 9,8; p<0,001	El IC informa magnitud y precisión; no basta con el p-valor.
7. Análisis de sensibilidad	Resultados bajo distintos escenarios	Excluyendo abandonos: diferencia se mantiene en 6	Si los resultados son consistentes, el hallazgo es robusto.
8. Representación gráfica	Líneas, barras, forest plots	Líneas divergen; barra PIRN más alta; IC sin cruzar el nulo	Gráficos ayudan a visualizar tendencias y la magnitud del efecto.
9. Tamaño muestral y potencia	Si hubo cálculo de muestra	n=100 puede detectar diferencias claras	Un estudio pequeño puede no mostrar efectos aunque existan.
10. Magnitud del efecto vs relevancia clínica	Diferencia significativa pero pequeña	Ganancia de 1 punto en Fugl-Meyer, aunque p<0,05	Puede ser irrelevante para el paciente. Importa la utilidad clínica.
11. Pérdidas de seguimiento	Porcentaje y manejo de abandonos	PIRN con más pérdidas por intensidad	Pérdidas desiguales sesgan; preferir análisis por intención de tratar.
12. Subgrupos y heterogeneidad	Análisis por edad, severidad, etc.	PIRN más efectivo en <65 años	Interpretar con cautela: muchos subgrupos aumentan riesgo de azar.

Fuente: Elaboración propia.

7. SESGOS Y LIMITACIONES

A diferencia de los ensayos clínicos aleatorizados, los estudios cuasiexperimentales son más vulnerables a sesgos. Estos sesgos no invalidan necesariamente el estudio, pero sí obligan al lector a interpretarlo con cautela. La clave de la lectura crítica está en identificar de qué manera esos sesgos pudieron influir en los resultados y si los autores tomaron medidas para reducirlos.

Uno de los más frecuentes es el sesgo de selección, que ocurre cuando los grupos no son equivalentes desde el inicio. Si un grupo comienza con pacientes más jóvenes o menos graves, sus resultados tenderán a ser mejores independientemente de la intervención.*Ejemplo PIRN:* los pacientes asignados al programa intensivo eran más jóvenes, tenían menor severidad de ACV y comenzaron la rehabilitación antes que los del grupo convencional.

El sesgo de confusión aparece cuando una variable externa, no medida, se asocia tanto con la intervención como con el desenlace, generando una falsa relación de causalidad.Ejemplo PIRN: el apoyo familiar podría haber sido mayor en los pacientes que ingresaron al programa intensivo, y este factor por sí solo facilita la recuperación funcional.

También es importante considerar el sesgo de información, que se produce cuando los desenlaces se miden de manera diferente entre grupos. Esto puede suceder si el evaluador conoce la intervención recibida y su expectativa influye en la medición.*Ejemplo PIRN:* si el evaluador sabe que un paciente participó en PIRN, podría calificar con mayor optimismo la escala Fugl-Meyer.

Otro riesgo es el **sesgo por pérdidas de seguimiento**. Si un grupo pierde más pacientes durante el seguimiento, los resultados pueden no reflejar la realidad de toda la cohorte, sino solo de quienes permanecieron.*Ejemplo PIRN:* algunos pacientes pueden abandonar el programa

https://doi.org/10.62486/978-9915-9928-3-9.Ch15

intensivo por su alta exigencia, quedando en el análisis final solo los más motivados.

Finalmente, se debe reconocer que existen **limitaciones propias del diseño cuasiexperimental**. Estos estudios reflejan mejor la práctica clínica cotidiana, pero no siempre logran controlar por completo los factores externos que ocurren de manera simultánea.*Ejemplo PIRN:* en paralelo a la implementación del programa, el hospital pudo haber contratado más fisioterapeutas o mejorado sus protocolos de atención, lo que también impactaría en los resultados.

8. COMPARACIONES CON OTROS ESTUDIOS

Una parte esencial de la lectura crítica es situar los resultados dentro del marco de la evidencia existente. Un estudio aislado difícilmente puede cambiar la práctica clínica; es al contrastarlo con otros trabajos similares donde se mide su verdadero valor.

Al comparar un estudio cuasiexperimental con otros, el lector debe preguntarse:

- ¿Los hallazgos son coherentes con lo que se ha observado en ensayos clínicos aleatorizados, cohortes o revisiones sistemáticas?
- ¿Existen diferencias importantes y, si es así, pueden explicarse por variaciones en la población estudiada, la intensidad de la intervención o la metodología empleada?
- ¿El contexto del estudio (hospital público, recursos limitados, población local) aporta información complementaria que no estaba presente en la literatura previa?

Un estudio cuasiexperimental bien conducido que coincide con la evidencia de mayor jerarquía puede reforzar la aplicabilidad de esa intervención en escenarios del mundo real. Por el contrario, si muestra resultados muy distintos, obliga al lector a reflexionar sobre posibles sesgos o sobre características propias del contexto que podrían explicar la discrepancia.

En nuestro caso clínico: El programa PIRN mostró una ganancia media de 6 puntos en la escala Fugl-Meyer a los seis meses. Al contrastar este resultado con la literatura internacional, se observa que varios ensayos clínicos controlados han reportado beneficios similares de la rehabilitación intensiva post-ACV, aunque en contextos con más recursos y bajo condiciones muy controladas. El estudio cuasiexperimental aporta así un valor añadido: confirma que un programa intensivo también puede ser efectivo en un hospital público de altura, con limitaciones de personal y cupos.

9. Checklist. En cuasiexperimentales la fortaleza no depende solo de “si hubo diferencia”, sino de cómo se asignaron los grupos, qué tanto se controló la confusión, qué tan bien se describió la intervención (para poder reproducirla) y qué tan transparente es el reporte. Para juzgar esto, conviene combinar estándares de reporte (qué debe describir el artículo: TREND/STROBE/TIDieR) con estándares de validez (qué riesgos de sesgo revisar: ROBINS-I) y criterios específicos por diseño (p. ej., ITS según EPOC/Cochrane).

Tabla 8. Checklist internacional para lectura crítica de estudios cuasiexperimentales

Sección	Pregunta clave	Fuente / Guía	Sí / No / N.A.
A. Identificación del diseño y encuadre	¿Se identifica claramente el diseño cuasiexperimental (grupos no equivalentes, antes-después, CBA, ITS) y la unidad de análisis?	EPOC (Cochrane)	□
	¿Se justifica por qué no fue posible aleatorizar (factibilidad/ ética)?	TREND (CDC)	□
B. Población, escenario y selección	¿Se describen criterios de inclusión/exclusión, fuente de pacientes, entorno y periodo de estudio?	STROBE (SciELO)	□
	¿Se explica el proceso de derivación/asignación administrativa y sus implicancias en comparabilidad basal?	TREND / ROBINS-I	□
C. Intervención y comparador (reproducibilidad)	¿La intervención y el comparador están descritos según estándar TIDieR (qué, quién, cómo, dónde, duración, adaptaciones, fidelidad)?	TIDieR (EQUATOR)	□
D. Medición de desenlaces y cegamiento	¿Se definen desenlaces primarios/secundarios, instrumentos validados, ventanas temporales y procedimientos de cegamiento de evaluadores?	STROBE / TREND	□
	¿Se reportan eventos adversos y adherencia/fidelidad de la intervención?	TREND / TIDieR	□
E. Comparabilidad basal y control de confusión	¿Se presentan características basales completas y balance entre grupos?	STROBE / TREND	□
	¿Se aplican métodos de control de confusión (regresión, propensity score, DiD, ITS)?	TREND / EPOC / Cochrane	□

https://doi.org/10.62486/978-9915-9928-3-9.Ch15

Sección	Pregunta clave	Fuente / Guía	Sí / No / N.A.
F. Supuestos del método y especificaciones por diseño	En DiD: ¿se evalúa la asunción de tendencias paralelas y se discuten common shocks?	SpringerLink	□
	En ITS: ¿hay suficientes puntos pre y post (≥3) y se usa regresión segmentada con autocorrelación/estacionalidad?	EPOC (Cochrane)	□
G. Estimación y presentación de efectos	¿Se reportan medidas de efecto apropiadas (DM, RR, OR) con IC95% y p-valores, y se discute relevancia clínica?	STROBE / TREND	□
	¿Se presentan tablas/figuras interpretables (líneas de tiempo, barras, forest plot con valores nulos 0/1)?	STROBE (buenas prácticas)	□
H. Manejo de datos y análisis complementarios	¿Se explican datos faltantes (método y justificación) y se hacen análisis de sensibilidad?	TREND / ROBINS-I	□
	¿Se reporta análisis por intención de tratar o por protocolo, y pérdidas de seguimiento cuantificadas?	STROBE / TREND	□
I. Riesgo de sesgo (ROBINS-I)	¿Se evalúan los siete dominios (confusión, selección, clasificación, desviaciones, datos faltantes, medición de desenlace, reporte selectivo) y se presenta juicio global?	ROBINS-I (BMJ/PMC)	□
J. Validez externa, equidad y contexto	¿Se discute la generalización, subgrupos relevantes y equidad?	STROBE	□
K. Transparencia y ética	¿Se informa financiación/conflictos, aprobación ética, protocolo/registro o disponibilidad de datos/código?	STROBE / TREND	□
L. Coherencia con la evidencia	¿Se comparan los resultados con RCTs, cohortes o revisiones y se discuten discrepancias plausibles?	Buenas prácticas / SQUIRE	□

Fuente: Elaboración propia.

10. UTILIDAD DE LOS ESTUDIOS CUASIEXPERIMENTALES

La principal fortaleza de los estudios cuasiexperimentales es su aplicabilidad en la práctica clínica real. Aunque no alcanzan el rigor metodológico de los ensayos clínicos aleatorizados, estos estudios son particularmente útiles para:

Generar evidencia en contextos reales.A diferencia de los ensayos clínicos, que se realizan en condiciones altamente controladas, los cuasiexperimentales reflejan las dinámicas cotidianas de hospitales, programas comunitarios o políticas de salud.

Evaluar intervenciones a nivel poblacional o de sistemas de salud.Muchas políticas públicas o programas hospitalarios no pueden asignarse al azar. Los cuasiexperimentales permiten valorar su impacto de manera estructurada.

Proveer evidencia preliminar para futuros ensayos clínicos.Un resultado consistente en un cuasiexperimental puede servir de base para diseñar un ensayo aleatorizado posterior.

Ofrecer respuestas rápidas. En situaciones donde no se puede esperar años para un RCT, un cuasiexperimental bien diseñado puede aportar información útil para orientar decisiones clínicas o de gestión.

En síntesis, la utilidad de los cuasiexperimentales radica en su capacidad de generar evidencia relevante, contextualizada y pragmática, que sirve tanto para la toma de decisiones clínicas y de políticas de salud, como para sentar las bases de investigaciones más robustas.

Tabla 9. Fortalezas y limitaciones de los estudios cuasiexperimentales

Aspecto	Fortalezas	Limitaciones
Aplicabilidad	Reflejan la práctica clínica y los escenarios reales (hospitales públicos, programas comunitarios, políticas de salud).	Resultados menos controlados, con mayor influencia de factores externos.
Factibilidad	Son posibles cuando la aleatorización no es viable por razones éticas, logísticas o administrativas.	Mayor riesgo de sesgo de selección y confusión al no existir asignación aleatoria.
Rapidez y costo	Generalmente más rápidos y menos costosos que los ensayos clínicos aleatorizados.	Pueden tener menor poder estadístico si la muestra es pequeña o las pérdidas son altas.
Generación de hipótesis	Proveen evidencia preliminar para diseñar RCTs o guías clínicas.	No establecen causalidad con la misma solidez que los ensayos aleatorizados.
Impacto en políticas de salud	Permiten evaluar intervenciones poblacionales o cambios de sistema (ej., programas de rehabilitación, leyes de salud pública).	Sus conclusiones suelen estar muy ligadas al contexto, lo que limita la generalización.

Fuente: Elaboración propia

https://doi.org/10.62486/978-9915-9928-3-9.Ch15

11. PREGUNTAS DE AUTOEVALUACIÓN

PREGUNTAS DE SELECCIÓN MÚLTIPLE

1. ¿Cuál es la principal característica de un estudio cuasiexperimental?
a) No interviene el investigador
b) La asignación a grupos no es aleatoria
c) Solo estudia variables descriptivas
d) No necesita grupo comparador

2. ¿Qué sesgo es más frecuente en los estudios cuasiexperimentales con grupos no equivalentes?
a) Sesgo de selección
b) Sesgo de información
c) Sesgo de publicación
d) Sesgo de memoria

3. Si la diferencia de medias entre dos grupos es 5 puntos, con un IC95%: –1 a 11, la conclusión correctaes:
a) La intervención es efectiva
b) El resultado no es significativo
c) El resultado es clínicamente relevante
d) El estudio demuestra causalidad

4. En el análisis de resultados de un estudio cuasiexperimental, ¿qué significa que un RR = 1,5 con IC95%: 1,1–2,0?
a) La intervención aumenta un 50% la probabilidad del desenlace, con efecto significativo
b) El desenlace es igual en ambos grupos
c) No hay diferencia estadística entre grupos
d) El resultado es clínicamente irrelevante

Respuesta correcta: a.El IC no incluye el 1, por lo que es significativo; el RR de 1,5 indica un 50% más de probabilidad.

5. ¿Cuál de las siguientes es una utilidad práctica de los estudios cuasiexperimentales?
a) Reemplazar los ensayos clínicos aleatorizados en todas las situaciones
b) Evaluar intervenciones en contextos reales donde la aleatorización no es posible
c) Confirmar causalidad con el mismo nivel que un ensayo clínico
d) Evitar el uso de análisis estadísticos complejos

Respuesta correcta: b.Son valiosos en la práctica real y en políticas de salud, cuando los RCTs no son factibles.

PREGUNTAS ABIERTAS

6. Mencione dos ventajas y dos limitaciones de los estudios cuasiexperimentales. Posible respuesta:

7. Explique con sus palabras por qué es importante revisar las características basales de los grupos en un estudio cuasiexperimental.

8. ¿Cómo debe interpretarse un intervalo de confianza que cruza el valor nulo (0 en diferencias,1 en riesgos)?

9. Describa una situación clínica en su hospital o servicio donde un estudio cuasiexperimental sería más factible que un ensayo clínico aleatorizado.

10. ¿Por qué los análisis de sensibilidad fortalecen las conclusiones de un estudio cuasiexperimental?

RESPUESTAS:

1. b). La característica principal de un estudio cuasiexperimental es que la asignación a los grupos no se realiza de manera aleatoria. Esto implica que, aunque existe una intervención por parte del investigador, los grupos pueden diferir desde el inicio, lo que introduce riesgos metodológicos que deben ser considerados en la interpretación.

2. a). El sesgo más frecuente en los estudios cuasiexperimentales con grupos no equivalentes es el sesgo de selección. Este ocurre porque los grupos pueden estar formados por pacientes con características distintas desde el inicio, lo que puede influir directamente en los resultados

https://doi.org/10.62486/978-9915-9928-3-9.Ch15

observados.

3. b). Cuando el intervalo de confianza incluye el valor nulo, en este caso el 0, no se puede afirmar que exista una diferencia real entre los grupos. Por lo tanto, el resultado no es estadísticamente significativo y debe interpretarse con cautela.

4. a). Un riesgo relativo de 1,5 con un intervalo de confianza que no incluye el 1 indica que la intervención aumenta en un 50% la probabilidad del desenlace y que este efecto es estadísticamente significativo. Esto sugiere una asociación real entre la intervención y el resultado observado.

5. b). Los estudios cuasiexperimentales son especialmente útiles para evaluar intervenciones en contextos reales donde la aleatorización no es posible, como en programas hospitalarios o políticas de salud pública, permitiendo generar evidencia aplicable a la práctica cotidiana

6. Los estudios cuasiexperimentales presentan ventajas como su capacidad para reflejar la práctica clínica real y su utilidad en situaciones donde la aleatorización no es factible. Sin embargo, también tienen limitaciones importantes, como un mayor riesgo de sesgos y una menor capacidad para establecer relaciones causales con la misma solidez que un ensayo clínico aleatorizado.

7. Revisar las características basales es fundamental porque permite determinar si los grupos eran comparables al inicio del estudio. Si existen diferencias importantes entre ellos, los resultados pueden estar influenciados por esas desigualdades y no por la intervención en sí.

8. Un intervalo de confianza que cruza el valor nulo indica que no se puede descartar la ausencia de efecto. En el caso de diferencias de medias, si el intervalo incluye el 0, y en el caso de riesgos relativos u odds ratio si incluye el 1, el resultado no es estadísticamente significativo.

9. Un ejemplo de situación en la que un estudio cuasiexperimental sería más factible es la implementación de un nuevo protocolo de tratamiento en un hospital, donde todos los pacientes deben recibir la intervención por razones éticas o administrativas, lo que impide la asignación aleatoria.

10. Los análisis de sensibilidad fortalecen las conclusiones porque permiten evaluar si los resultados se mantienen consistentes bajo diferentes condiciones o supuestos. Si los hallazgos persisten al modificar el análisis, se incrementa la confianza en que el efecto observado es real y no producto de un artefacto metodológico.

https://doi.org/10.62486/978-9915-9928-3-9.Ch15

REFERENCIAS

Shadish WR, Cook TD, Campbell DT. Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin; 2002. Disponible en: https://books.google.com/books?id=o7jaAAAAMAAJ
Harris AD, McGregor JC, Perencevich EN, Furuno JP, Zhu J, Peterson DE, et al. The use and interpretation of quasi-experimental studies in medical informatics. J Am Med InformAssoc. 2006;13(1):16-23. https://doi.org/10.1197/jamia.m1749
Des Jarlais DC, Lyles C, Crepaz N; TREND Group. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. Am J Public Health. 2004;94(3):361-366. https://doi.org/10.2105/ajph.94.3.361
Craig P, Cooper C, Gunnell D, Haw S, Lawson K, Macintyre S, Ogilvie D, et al. Using natural experiments to evaluate population health interventions: new guidance. J Epidemiol Community Health. 2012;66(12):1182-1186. https://doi.org/10.1136/jech-2011-200375
Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355:i4919. https://doi.org/10.1136/bmj.i4919
von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med. 2007;4(10):e296. https://doi.org/10.1371/journal.pmed.0040296
Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ. 2014;348:g1687. https://doi.org/10.1136/bmj.g1687
Fuller T, Peters J, Pearson M, Anderson R. Evaluating the impact and use of Transparent Reporting of Evaluations with Non-randomised Designs (TREND) reporting guidelines. BMJ Open. 2012;2(6):e002073. doi:10.1136/bmjopen-2012-002073
Campbell DT, Stanley JC. Experimental and quasi-experimental designs for research. Chicago: Rand McNally; 1963.
Wong S, Lee K-P, Wong JY, Chan AW, Chan HY, Pang SM, et al. Applying difference-in-differences design in quality improvement and health systems research. J Am Geriatr Soc. 2025;73(4):1035–44. Available from: https://doi.org/10.1111/jgs.19180.
Alsaggaf R, O'Hara LM, Stafford KA, Leekha S, Harris AD. Quasi-experimental studies in the fields of infection control and antibiotic resistance, ten years later: A systematic review. Infect Control Hosp Epidemiol. 2018;39(2):170–6. Available at: https://doi.org/10.1017/ice.2017.296.
Huang HL, Chou CP, Leu S, et al. Effects of a quasi-experimental study of using flipped classroom approach to teach evidence-based medicine to medical technology students. BMC Med Educ. 2020;20(1):31. Available at: https://doi.org/10.1186/s12909-020-1946-7.
Couto LA, Nieuwlaat R, Nolan R, et al. A quasi-experimental evaluation of a primary care behavioral health integration program based on the chronic care model. J Gen Intern Med. 2025. Available at: https://doi.org/10.1007/s11606-025-09641-0.
Caniglia EC, Murray EJ. Difference-in-difference in the time of cholera: A gentle introduction for epidemiologists. Curr Epidemiol Rep. 2020;7(4):203–11. Available at: https://doi.org/10.1007/s40471-020-00245-2.

Chapter 16 / Capítulo 16

Critical reading of scientific articles on health: A practical guide (Spanish Edition)
ISBN: 978-9915-9928-3-9
DOI: 10.62486/978-9915-9928-3-9.Ch16
Pages: 263-298

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Systematic Reviews and Meta-Analyses

Revisiones Sistemáticas y Metaanálisis

José Bernardo Antezana-Muñoz
Adolfo Israel Vasquez Cuellar
Freddy Ednildon Bautista-Vanegas
Eloy Paycho Anagua

1 Centro de Alta Complejidad Cardiovascular Presidente Juan Domingo Perón. Argentina
2 Hospital Materno Infantil, Caja Nacional de Salud (CNS). La Paz, Bolivia.
3 Neurokliniken Beelitz GmbH Neurologische Rehabilitationsklinik: Beelitz Heilstätten, Brandenburg, DE. Germany.
4 Neurologisches Fachkrankenhaus für Bewegungsstörungen/Parkinson: Beelitz Heilstätten, Brandenburg, DE. Germany.
5 Hospital de Tercer Nivel Dr. Hernán Messuti Ribera. Pando, Bolivia.

ABSTRACT

Introduction: Systematic reviews and meta-analyses represent the highest level of evidence in Evidence-Based Medicine, rigorously, transparently, and reproducibly integrating the results of multiple primary studies to answer specific clinical questions. Their application reduces selection bias, increases the precision of estimates, and strengthens clinical and healthcare decision-making. Development: This chapter describes the essential methodological steps for the development and critical interpretation of systematic reviews and meta-analyses, including formulating the PICO question, registering the protocol, conducting a comprehensive literature search, selecting and extracting data, assessing the risk of bias, performing qualitative and quantitative synthesis, analyzing heterogeneity, addressing publication bias, and evaluating the certainty of evidence using GRADE. It also presents tools for interpreting forest plots, funnel plots, and network meta-analyses, highlighting their usefulness in generating robust and clinically applicable evidence. Conclusions: Systematic reviews and meta-analyses are indispensable tools for transforming scientific evidence into well-founded clinical decisions. However, its true value depends on methodological rigor, critical appraisal of study quality, and the proper interpretation of the magnitude and certainty of the effects. A thoughtful reading allows us to distinguish between statistical significance and clinical relevance, promoting safer, more efficient, and patient-centered healthcare practices.

Keywords: Systematic Review; Meta-analysis; Evidence-Based Medicine; Publication Bias; GRADE System

RESUMEN

Introducción: Las revisiones sistemáticas y los metaanálisis representan el nivel más alto de evidencia en la Medicina Basada en la Evidencia, al integrar de forma rigurosa, transparente y reproducible los resultados de múltiples estudios primarios para responder preguntas clínicas específicas. Su aplicación reduce el sesgo de selección, incrementa la precisión de las estimaciones y fortalece la toma de decisiones clínicas y sanitarias. Desarrollo: El capítulo describe las etapas metodológicas esenciales para la elaboración e interpretación crítica de revisiones sistemáticas y metaanálisis, incluyendo la formulación de la pregunta PICO, el registro del protocolo, la búsqueda bibliográfica exhaustiva, la selección y extracción de datos, la evaluación del riesgo de sesgo, la síntesis cualitativa y cuantitativa, el análisis de heterogeneidad, el sesgo de publicación y la valoración de la certeza de la evidencia mediante GRADE. De esta manera, expone herramientas para interpretar forest plots, funnel plots y metaanálisis en red, destacando su utilidad para generar evidencia robusta y clínicamente aplicable. Conclusiones: Las revisiones sistemáticas y los metaanálisis constituyen herramientas indispensables para transformar la evidencia científica en decisiones clínicas fundamentadas. No obstante, su verdadero valor depende de la rigurosidad metodológica, la evaluación crítica de la calidad de los estudios y la adecuada interpretación de la magnitud y certeza de los efectos. Una lectura reflexiva permite distinguir entre significancia estadística y relevancia clínica, favoreciendo una práctica asistencial más segura, eficiente y centrada en el paciente.

Palabras clave: Revisión Sistemática; Metaanálisis; Medicina Basada en la Evidencia; Sesgo de Publicación; Sistema GRADE

1. CASO CLÍNICO

Un médico especialista en Medicina Interna atiende a un paciente de 55 años con diagnóstico de diabetes mellitus tipo 2 y obesidad. Durante la consulta, el paciente refiere haber consultado diversas fuentes de información científica y de divulgación médica, en las cuales se señala que los inhibidores del cotransportador sodio-glucosa tipo 2 (iSGLT2) no solo mejoran el control glucémico, sino que también reducen la mortalidad cardiovascular, las hospitalizaciones por insuficiencia cardíaca y otros desenlaces cardiovasculares mayores. Ante esta información, plantea una pregunta clínicamente relevante: ¿la evidencia disponible demuestra que estos beneficios son superiores a los obtenidos con otros antidiabéticos orales o con placebo?

El médico reconoce que, durante la última década, múltiples ensayos clínicos aleatorizados de gran tamaño entre ellos EMPA-REG OUTCOME, CANVAS Program, DECLARE–TIMI 58 y VERTIS-CV han evaluado la eficacia y seguridad cardiovascular de los iSGLT2 en diferentes poblaciones de pacientes con diabetes mellitus tipo 2. Sin embargo, también advierte que, aunque los estudios muestran una tendencia consistente hacia el beneficio, existen diferencias en la magnitud del efecto, la significancia estadística de algunos desenlaces, las características basales de las poblaciones incluidas y los criterios metodológicos empleados. Estas discrepancias dificultan establecer una conclusión definitiva a partir de un único ensayo clínico.

Frente a este escenario, surgen varias interrogantes propias de la práctica de la Medicina Basada en la Evidencia:

- ¿Cómo puede integrarse de manera rigurosa la totalidad de la evidencia disponible para responder esta pregunta clínica?
- ¿Qué importancia relativa debe otorgarse a cada ensayo clínico considerando sus diferencias metodológicas, tamaño muestral y riesgo de sesgo?
- ¿Es posible obtener una estimación global del efecto terapéutico que incremente la precisión de los resultados y reduzca la incertidumbre inherente a los estudios individuales?
- ¿Cuál es el nivel de certeza de la evidencia que sustenta la recomendación clínica?

Para responder estas interrogantes, el médico decide recurrir a una revisión sistemática con metaanálisis, considerada el diseño de investigación secundaria con mayor nivel de evidencia para sintetizar los resultados de múltiples estudios primarios. Mediante una metodología estructurada, transparente y reproducible, este enfoque permite identificar, seleccionar, evaluar críticamente y combinar cuantitativamente la evidencia disponible, disminuyendo el riesgo de sesgos, aumentando la precisión de las estimaciones del efecto y proporcionando una base más sólida para la toma de decisiones clínicas.

Como primer paso, formula la pregunta de investigación utilizando el modelo PICO, ampliamente recomendado para estructurar preguntas clínicas de intervención:

- P (Población): adultos con diabetes mellitus tipo 2.
- I (Intervención): tratamiento con inhibidores del cotransportador sodio-glucosa tipo 2 (iSGLT2).
- C (Comparación): placebo u otros fármacos antidiabéticos.
- (Outcomes o desenlaces): mortalidad cardiovascular, hospitalización por insuficiencia cardíaca y eventos cardiovasculares mayores (Major Adverse Cardiovascular Events, MACE).

Con base en esta formulación, el internista plantea la hipótesis de que una revisión sistemática acompañada de un metaanálisis de ensayos clínicos aleatorizados permitirá obtener una estimación global más precisa del efecto de los iSGLT2 sobre los principales desenlaces cardiovasculares, establecer la consistencia de la evidencia disponible y valorar el grado de certeza de los resultados, proporcionando así un fundamento científico robusto para orientar la toma de decisiones en la práctica clínica basada en la evidencia.

Tabla 1. Comparación entre ensayo clínico individual, revisión sistemática y metaanálisis

Característica	Ensayo clínico individual (ECA)	Revisión sistemática	Metaanálisis
Tipo de estudio	Investigación primaria, prospectiva, que evalúa la eficacia o seguridad de una intervención en un grupo específico de pacientes.	Investigación secundaria, que recopila, evalúa y sintetiza de forma estructurada toda la evidencia disponible sobre una pregunta clínica.	Técnica estadística que forma parte de una revisión sistemática y combina cuantitativamente los resultados de múltiples estudios.
Objetivo principal	Determinar si una intervención es eficaz y segura en comparación con un control (placebo u otro tratamiento).	Ofrecer una visión integral y crítica de la evidencia, reduciendo sesgos mediante criterios predefinidos.	Incrementar la precisión del estimador de efecto y explorar la consistencia entre estudios.
Unidad de análisis	Pacientes o sujetos incluidos en el estudio.	Estudios publicados y no publicados que cumplen con criterios de inclusión.	Resultados cuantitativos de los estudios individuales (ej. riesgo relativo, odds ratio, hazard ratio).

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

Característica	Ensayo clínico individual (ECA)	Revisión sistemática	Metaanálisis
Nivel de evidencia	Alto, pero limitado a la población y condiciones del estudio.	Superior, al integrar múltiples estudios y evaluar su calidad metodológica.	Muy alto, siempre que los estudios incluidos sean de buena calidad y suficientemente homogéneos.
Fortalezas	Control riguroso de variables, aleatorización y seguimiento prospectivo.	Minimiza el sesgo de selección y ofrece síntesis estructurada de la literatura.	Aumenta la potencia estadística, permite detectar efectos pequeños y realizar análisis de subgrupos.
Limitaciones	Puede no ser generalizable; resultados aislados y a veces contradictorios.	Depende de la calidad de los estudios incluidos y del rigor metodológico.	Sensible a la heterogeneidad y al sesgo de publicación; no corrige debilidades de los estudios primarios.
Aplicación clínica	Informa decisiones en contextos específicos, pero no siempre extrapolables.	Informa guías clínicas, políticas de salud y decisiones basadas en el conjunto de la evidencia.	Permite cuantificar el beneficio o riesgo promedio y explorar la consistencia de los hallazgos.

Fuente: Elaboración propia.

2. OBJETIVOS DE UNA REVISIÓN SISTEMÁTICA Y METAANÁLISIS

Los objetivos principales pueden resumirse en los siguientes ejes:

Sintetizar de manera estructurada y transparente la evidencia disponible: Una revisión sistemática recoge, evalúa y organiza la información publicada (y no publicada) sobre una pregunta clínica específica.El objetivo es reducir el sesgo, evitando que la selección de estudios dependa de la preferencia del investigador.

Cuantificar el efecto de una intervención o exposición: El metaanálisis combina estadísticamente los resultados de los estudios incluidos.Esto permite obtener una estimación puntual y su intervalo de confianza, que representa de manera más precisa el efecto global que cualquiera de los estudios individuales.

Incrementar la potencia estadística y la precisión de los hallazgos: Al unir los datos de múltiples ensayos, se aumenta la capacidad de detectar efectos pequeños, pero clínicamente relevantes. Esto es especialmente útil cuando los estudios individuales tienen muestras pequeñas o resultados aparentemente contradictorios.Explorar la heterogeneidad entre estudios: Un metaanálisis no solo ofrece un “resultado promedio”, sino que permite analizar si los efectos difieren entre subgrupos (ej., hombres vs. mujeres, pacientes jóvenes vs. mayores, poblaciones con distintas comorbilidades).De esta manera, ayuda a comprender qué intervenciones funcionan mejor en qué contextos.

Evaluar la calidad y certeza de la evidencia: Una revisión sistemática rigurosa incorpora herramientas como GRADE o AMSTAR 2, que valoran la calidad de los estudios y la confianza en los resultados.El objetivo es que las conclusiones no solo sean claras, sino también fiables y aplicables a la práctica clínica.Identificar vacíos de conocimiento e impulsar nuevas investigaciones: Al mapear lo que existe y lo que falta, las revisiones sistemáticas orientan a los investigadores hacia preguntas aún no resueltas o hacia áreas donde los resultados siguen siendo inciertos.

3. INTRODUCCIÓN

En la práctica clínica diaria, los profesionales de la salud se enfrentan constantemente a preguntas que requieren respuestas fundamentadas en la mejor evidencia disponible. Sin embargo, la literatura científica crece de manera exponencial: cada año se publican miles de ensayos clínicos, estudios observacionales y reportes en distintas áreas de la medicina. Este volumen de información hace que sea difícil, e incluso imposible, para un médico individual leer y analizar críticamente todos los estudios relevantes sobre un tema específico.

En este contexto, las revisiones sistemáticas y los metaanálisis se han convertido en herramientas fundamentales para la medicina basada en la evidencia (MBE). Una revisión sistemática permite reunir, evaluar y sintetizar, de forma estructurada y reproducible, todos los estudios pertinentes a una pregunta clínica claramente definida. De esta manera, se busca minimizar el sesgo y proporcionar una visión integral de la evidencia.

El metaanálisis, por su parte, aplica técnicas estadísticas que permiten combinar cuantitativamente los resultados de los estudios individuales. Esto no solo aumenta la precisión de las estimaciones del efecto de una intervención o exposición, sino que también mejora la capacidad de detectar diferencias sutiles que podrían pasar inadvertidas en estudios aislados.

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

La importancia de estas metodologías trasciende la investigación académica: sus conclusiones constituyen la base de las guías de práctica clínica, orientan políticas de salud pública y permiten a médicos, pacientes y gestores tomar decisiones informadas. Al mismo tiempo, identifican áreas donde la evidencia es débil o insuficiente, señalando así la necesidad de nuevas investigaciones.

A lo largo de este capítulo se explorarán en detalle los fundamentos, pasos metodológicos, herramientas estadísticas, fuentes de sesgo y aplicaciones clínicas de las revisiones sistemáticas y metaanálisis. El objetivo es que el lector no solo comprenda cómo interpretar sus resultados, sino que también adquiera la capacidad crítica de valorar su calidad y aplicabilidad en la práctica diaria.

4. METODOLOGÍA

La metodología de una revisión sistemática con o sin metaanálisis sigue un proceso riguroso, diseñado para reducir el sesgo y garantizar que las conclusiones sean reproducibles y útiles en la práctica clínica. Aunque existen variantes según el tipo de revisión, los pasos fundamentales pueden resumirse en los siguientes:

4.1 DEFINICIÓN DE LA PREGUNTA DE INVESTIGACIÓN

Al analizar críticamente una revisión sistemática, el primer aspecto que debe evaluarse es si la pregunta de investigación está claramente definida y estructurada. Este punto inicial orienta toda la revisión: qué estudios se buscan, cuáles se incluyen y cómo se interpretan los hallazgos.

Las revisiones de calidad suelen utilizar el modelo PICO (Population, Intervention, Comparator, Outcome), ampliamente recomendado por el Cochrane Handbook y las guías PRISMA 2020. Una pregunta PICO bien planteada asegura que la revisión sea específica, reproducible y clínicamente relevante.

Aspectos que debe valorar el lector crítico:

- Claridad: ¿La población, la intervención, el comparador y los desenlaces están bien delimitados?
- Relevancia clínica: ¿Los desenlaces son realmente importantes para la práctica (ej. mortalidad, calidad de vida) o se limitan a variables intermedias?
- Pertinencia: ¿La pregunta responde a una necesidad clínica concreta o es demasiado amplia/difusa?
- Adecuación del diseño: ¿El modelo PICO u otra variante (PICOTS, PEO) se ajusta al tipo de revisión (tratamiento, diagnóstico, pronóstico, cualitativa)?

Para nuestro ejemplo: Un paciente con diabetes mellitus tipo 2 consulta a su médico sobre si los inhibidores de SGLT2 realmente reducen la mortalidad cardiovascular en comparación con otros antidiabéticos. Al buscar evidencia, el médico encuentra una revisión sistemática con metaanálisis que formula su pregunta así:

- P (Población): Adultos con diabetes tipo 2.
- I (Intervención): Tratamiento con inhibidores de SGLT2.
- C (Comparador): Placebo o antidiabéticos orales distintos a iSGLT2.
- (Outcomes): Mortalidad cardiovascular, hospitalización por insuficiencia cardíaca y eventos cardiovasculares mayores (MACE).

El lector crítico debe preguntarse: ¿Está claramente definida la población (ej. pacientes con y sin enfermedad cardiovascular previa) ?, ¿Los desenlaces son clínicamente relevantes o solo incluyen variables intermedias como niveles de glucosa?, ¿La comparación es pertinente para la práctica clínica local? De esta manera, la lectura crítica de la pregunta de investigación permite valorar desde el inicio si la revisión sistemática está bien enfocada y si sus resultados serán aplicables a la toma de decisiones clínicas.

4.2 ELABORACIÓN Y REGISTRO DEL PROTOCOLO

Un criterio fundamental al leer críticamente una revisión sistemática es comprobar si los autores elaboraron y registraron un protocolo previo. Este documento es la garantía de que la revisión se desarrolló con transparencia y rigor metodológico, ya que evita cambios arbitrarios en la pregunta, en los desenlaces o en los métodos de análisis una vez conocidos los resultados. La ausencia de un protocolo suele asociarse con mayor riesgo de sesgo de reporte.

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

Las guías PRISMA-P y el Cochrane Handbook recomiendan que un protocolo incluya, de forma detallada, la pregunta de investigación, los criterios de inclusión y exclusión de estudios, la estrategia de búsqueda, los métodos para la extracción de datos, la forma de evaluar el riesgo de sesgo y el plan de análisis estadístico. Asimismo, debe declarar potenciales conflictos de interés y describir cómo se resolverán las discrepancias entre revisores.

Existen plataformas reconocidas para su registro, entre ellas PROSPERO (la más utilizada en salud), Open Science Framework (OSF) y los protocolos de revisiones Cochrane. Al leer un artículo, el crítico debe verificar si aparece un número de registro y si lo publicado coincide con lo planificado.

Por ejemplo, en el caso del paciente con diabetes tipo 2 que consulta sobre los inhibidores de SGLT2, supongamos que el médico revisa un metaanálisis registrado en PROSPERO (N° CRD42021234567). Al compararlo con el artículo final, observa que los desenlaces principales fueron respetados (mortalidad cardiovascular y hospitalización por insuficiencia cardíaca), pero también identifica que se añadió como desenlace primario la reducción de hemoglobina glicosilada, originalmente listado como secundario. Esta modificación puede interpretarse como un cambio post hoc, lo que reduce la confianza en la objetividad de los resultados.

4.3 ESTRATEGIA DE BÚSQUEDA

Otro aspecto clave en la lectura crítica de una revisión sistemática es la estrategia de búsqueda utilizada para identificar los estudios. Una revisión es tan sólida como lo sea su búsqueda: si esta es incompleta o sesgada, los resultados pueden ser engañosos.

Una estrategia rigurosa debe ser amplia, transparente y reproducible. Esto implica detallar las bases de datos consultadas, los términos de búsqueda empleados y los filtros aplicados. Las principales bases utilizadas en salud son PubMed/MEDLINE, Cochrane Library, Embase, Scopus y Web ofScience, aunque en Latinoamérica adquieren gran relevancia SciELO y la Biblioteca Virtual en Salud (BVS). Además, se recomienda explorar la literatura gris (tesis, actas de congresos, preprints, registros de ensayos como ClinicalTrials.gov) para reducir el sesgo de publicación.

El uso de términos MeSH (Medical Subject Headings) y DeCS (Descriptores en Ciencias de la Salud), combinados con operadores booleanos (AND, OR, NOT), asegura que la búsqueda capture con precisión la mayor parte de los estudios relevantes. Por ejemplo, un metaanálisis sobre inhibidores de SGLT2 debería mostrar en su sección de métodos las palabras clave utilizadas ("Sodium-GlucoseTransporter 2 Inhibitors", "Diabetes Mellitus, Type 2", "Cardiovascular Outcomes"), así como los límites temporales o de idioma aplicados.

Desde la perspectiva del lector crítico, hay preguntas indispensables: ¿se explicaron claramente las bases de datos consultadas?, ¿se incluyó literatura en distintos idiomas o solo en inglés?, ¿se aportó la estrategia completa de búsqueda como anexo para permitir su reproducción?, ¿se consideraron estudios no publicados o literatura gris? Una revisión que no aclare estos aspectos puede estar sesgada hacia estudios positivos y sobrestimar los beneficios de una intervención.

Volviendo al caso del paciente con diabetes tipo 2, si el médico encuentra una revisión sobre inhibidores de SGLT2 que solo buscó en PubMed y en inglés, debería desconfiar de su exhaustividad. Es posible que estudios relevantes publicados en otras bases o en otros idiomas por ejemplo, en población latinoamericana— hayan quedado fuera, limitando la aplicabilidad de los resultados a su contexto clínico, la forma correcta debería ser descrita de esta forma:

- "Se realizó una búsqueda sistemática en PubMed/MEDLINE, Embase, Cochrane Central Register ofControlledTrials (CENTRAL), Scopus y Web ofScience hasta marzo de 2023. Se utilizaron términos MeSH y DeCS relacionados con 'Diabetes Mellitus, Type 2', 'Sodium-GlucoseTransporter 2 Inhibitors', 'Cardiovascular Outcomes', combinados con operadores booleanos. No se aplicaron restricciones de idioma. Además, se consultaron registros de ensayos clínicos (ClinicalTrials.gov) y literatura gris.

4.4 SELECCIÓN DE ESTUDIOS

Una vez realizada la búsqueda bibliográfica, el siguiente paso en una revisión sistemática es la selección de los estudios que serán incluidos en el análisis. Este proceso es crucial porque determina la calidad y la relevancia de la evidencia que finalmente se sintetizará.

Desde la perspectiva de la lectura crítica, lo importante es verificar si los autores describen

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

con claridad:

Criterios de inclusión y exclusión: deben estar definidos antes de comenzar la selección. Estos criterios suelen basarse en la población, la intervención, el comparador, los desenlaces y el diseño de estudio (por ejemplo, ensayos clínicos aleatorizados para preguntas de eficacia).

Proceso de selección: idealmente, dos revisores deberían evaluar de forma independiente los títulos y resúmenes, para luego revisar el texto completo. Las discrepancias deben resolverse mediante consenso o con un tercer revisor.

Transparencia: la revisión debe mostrar cuántos estudios se identificaron inicialmente, cuántos se excluyeron y por qué motivos. Esto se representa en un diagrama de flujo PRISMA, que facilita al lector entender el proceso.

Si una revisión no especifica los criterios de selección o no explica cómo se resolvieron las discrepancias entre revisores, su validez metodológica se ve comprometida. Del mismo modo, si se excluyeron estudios relevantes sin justificación clara, existe un riesgo de sesgo de selección.

En nuestro ejemplo aplicado al caso clínico, se describe lo siguiente en su metodología:

- "Se identificaron 2.345 referencias tras la búsqueda inicial. Después de eliminar duplicados, 1.876 estudios fueron evaluados por título y resumen. De ellos, 1.820 fueron excluidos por no cumplir los criterios de inclusión (estudios en animales, revisiones narrativas, reportes de casos, población sin diabetes tipo 2). Un total de 56 artículos fueron evaluados en texto completo, y 10 ensayos clínicos aleatorizados cumplieron con los criterios para la síntesis cuantitativa. La selección fue realizada por dos revisores de manera independiente; las discrepancias se resolvieron mediante consenso. El proceso completo se muestra en el diagrama PRISMA del anexo 2."

Un lector crítico que revisa esta sección puede valorar positivamente la claridad y transparencia de la selección. Sin embargo, si el artículo no incluyera los motivos de exclusión o no presentara un diagrama PRISMA, la confianza en la exhaustividad de la revisión sería menor.

4.5 EXTRACCIÓN DE DATOS

Una vez seleccionados los estudios, el paso siguiente en una revisión sistemática es la extracción de datos. Esta etapa consiste en trasladar de cada artículo los elementos esenciales que permitan compararlos y, posteriormente, sintetizarlos en un análisis narrativo o cuantitativo.

Desde la perspectiva de la lectura crítica, lo importante no es memorizar cómo se llena una tabla de extracción, sino saber identificar si el artículo lo hizo de manera rigurosa y transparente. Una revisión sistemática de calidad siempre presenta una plantilla estandarizada —a veces como tabla en el texto principal, otras en material suplementario— donde resume las características de cada estudio incluido.La información mínima que se espera encontrar comprende:

Características de la población: edad promedio, proporción de hombres y mujeres, comorbilidades, criterios de inclusión y exclusión.

Detalles de la intervención y comparadores: fármaco, dosis, duración del tratamiento o tipo de exposición analizada.

Desenlaces primarios y secundarios: aquellos que guían la pregunta de investigación (mortalidad, eventos cardiovasculares) y los adicionales (biomarcadores, variables intermedias).

Diseño y tamaño muestral: número de participantes, tipo de estudio (ECA, cohorte, etc.) y duración del seguimiento.

Resultados principales: medidas de efecto (RR, OR, HR, medias con desviación estándar) con sus intervalos de confianza.

Riesgo de sesgo y calidad metodológica: evaluación realizada con herramientas reconocidas (RoB 2, ROBINS-I, QUADAS-2, etc.).

Cuando un artículo no detalla esta información, el lector debe ser cauteloso: la falta de transparencia en la extracción de datos puede indicar riesgo de sesgo o incluso errores en la síntesis.

Ejemplo aplicado al caso clínico:

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

Tabla 2. Características de los estudios incluidos en el metaanálisis sobre inhibidores de SGLT2 en diabetes tipo 2

Estudio (año)	Población	Intervención / Comparador	Desenlaces principales	N (seguimiento)	Resultados	Riesgo de sesgo
Zheng et al., JAMA Cardiol 2020 PMID: 33031522	46.969 pacientes con DM2; edad media 63.7 años; 65.9 % hombres; 66 % con enfermedad CV establecida	iSGLT2 (empagliflozina, dapagliflozina, canagliflozina) vs placebo	Mortalidad CV, MACE, hospitalización por IC, desenlaces renales compuestos	6 ECA; 2-4 años	HR MACE 0.90 (IC95 % 0.85–0.95); HR hospitalización por IC/muerte CV 0.78; HR renal 0.62	Bajo (RoB 2)
Castellana et al., Cardiovasc Diabetol 2020 DOI: 10.1186/s12933-020-01067-8	1.703.519 pacientes de múltiples cohortes; edad media variable; proporción de CV previa dependiente de cohorte	Evaluación de elegibilidad de pacientes para criterios de EMPA-REG, CANVAS, DECLARE-TIMI 58 y VERTIS-CV	Proporción de pacientes que cumplirían criterios de inclusión en CVOTs	5 estudios de cohortes	Elegibilidad: EMPA-REG 17 %, CANVAS 36.4 %, DECLARE 49.5 %, VERTIS-CV 19 %	Moderado (heterogeneidad alta)

Fuente: Elaboración propia.

Cómo interpretarla críticamente en nuestro caso clínico:

- Ver si los artículos incluidos muestran todos los elementos clave (como en la tabla). Si falta alguno (por ejemplo, duración de seguimiento, desenlace renal, subgrupos), eso disminuye la aplicabilidad de los resultados al paciente real.
- Observar si los desenlaces secundarios son relevantes (por ejemplo, reducción de azúcar, peso) pero no sobredimensionarlos frente a los desenlaces clínicos críticos (muerte cardiovascular, hospitalización por ICC).
- Ver el riesgo de sesgo: si los estudios son ensayos aleatorizados, placebo-control, con buen seguimiento, poco desenmascaramiento, y con criterios bien definidos; si no, aplicar reservas al interpretar los resultados.
- El cuadro también muestra que los estudios de mayor tamaño (ej. EMPA-REG, CANVAS) aportan más peso estadístico en el metaanálisis, y que el riesgo de sesgo global fue bajo según la herramienta RoB 2. Para el lector crítico, este nivel de detalle confirma que la extracción de datos fue meticulosa, transparente y útil para interpretar la síntesis de resultados.

4.6 EVALUACIÓN DEL RIESGO DE SESGO Y CALIDAD DE LOS ESTUDIOS

Una revisión sistemática es tan sólida como la calidad de los estudios que la conforman. Por eso, un aspecto esencial en la lectura crítica es comprobar si los autores evaluaron de manera rigurosa el riesgo de sesgo y la calidad metodológica de cada estudio incluido.

El sesgo se entiende como cualquier desviación sistemática que pueda distorsionar la estimación real del efecto de una intervención. La presencia de sesgo en los estudios primarios compromete directamente la validez de los resultados del metaanálisis. Para evitar interpretaciones erróneas, los revisores deben utilizar herramientas estandarizadas, distintas según el tipo de estudio:

- **RoB 2.0 (Risk of Bias 2.0):** para ensayos clínicos aleatorizados. Evalúa aspectos como la generación de la secuencia aleatoria, la ocultación de la asignación, el cegamiento, la integridad de los datos y el reporte selectivo.
- **ROBINS-I (Risk Of Bias In Non-randomizedStudiesofInterventions):** para estudios no aleatorizados, analiza confusión, selección de participantes, clasificación de intervenciones y medición de desenlaces.
- **QUADAS-2:** diseñado para estudios de pruebas diagnósticas, evalúa riesgo de sesgo y aplicabilidad en cuatro dominios (selección de pacientes, prueba índice, estándar de referencia, flujo de pacientes).
- **AMSTAR 2:** orientado a evaluar la calidad de revisiones sistemáticas publicadas, considerando aspectos como la estrategia de búsqueda, la duplicación en selección y extracción, y la evaluación de sesgo en los estudios primarios.

Además de estas herramientas, la certeza global de la evidencia debe valorarse con el sistema GRADE (GradingofRecommendations, Assessment, Development and Evaluation). Este enfoque permite calificar la evidencia por cada desenlace en cuatro niveles (alta, moderada, baja, muy baja), considerando factores como limitaciones metodológicas, inconsistencia,

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

indirectitud, imprecisión y sesgo de publicación. Los resultados suelen presentarse en una tabla SummaryofFindings (SoF), que facilita al lector identificar de manera rápida la calidad de la evidencia y la magnitud del efecto.

Ejemplo aplicado al caso clínico:

Tabla 3. Evaluación del riesgo de sesgo (RoB 2.0) en los principales ensayos clínicos de inhibidores de SGLT2 en diabetes tipo 2

Estudio	Aleatorización	Desviaciones de la intervención	Datos incompletos	Medición de desenlaces	Reporte selectivo	Juicio global
EMPA-REG OUTCOME	Bajo riesgo	Bajo riesgo	Bajo riesgo	Bajo riesgo	Bajo riesgo	Bajo riesgo
CANVAS Program	Bajo riesgo	Algunas preocupaciones	Bajo riesgo	Bajo riesgo	Bajo riesgo	Algunas preocupaciones
DECLARE–TIMI 58	Bajo riesgo	Bajo riesgo	Bajo riesgo	Bajo riesgo	Bajo riesgo	Bajo riesgo

Fuente: Elaboración propia.

Su descripción tendría que ser: "En la revisión sistemática que consulta el médico para responder al paciente con diabetes tipo 2, los autores evaluaron los diez ensayos clínicos incluidos mediante la herramienta RoB 2.0. El resultado fue que la mayoría presentaban bajo riesgo de sesgo, aunque dos estudios mostraron cierta preocupación por la falta de cegamiento de participantes. En el análisis con GRADE, la certeza de la evidencia para el desenlace "mortalidad cardiovascular" fue calificada como alta, mientras que para "hospitalización por insuficiencia cardíaca" se consideró moderada debido a la heterogeneidad entre estudios".

Para el lector crítico, estos hallazgos permiten interpretar con mayor confianza los resultados del metaanálisis. Una revisión que no aplica herramientas de evaluación de sesgo ni califica la certeza de la evidencia debe considerarse metodológicamente débil, por más sofisticado que sea su análisis estadístico.

4.7 SÍNTESIS DE LA EVIDENCIA

El propósito final de una revisión sistemática es integrar y resumir los hallazgos de los estudios incluidos para responder de manera clara a la pregunta de investigación. Esta síntesis puede adoptar dos formas principales: narrativa o cuantitativa, cada una con sus ventajas y limitaciones.

A) SÍNTESIS NARRATIVA (CUALITATIVA)

Se utiliza cuando los estudios incluidos son demasiado heterogéneos (distintos) entre sí o cuando los desenlaces no son comparables de manera directa. La síntesis narrativa no combina los resultados estadísticamente, sino que describe las tendencias generales, las fortalezas de los estudios, las limitaciones metodológicas y las brechas de conocimiento.

En la lectura crítica, el lector debe valorar si los autores presentan una narrativa equilibrada, sin exagerar resultados positivos ni minimizar los negativos. Además, debe revisar si se discuten de forma explícita las causas de la heterogeneidad (poblaciones distintas, intervenciones con variaciones de dosis, desenlaces definidos de manera desigual).

Este formato es habitual en revisiones narrativas o en scopingreviews, cuyo objetivo es ofrecer un panorama amplio más que una estimación precisa del efecto.

B) METAANÁLISIS (SÍNTESIS CUANTITATIVA)

Cuando los estudios son lo suficientemente homogéneos en diseño, población, intervención y desenlaces, es posible realizar un metaanálisis, que combina estadísticamente los resultados individuales para obtener un estimador global.

Existen dos modelos principales:

- Modelo de efectos fijos: asume que todos los estudios estiman el mismo efecto verdadero, y las diferencias entre ellos se deben únicamente al azar.
- Modelo de efectos aleatorios: considera que los estudios pueden tener efectos verdaderos distintos, por lo que incorpora la variabilidad entre ellos. Este suele ser el modelo más

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

realista en medicina, donde rara vez los estudios son completamente idénticos.

El metaanálisis permite calcular medidas de efecto como riesgo relativo (RR), odds ratio (OR), hazard ratio (HR), diferencia de medias (MD) o diferencia de medias estandarizada (SMD). Los resultados se presentan con su intervalo de confianza, y la forma gráfica clásica de mostrarlos es el forestplot, que resume visualmente los resultados de cada estudio y la estimación combinada.

Desde la perspectiva crítica, el lector debe preguntarse:¿Los estudios eran lo suficientemente homogéneos para ser combinados?, ¿El modelo elegido (fijo o aleatorio) fue el más apropiado?, ¿El forestplot muestra resultados consistentes o hay heterogeneidad marcada?, ¿Los autores interpretan los resultados con cautela, o presentan la síntesis como una verdad absoluta?

4.8 ANÁLISIS DE HETEROGENEIDAD Y SESGOS

Al interpretar un metaanálisis, el lector crítico debe recordar que no todos los estudios incluidos son idénticos. Variaciones en la población, la intervención, la forma de medir los desenlaces o la calidad metodológica pueden generar heterogeneidad. Asimismo, la publicación selectiva de resultados positivos puede introducir sesgos de publicación que distorsionen las conclusiones.

Heterogeneidad: La heterogeneidad mide hasta qué punto los resultados de los estudios difieren más de lo esperado por el azar. Una revisión rigurosa debe reportar:

- Prueba de Cochran-Q: evalúa si la variabilidad observada es mayor a la esperada por azar. Su valor p indica significancia, aunque es sensible al número de estudios.
- Estadístico I^2: expresa en porcentaje la proporción de variabilidad atribuible a heterogeneidad real
- Estrategias de exploración: análisis de subgrupos y metarregresión, que permiten investigar si la variabilidad se explica por factores como edad, presencia de comorbilidades, duración del seguimiento o dosis de intervención.

Desde la perspectiva crítica, el lector debe revisar si la heterogeneidad fue cuantificada, si se interpretó con cautela y si se justificó adecuadamente el uso de modelos de efectos fijos o aleatorios.

Sesgo de publicación: El sesgo de publicación ocurre cuando los estudios con resultados positivos tienen más probabilidad de ser publicados, mientras que los negativos o neutros permanecen inéditos. Este fenómeno puede llevar a que un metaanálisis sobrestime el efecto de una intervención.Herramientas para detectarlo:

- Gráfico en embudo (funnelplot): representa el tamaño del efecto frente a la precisión del estudio. Un gráfico simétrico sugiere ausencia de sesgo; la asimetría puede indicar estudios faltantes.
- Prueba de Egger: análisis estadístico que cuantifica la asimetría del funnelplot; un $p < 0,05$ sugiere sesgo de publicación.
- Método trim and fill: estima cuántos estudios faltan debido a sesgo de publicación y recalcula la estimación global ajustada.

En lectura crítica, es fundamental revisar si el artículo aplicó estos métodos, cuántos estudios incluyó (el funnelplot y la prueba de Egger no son confiables si hay menos de 10 estudios), y si los autores discutieron las limitaciones relacionadas con el sesgo de publicación.Estas métricas serán analizadas a detalle en el acápite de interpretación estadística.

4.9. OTRAS MODALIDADES DE SÍNTESIS DE EVIDENCIA

Si bien la mayoría de las revisiones sistemáticas y metaanálisis se centran en intervenciones terapéuticas, existen diversas modalidades de síntesis de evidencia que responden a preguntas distintas y que el lector crítico debe aprender a reconocer. Cada una tiene fortalezas y limitaciones que impactan en la confianza de sus conclusiones.

METAANÁLISIS EN RED (NETWORK META-ANALYSIS, NMA)

El NMA permite comparar múltiples tratamientos de manera simultánea, incluso cuando no existen ensayos que enfrenten directamente todas las opciones. Se basa en una red de comparaciones directas e indirectas, bajo el supuesto de que los estudios son suficientemente comparables (transitividad).

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

Su utilidad radica en que genera rankings de eficacia y seguridad (ej. mediante SUCRA: Surface Underthe Cumulative Ranking).

- Ejemplo: en hipertensión arterial, un NMA puede integrar ensayos que comparan A vs B y B vs C, para inferir la eficacia relativa de A vs C. El lector crítico debe preguntarse si se evaluó la consistencia entre comparaciones directas e indirectas, y si los supuestos de homogeneidad se cumplen.

REVISIONES NARRATIVAS

Son revisiones más tradicionales, que resumen un tema amplio sin seguir una metodología estandarizada. Su fortaleza es ofrecer contexto y actualización general en campos emergentes, pero son muy susceptibles a sesgo de selección, ya que dependen del criterio subjetivo del autor.

- Ejemplo: una revisión narrativa sobre microbiota intestinal y enfermedades autoinmunes puede ser útil para explorar hipótesis, pero no debe utilizarse como base exclusiva para guías de práctica clínica. El lector crítico debe reconocerlas como informativas, pero no concluyentes.

REVISIONES DE ALCANCE (SCOPING REVIEWS)

Exploran un tema amplio sin intentar sintetizar cuantitativamente los resultados. Su objetivo es mapear la literatura existente, identificar vacíos de conocimiento y proponer futuras preguntas de investigación. Se apoyan en guías como PRISMA-ScR.

- Ejemplo: una revisión de alcance podría explorar el uso de inteligencia artificial en diagnóstico radiológico, identificando qué áreas han sido más investigadas y cuáles permanecen inexploradas.

REVISIONES SISTEMÁTICAS CUALITATIVAS

Integran los hallazgos de estudios cualitativos (entrevistas, grupos focales, etnografías). No buscan estimaciones numéricas, sino comprender experiencias, percepciones y contextos de los pacientes.

- Ejemplo: una revisión cualitativa puede analizar las barreras percibidas por los pacientes con diabetes para adherirse a su tratamiento. Para el lector crítico, la clave es valorar si la síntesis cualitativa fue rigurosa, transparente y coherente con la pregunta planteada.

METAANÁLISIS DE PRUEBAS DIAGNÓSTICAS

Evalúan la exactitud diagnóstica de una prueba en comparación con un estándar de referencia. Utilizan modelos bivariados o HSROC (Hierarchical Summary Receiver Operating Characteristic), que permiten estimar sensibilidad, especificidad y curvas resumen ajustadas por heterogeneidad y efecto de umbral.

- Ejemplo: un metaanálisis de troponina ultrasensible para diagnóstico de infarto agudo de miocardio combina estudios con distintos puntos de corte y genera una curva HSROC que permite estimar el rendimiento global de la prueba. En la lectura crítica, se debe verificar si se aplicó QUADAS-2 para evaluar riesgo de sesgo en cada estudio diagnóstico.

4.10 INTERPRETACIÓN Y APLICABILIDAD

El paso final en la lectura crítica de una revisión sistemática o metaanálisis es interpretar los hallazgos y valorar su aplicabilidad en la práctica clínica. No basta con conocer los resultados numéricos; lo esencial es integrarlos al contexto del paciente y del sistema de salud.

En primer lugar, debe analizarse la magnitud y dirección del efecto. Un resultado estadísticamente significativo puede carecer de relevancia clínica si la magnitud del beneficio es pequeña. Aquí cobra importancia el concepto de mínima diferencia clínicamente importante (MCID), que ayuda a determinar si la diferencia observada tiene un impacto real en la vida del paciente.

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

Otro elemento es la certeza de la evidencia, que debe estar valorada con un sistema estandarizado como GRADE. Una conclusión basada en evidencia de alta certeza ofrece mayor confianza que aquella apoyada en estudios de baja calidad o con alta heterogeneidad.

La relevancia clínica debe evaluarse en función de los desenlaces que importan al paciente: mortalidad, hospitalización, calidad de vida y seguridad del tratamiento. No siempre los resultados en variables intermedias (como biomarcadores) se traducen en beneficios tangibles.

Finalmente, la aplicabilidad depende de factores como el perfil de los pacientes estudiados, el contexto local y el balance entre beneficios y riesgos. Incluso un tratamiento eficaz puede no ser la mejor opción si los efectos adversos, los costos o las barreras de acceso superan a sus ventajas.

En síntesis, la interpretación de una revisión sistemática no se limita a los números: exige un juicio clínico informado, capaz de ponderar la calidad de la evidencia, su relevancia y su aplicabilidad a la realidad del paciente frente al médico.

5. EVALUACIÓN DE LA VALIDEZ Y RELEVANCIA

Al leer una revisión sistemática o un metaanálisis, el primer paso del análisis crítico es preguntarse: ¿es válido lo que estoy leyendo? y, en segundo lugar, ¿es relevante para mis pacientes y mi contexto? Una revisión puede estar impecablemente realizada desde el punto de vista metodológico, pero si no responde a una necesidad clínica concreta o si la población analizada difiere demasiado de la práctica habitual, su utilidad será limitada.

5.1 VALIDEZ INTERNA

La validez interna hace referencia a la credibilidad de los resultados y depende de que la revisión haya seguido una metodología rigurosa y transparente. Para valorarla, el lector debe fijarse en:

- Pregunta claramente definida: si se estructuró en formato PICO o similar.
- Protocolo registrado: idealmente en PROSPERO u OSF, con criterios de inclusión y exclusión preespecificados.
- Búsqueda exhaustiva y reproducible: que abarque varias bases de datos, con uso de términos MeSH/DeCS, sin restricciones arbitrarias de idioma.
- Selección y extracción de datos duplicada: realizada por al menos dos revisores de forma independiente.
- Evaluación sistemática del riesgo de sesgo: con herramientas estandarizadas (RoB 2, ROBINS-I, QUADAS-2, AMSTAR 2).
- Síntesis adecuada: narración clara cuando hay heterogeneidad, metaanálisis cuando los estudios son comparables.
- Análisis de heterogeneidad y sesgos de publicación: reportados con I^2, Cochran-Q, funnelplot y pruebas estadísticas apropiadas.

Si alguno de estos elementos falta o se reporta de manera superficial, la confianza en los resultados de la revisión disminuye.

5.2 RELEVANCIA CLÍNICA

Una vez valorada la validez interna, el siguiente paso es considerar la relevancia clínica, es decir, si los hallazgos son aplicables al paciente real. Para ello, conviene reflexionar sobre tres aspectos:

- Población: ¿los pacientes incluidos en los estudios se parecen a los que atiendo en mi práctica diaria?
- Desenlaces: ¿los resultados medidos son clínicamente importantes (ej. mortalidad, hospitalización, calidad de vida) o solo variables intermedias?
- Balance riesgo/beneficio y aplicabilidad local: ¿el beneficio supera los riesgos y los costos? ¿el tratamiento está disponible en mi entorno?

Una revisión puede mostrar un beneficio claro, pero si los pacientes estudiados son muy distintos a los de mi contexto, la aplicabilidad será limitada.

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

6. INTERPRETACIÓN ESTADÍSTICA

La fortaleza de una revisión sistemática o de un metaanálisis no reside únicamente en la calidad de los estudios incluidos ni en la transparencia de la metodología. El verdadero valor surge en el momento de interpretar los resultados estadísticos y comprender lo que significan para la práctica clínica.

Muchos médicos y estudiantes se enfrentan a tablas llenas de siglas (RR, OR, HR, MD, SMD), gráficos con diamantes y cuadrados de distintos tamaños (forestplots), o curvas complejas (HSROC, SUCRA), y sienten que basta con leer la conclusión final del artículo. Sin embargo, la medicina basada en la evidencia exige más que aceptar resúmenes: demanda la capacidad de leer críticamente los números y figuras para responder tres preguntas esenciales:

- ¿Qué mide realmente el resultado estadístico? (ej. un riesgo relativo, una odds ratio o una diferencia de medias).
- ¿Qué tan confiable es esa estimación? (intervalos de confianza, heterogeneidad, sesgo de publicación).
- ¿Qué significa para mi paciente? (certeza de la evidencia, mínima diferencia clínicamente importante, aplicabilidad).

La interpretación estadística, por tanto, es el puente entre los resultados de la investigación y la toma de decisiones clínicas informadas. Sin ella, se corre el riesgo de basarse en conclusiones incompletas o incluso engañosas.

6.1 INTERPRETACIÓN DE LAS REVISIONES NARRATIVAS

Las revisiones narrativas son el formato más común en la literatura médica tradicional. A diferencia de una revisión sistemática, no siempre siguen un protocolo riguroso ni aplican criterios de búsqueda exhaustivos. Suelen tener un carácter más descriptivo y explicativo, reuniendo estudios seleccionados por el autor para ofrecer una visión general sobre un tema.

Desde la perspectiva de la lectura crítica, lo fundamental es entender que la interpretación de una revisión narrativa depende casi por completo de la coherencia argumentativa del autor y de la amplitud de la literatura considerada.

Aspectos clave que debe analizar el lector crítico:

Alcance y objetivos: ¿la revisión está orientada a un área concreta o se limita a un resumen superficial?

Selección de la literatura: ¿el autor explica cómo eligió los artículos o solo cita los que apoyan su postura?

Equilibrio en la presentación: ¿se incluyen tanto estudios positivos como negativos, o se privilegian resultados favorables?

Identificación de limitaciones: ¿el autor señala las brechas de conocimiento y los posibles sesgos, o presenta sus conclusiones como verdades absolutas?

La lectura crítica de una revisión narrativa debe enfocarse en extraer las tendencias generales: qué aspectos parecen más consistentes en la literatura y cuáles son aún inciertos o contradictorios. No se debe asumir que una conclusión es definitiva, sino considerarla como un punto de partida para el razonamiento clínico o la generación de nuevas preguntas de investigación.

6.2 INTERPRETACIÓN DE UN METAANÁLISIS CLÁSICO

Cuando los estudios primarios son comparables en diseño, población y desenlaces, los autores pueden realizar un metaanálisis, es decir, una síntesis estadística que combina cuantitativamente sus resultados. Para el lector crítico, el reto es comprender qué significa el estimador global y cómo interpretarlo clínicamente.

Medidas de efecto: Las más utilizadas en medicina son:

- Riesgo relativo (RR): indica cuántas veces es más probable que ocurra un evento en el grupo intervención respecto al comparador.
- Odds ratio (OR): compara las probabilidades relativas; más común en estudios de casos y controles.
- Hazard ratio (HR): se usa en análisis de supervivencia, donde importa no solo si ocurre el evento, sino cuándo.
- Diferencia de medias (MD): compara directamente medias entre dos grupos (ej. reducción en mmHg de presión arterial).

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

- Diferencia de medias estandarizada (SMD): se utiliza cuando los estudios miden el mismo desenlace con escalas diferentes (ej. dolor en diferentes escalas).

Tabla 4. Medidas de efecto en metaanálisis clásicos

Medida	Uso principal	Valor de no efecto	Interpretación crítica
Riesgo relativo (RR)	Ensayos clínicos; cohorte	1	RR < 1: reducción de riesgo; RR > 1: aumento de riesgo
Odds ratio (OR)	Casos y controles	1	OR < 1: menor probabilidad; OR > 1: mayor probabilidad
Hazard ratio (HR)	Estudios de supervivencia	1	HR < 1: menor riesgo en el tiempo; HR > 1: mayor riesgo
Diferencia de medias (MD)	Comparaciones directas en misma escala	0	MD ≠ 0: diferencia en unidades reales (ej. mmHg, kg)
Diferencia de medias estandarizada SMD	Comparaciones con distintas escalas	0	Estándar: 0.2 pequeño, 0.5 moderado, 0.8 grande (Cohen)

Fuente: Elaboración propia.

Intervalos de confianza: El IC 95 % no solo indica significación estadística, sino también precisión. Intervalos estrechos refieren estimación confiable. Intervalos amplios refieren gran incertidumbre.

Para comprender su interpretación presentamos la siguiente tabla:

Tabla 5. Efecto de los inhibidores SGLT2 en DM2 (síntesis ficticia para fines didácticos)

Desenlace	Medida	Estimación (IC95%)	Valor de no efecto	Modelo	Estudios (n)	I^2 (%)
MACE (ictus/IM/muerte CV)	RR	0.90 (0.85–0.95)	1.00	Aleatorio	10	18
Hospitalización por IC	HR	0.78 (0.72–0.84)	1.00	Aleatorio	10	42
Mortalidad por cualquier causa	OR	0.92 (0.86–0.99)	1.00	Aleatorio	10	20
HbA1c (%)	MD	−0.35 (−0.45 a −0.25)	0.00	Aleatorio	9	30
Peso corporal (kg)	MD	−2.1 (−2.8 a −1.4)	0.00	Aleatorio	8	36
PAS (mmHg)	MD	−3.5 (−5.0 a −2.0)	0.00	Aleatorio	7	28
Calidad de vida (escala variable)	SMD	0.20 (0.10–0.31)	0.00	Aleatorio	5	24

Notas: Datos simulados con fines pedagógicos. RR: riesgo relativo; OR: odds ratio; HR: hazard ratio; MD: diferencia de medias; SMD: diferencia de medias estandarizada; IC95%: intervalo de confianza del 95%; PAS: presión arterial sistólica; IM: infarto de miocardio; CV: cardiovascular; IC: insuficiencia cardíaca.

Para su interpretación: el riesgo relativo (RR) para eventos cardiovasculares mayores (MACE) fue de 0.90 (IC95% 0.85–0.95). Esto significa que los pacientes tratados con estos fármacos tuvieron un 10 % menos de riesgo relativo de presentar un MACE en comparación con el grupo control. Como el intervalo de confianza no cruza la unidad, el resultado es estadísticamente significativo y, además, clínicamente relevante, pues la reducción es consistente y se observa con baja heterogeneidad.

En cuanto a la hospitalización por insuficiencia cardíaca, el análisis mostró un hazard ratio (HR) de 0.78 (IC95% 0.72–0.84). A diferencia del RR, el HR incorpora la dimensión temporal, de modo que este resultado indica que los iSGLT2 no solo reducen la incidencia de hospitalizaciones, sino que también retrasan su aparición en el tiempo. El beneficio es claro y robusto, con una reducción relativa del 22 %.

Respecto a la mortalidad por cualquier causa, la medida utilizada fue el odds ratio (OR), que resultó en 0.92 (IC95% 0.86–0.99). Este hallazgo refleja una reducción modesta en la probabilidad de muerte. Aunque estadísticamente significativo, el límite inferior del intervalo de confianza se acerca mucho a la unidad, lo que invita a interpretar el efecto con cautela, pues la magnitud del beneficio es pequeña y su relevancia clínica discutible.

Para los desenlaces intermedios se emplearon medidas de diferencia de medias. En el caso de la hemoglobina glicosilada, la diferencia de medias (MD) fue de −0.35 % (IC95% −0.45 a −0.25), lo que implica una reducción modesta pero consistente en el control glucémico. En el peso corporal, la MD fue de −2.1 kg (IC95% −2.8 a −1.4), un efecto clínicamente más significativo, pues la pérdida de alrededor de dos kilogramos puede impactar en el control metabólico y cardiovascular a largo plazo.

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

Finalmente, en desenlaces más subjetivos como la calidad de vida se utilizó la diferencia de medias estandarizada (SMD), que fue de 0.20 (IC95% 0.10–0.31). Este valor corresponde a un efecto pequeño según los criterios convencionales, pero relevante al indicar que los pacientes tratados con iSGLT2 reportaron una mejoría consistente en su percepción de salud.

En resumen, la interpretación de estas medidas depende siempre de dos elementos: el valor puntual de la estimación y su intervalo de confianza. En RR, OR y HR, valores por debajo de 1 favorecen a la intervención; en MD y SMD, valores por debajo de 0 indican reducción y por encima de 0, incremento. Sin embargo, la verdadera lectura crítica consiste en valorar no solo la significación estadística, sino también la magnitud del efecto y su pertinencia clínica en relación con el paciente real.

HETEROGENEIDAD

Uno de los errores más frecuentes al leer un metaanálisis es asumir que la estimación global representa un efecto único y aplicable a todos los contextos. En realidad, los estudios incluidos casi nunca son idénticos: difieren en características de los pacientes, intervenciones, dosis, desenlaces medidos o duración del seguimiento. Esta variabilidad se conoce como heterogeneidad y puede modificar de forma importante la interpretación de los resultados.

Existen diferentes tipos de heterogeneidad:

- Clínica: se refiere a diferencias en la población (edad, comorbilidades), en la intervención (fármacos, dosis, duración), en el comparador o en los desenlaces evaluados.
- Metodológica: proviene de diferencias en el diseño de los estudios (ensayos ciegos vs. abierto, tamaños muestrales, criterios de inclusión y exclusión).
- Estadística: es la variabilidad numérica observada en los resultados de los estudios, que se mide con pruebas específicas.

El lector crítico debe tener presente que la heterogeneidad clínica y metodológica son inevitables en investigación, pero lo importante es valorar si estas diferencias son lo suficientemente grandes como para invalidar la síntesis cuantitativa.

Para la medición de la heterogeneidad se utilizan las siguientes medidas estadísticas:

Prueba de Cochran-Q: Es una prueba de hipótesis que evalúa si la variabilidad entre los estudios es mayor de lo que cabría esperar por azar. Un valor de $p < 0,05$ indica heterogeneidad significativa. Sin embargo, tiene limitaciones: cuando hay pocos estudios puede no detectar diferencias reales, y cuando hay muchos estudios tiende a detectar heterogeneidad incluso si esta es clínicamente irrelevante.

Estadístico I^2: Es el indicador más utilizado en la práctica. Expresa en porcentaje la proporción de la variabilidad total que se debe a heterogeneidad real y no al azar, para su interpretacion:

- 0–25 %: heterogeneidad baja.
- 25–50 %: heterogeneidad moderada.
- 50–75 %: heterogeneidad alta.
- 75 %: heterogeneidad muy alta.

Aunque estas categorías son orientativas, el juicio clínico es clave: un I^2 del 40 % puede ser aceptable si los estudios son consistentes en la dirección del efecto, mientras que un I^2 bajo no garantiza ausencia de problemas si los estudios difieren en calidad o población.

ESTRATEGIAS PARA EXPLORAR LA HETEROGENEIDAD

Cuando se detecta heterogeneidad, el metaanálisis no debe detenerse en el reporte del I^2, sino explorar sus causas. Las principales estrategias son:

- Análisis de subgrupos: divide los estudios según una característica específica (ej. pacientes con vs. sin enfermedad cardiovascular previa, dosis alta vs. baja del fármaco) y analiza si el efecto cambia entre subgrupos.
- Metarregresión: técnica estadística más avanzada que permite evaluar la relación entre características de los estudios (ej. edad media de los pacientes, duración del seguimiento) y el tamaño del efecto.
- Análisis de sensibilidad: repite el metaanálisis excluyendo ciertos estudios (ej. aquellos con alto riesgo de sesgo o tamaños extremos) para comprobar si los resultados globales cambian.

La heterogeneidad no invalida automáticamente un metaanálisis, pero obliga a interpretarlo con cautela. Lo importante no es solo el número (I^2 o p del Q), sino si existe una explicación clínica

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

plausible para esa variabilidad y si los autores exploraron adecuadamente sus causas.

FOREST PLOT

El forestplot (o gráfico de bosque) es la representación visual más utilizada para resumir los resultados de un metaanálisis. Su valor radica en que, de un solo vistazo, permite observar los efectos de los estudios individuales y la estimación global combinada.

Los elementos principales de un forestplot son:

- Estudios individuales: cada ensayo aparece representado por un cuadrado en el eje horizontal. La posición del cuadrado indica el valor del estimador puntual (RR, OR, HR, MD o SMD), mientras que la línea horizontal que lo acompaña corresponde a su intervalo de confianza (IC 95 %).
- Tamaño del cuadrado: refleja el peso estadístico del estudio dentro del metaanálisis, que depende sobre todo del tamaño de la muestra y de la precisión de sus resultados. Un ensayo grande como EMPA-REG OUTCOME suele tener un cuadrado más grande que un estudio pequeño.
- Línea vertical de no efecto: situada en el valor 1 cuando se representan RR, OR o HR, o en 0 cuando se representan diferencias de medias. Si la línea de IC de un estudio cruza esta línea de no efecto, se interpreta que ese estudio por sí solo no tiene un resultado estadísticamente significativo.
- Diamante final: representa la estimación global del metaanálisis. Su centro indica el valor combinado del efecto y sus extremos el intervalo de confianza. Si el diamante no cruza la línea de no efecto, el resultado global es estadísticamente significativo.
- Medidas de heterogeneidad: habitualmente se muestran en una esquina del gráfico (I^2, Cochran-Q), para que el lector pueda valorar la consistencia entre estudios.

Así mismo para interpretar un Forest plot se presentan los siguientes tips:

- Observar la dirección de los efectos individuales: si la mayoría de los estudios apuntan hacia el mismo lado del gráfico, la conclusión será más robusta.
- Revisar la magnitud de los efectos: aunque todos vayan en la misma dirección, algunos pueden mostrar beneficios grandes y otros más modestos.
- Valorar los intervalos de confianza: si son amplios, los resultados son menos precisos; si son estrechos, más confiables.
- Identificar el peso relativo de cada estudio: un ensayo de gran tamaño puede dominar la estimación global; el lector debe ser consciente de ello.
- Analizar el diamante global: ¿cruza o no la línea de no efecto?, ¿su IC es estrecho o amplio?
- Relacionar con la heterogeneidad: incluso si el diamante muestra un beneficio, una alta heterogeneidad reduce la certeza del resultado.

Para su correcta interpretación se presentan las siguiente figura:

Figura 1: Forest plot de eventos cardiovasculares mayores (MACE) con inhibidores de SGLT2 (desenlace dicotómico, medida = RR)

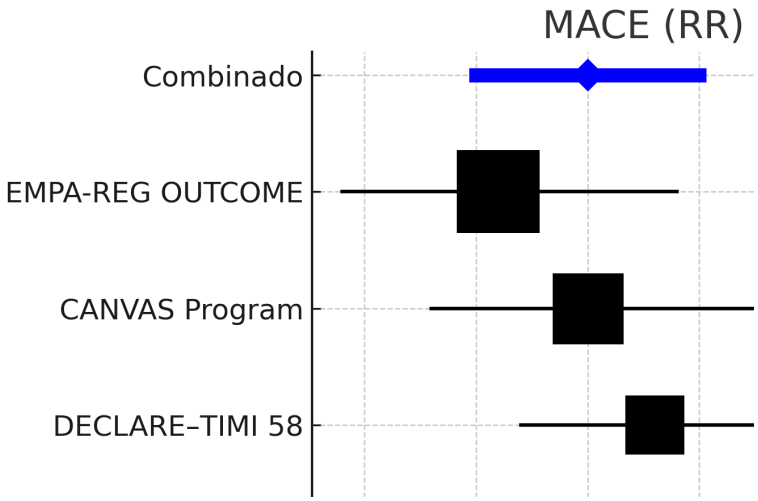


Figura 1. Forest plot de eventos cardiovasculares mayores (MACE) con inhibidores de SGLT2.

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

El resultado global mostró un RR = 0.90 (IC95% 0.85–0.95). (5)Heterogeneidad: I^2 = 18 %, p = 0.21.

Para su interpretación de estudios individuales:

- EMPA-REG OUTCOME: RR ≈ 0.86 con IC95% 0.79–0.94. No cruza la línea de no efecto → resultado significativo, con un beneficio claro.
- CANVAS Program: RR ≈ 0.90 (IC95% 0.83–0.98). También significativo, aunque con un efecto más cercano a 1.
- DECLARE–TIMI 58: RR ≈ 0.93 (IC95% 0.87–1.00). Resultado en el límite; el IC toca la línea de no efecto, lo que indica un beneficio modesto y con menor certeza.
- VERTIS-CV: RR ≈ 0.97 (IC95% 0.89–1.06). El intervalo cruza la línea de no efecto → no significativo.

Para la estimación combinada:

El diamante resume el efecto global: RR = 0.90 (IC95% 0.85–0.95), el centro del diamante indica la estimación puntual (10 % de reducción en riesgo relativo), los extremos corresponden al IC95% que no cruza la línea de no efecto, el beneficio global es estadísticamente significativo.

Para la heterogeneidad:

El análisis reporta I^2 = 18 %, p = 0.21, lo que indica baja heterogeneidad. Esto significa que los estudios son consistentes entre sí y que las diferencias observadas probablemente se deban al azar más que a variaciones clínicas importantes.

Como interpretación final: Este forestplot muestra que los inhibidores de SGLT2 reducen el riesgo de eventos cardiovasculares mayores (MACE) en un 10 % relativo, con un intervalo de confianza estrecho y baja heterogeneidad. Aunque algunos estudios aislados no alcanzan significación (DECLARE, VERTIS), la dirección del efecto es consistente y la síntesis global es robusta.

Figura 2. Forest plot de cambio en HbA1c (%) con inhibidores de SGLT2 (desenlace continuo, medida = diferencia de medias)

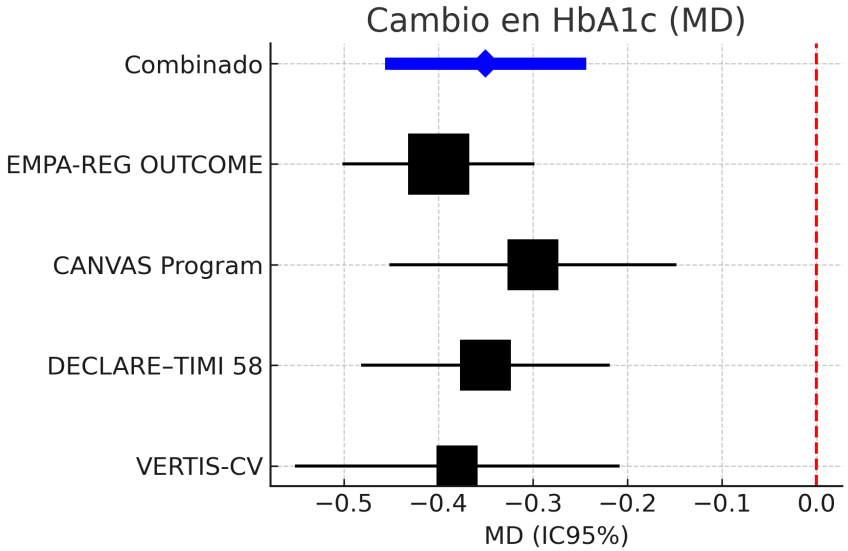


Figura 2. Forest plot de cambio en HbA1c (%) con inhibidores de SGLT2. El resultado global mostró una diferencia de medias (MD) = −0.35 % (IC95% −0.45 a −0.25). Heterogeneidad: I^2 = 30 %, p = 0.15.

Para la interpretación de los estudios individuales:

- EMPA-REG OUTCOME: MD ≈ −0.40 % (IC95% −0.50 a −0.30). Reducción clara y significativa de HbA1c.
- CANVAS Program: MD ≈ −0.30 % (IC95% −0.45 a −0.15). También reducción significativa.
- DECLARE–TIMI 58: MD ≈ −0.35 % (IC95% −0.48 a −0.22). Efecto consistente.
- VERTIS-CV: MD ≈ −0.38 % (IC95% −0.55 a −0.21). Beneficio significativo.

Para la interpretación de estudios combinados:

- El diamante que representa la diferencia de medias combinada, cuenta con MD = −0.35 % (IC95% −0.45 a −0.25), como no cruza la línea de no efecto, el resultado es estadísticamente significativo, la magnitud del efecto es modesta (una reducción de HbA1c de 0.35 %), aunque clínicamente puede ser relevante si se suma a otros beneficios cardiovasculares y renales de estos fármacos.

Para el análisis de heterogeneidad:

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

- El análisis reporta I^2 = 30 %, p = 0.15, lo que corresponde a heterogeneidad moderada. Esto indica que hay cierta variabilidad entre estudios, pero no lo suficiente para invalidar la síntesis. La consistencia general es buena.

El forestplot muestra que los inhibidores de SGLT2 reducen la HbA1c en un promedio de 0.35 puntos porcentuales respecto al comparador, con un intervalo de confianza estrecho y resultados consistentes entre estudios. Aunque la magnitud del beneficio es estadísticamente significativa y clínicamente modesta, cobra importancia cuando se interpreta junto con otros efectos metabólicos (pérdida de peso, reducción de presión arterial) y cardiovasculares.

FUNNELPLOT

El Funnelplot (o gráfico en embudo) es una herramienta visual diseñada para detectar posibles sesgos de publicación en un metaanálisis. Su lógica es sencilla: si todos los estudios realizados sobre un tema se publicaran, sus resultados deberían distribuirse de manera equilibrada alrededor del efecto global. Sin embargo, cuando los estudios pequeños con resultados negativos o neutros no llegan a publicarse, la distribución se vuelve asimétrica.

Para que se entienda mejor, el sesgo de publicación ocurre cuando los resultados de los estudios influyen en su probabilidad de ser publicados. En la práctica, los estudios con resultados positivos (es decir, aquellos que muestran un beneficio claro de la intervención) tienen más posibilidades de ser aceptados en revistas científicas, mientras que los estudios con resultados negativos o neutros suelen publicarse con menor frecuencia o incluso no llegar a difundirse.

Esto genera un problema importante: si los lectores solo tienen acceso a los estudios publicados, la impresión general puede ser que una intervención es más eficaz o más segura de lo que realmente es. En otras palabras, la evidencia disponible está sesgada hacia los resultados favorables.

Los estudios negativos no deben interpretarse como fracasos, sino como piezas fundamentales para conocer el verdadero efecto de una intervención. Un metaanálisis sin ellos puede sobreestimar los beneficios y pasar por alto riesgos relevantes.

El funnelplot busca precisamente detectar este sesgo. Si los estudios negativos pequeños no se publican, el gráfico pierde simetría y aparece un “hueco” en uno de los lados del embudo. Esa asimetría alerta al lector crítico de que la evidencia presentada podría no reflejar toda la realidad, sino solo la parte más favorable.

Las características principales de este grafico son:

Ejes:

- El eje horizontal representa el estimador de efecto (RR, OR, HR o MD).
- El eje vertical muestra la precisión del estudio, generalmente expresada como la inversa del error estándar (1/SE). Así, los estudios grandes y precisos aparecen en la parte superior, mientras que los estudios pequeños y menos precisos quedan en la parte inferior.
- Línea central: corresponde al efecto global estimado por el metaanálisis.
- Líneas punteadas en forma de embudo: delimitan el intervalo de confianza esperado (95 %) alrededor del efecto global. Si no hay sesgo, los estudios deberían agruparse dentro de este embudo de manera simétrica.

Simetría vs. asimetría:

- Simetría: sugiere que no hay evidencia de sesgo de publicación; tanto estudios positivos como neutros aparecen distribuidos a ambos lados del efecto global.
- Asimetría: cuando se observa un “hueco” en uno de los lados, suele indicar que faltan estudios (generalmente los pequeños con resultados negativos), lo cual puede sobreestimar el efecto global.

El lector crítico debe recordar que igual el Funnelplot tiene limitaciones:

- Solo es confiable cuando hay al menos 10 estudios; con menos, la asimetría puede ser producto del azar.
- La asimetría no siempre se debe a sesgo de publicación: también puede reflejar heterogeneidad clínica, diferencias metodológicas o variaciones en la calidad de los estudios.
- Para cuantificar formalmente la asimetría se emplean pruebas estadísticas como la prueba de Egger.

Prueba de Egger: Además de la inspección visual del funnelplot, existen métodos estadísticos para evaluar formalmente la presencia de sesgo de publicación. El más utilizado es la prueba de Egger, que analiza la relación entre el tamaño del efecto y su precisión.

En términos simples, la prueba evalúa si los estudios pequeños tienden a mostrar resultados

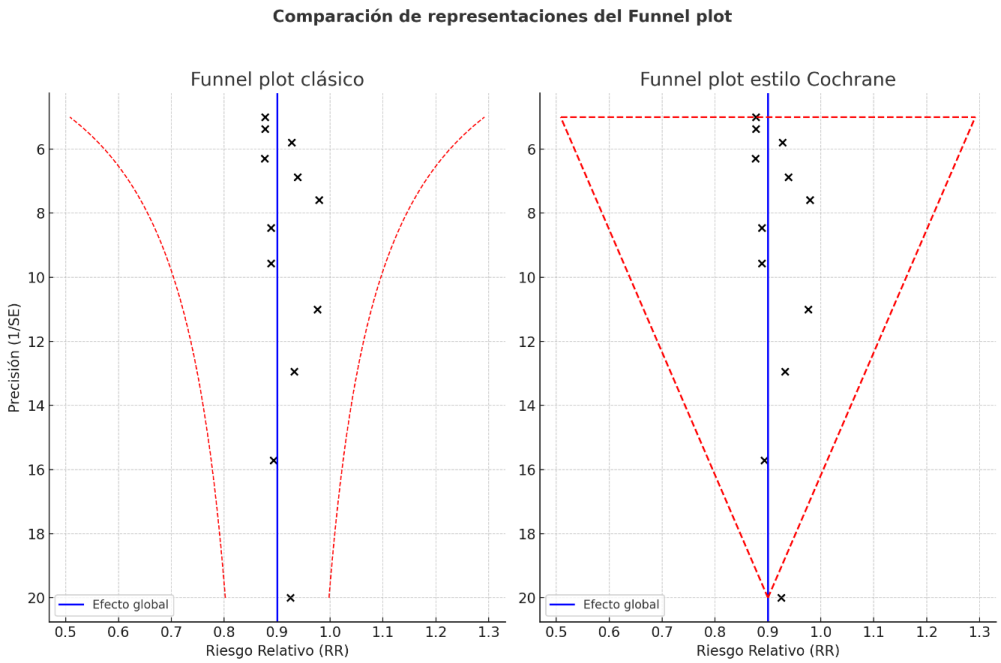
https://doi.org/10.62486/978-9915-9928-3-9.Ch16

diferentes de los estudios grandes. Cuando esto ocurre, puede ser señal de que los estudios pequeños con resultados negativos no fueron publicados.

Para su interpretación: un valor de $p < 0.05$ en la prueba de Egger sugiere asimetría significativa en el funnelplot y, por tanto, posible sesgo de publicación. Su s principales limitaciones radican que, al igual que el funnelplot, la prueba de Egger no es confiable si hay menos de 10 estudios en el metaanálisis, además, la asimetría detectada no siempre se debe a sesgo de publicación, pues también puede reflejar heterogeneidad u otros factores metodológicos.

Con todo esto podemos interpretar el siguiente Funnelplot:

Figura 3. Representaciones del funnelplot para evaluar sesgo de publicación en metaanálisis



Para su interpretación:

En este gráfico cada punto representa un estudio sobre inhibidores de SGLT2 y eventos cardiovasculares mayores (datos ficticios para fines didácticos). Los estudios grandes y más precisos aparecen en la parte superior, muy cercanos al efecto global (línea azul), mientras que los estudios pequeños, menos precisos, se ubican en la parte inferior y presentan mayor dispersión.

La distribución observada es simétrica a ambos lados de la línea de efecto global (RR = 0.90). Esta simetría indica que los estudios publicados incluyen tanto resultados más favorables como menos favorables, pero en conjunto forman la figura de un embudo regular.

Las líneas rojas punteadas delimitan el intervalo esperado del 95 %. La mayoría de los estudios se concentran dentro de esta área, lo que refuerza la impresión de que no existe un sesgo de publicación evidente.

En términos prácticos, no parece que falten “estudios negativos” en un lado del embudo. De existir sesgo de publicación, el gráfico se vería asimétrico, con una acumulación de estudios pequeños solo en el lado de los resultados positivos.

Finalmente, la prueba de Egger aplicada a estos datos simulados fue no significativa ($p = 0.28$), lo que coincide con la inspección visual: no se detecta asimetría relevante y, por tanto, no hay evidencia de sesgo de publicación.

Conclusión: en este ejemplo, el funnelplot da confianza al lector crítico de que el efecto global del metaanálisis (RR = 0.90 para MACE) no está distorsionado por ausencia sistemática de estudios neutros o negativos.

ANÁLISIS DE SENSIBILIDAD

El análisis de sensibilidad es un paso esencial en todo metaanálisis porque permite responder a una pregunta clave: ¿son las conclusiones estables o dependen de ciertas decisiones metodológicas? En otras palabras, nos ayuda a saber si el resultado es robusto. Si los resultados

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

se mantienen estables ante cambios metodológicos plausibles, la confianza del lector aumenta; si cambian de forma relevante, las conclusiones deben matizarse.

DIFERENCIAS ENTRE SUBGRUPOS, META REGRESIÓN Y ANÁLISIS DE SENSIBILIDAD

En la lectura crítica de un metaanálisis es común encontrar tres tipos de análisis complementarios: subgrupos, metarregresión y sensibilidad. Aunque a veces se mencionan juntos, cumplen propósitos distintos y conviene diferenciarlos claramente.

Los análisis de subgrupos dividen los estudios según una característica clínica o metodológica y recalculan el efecto dentro de cada grupo. Por ejemplo, en nuestro estudio puede presentar un análisis separado para pacientes con enfermedad cardiovascular establecida frente a aquellos sin antecedentes. El objetivo aquí es explorar si la magnitud del beneficio varía en función de una característica relevante de la población, la intervención o el diseño.

La **metarregresión** va un paso más allá: en lugar de dividir en categorías, utiliza un modelo estadístico para evaluar si una variable continua (edad media, duración del seguimiento, proporción de mujeres, dosis del fármaco, etc.) se asocia con la magnitud del efecto. Es una técnica poderosa, pero requiere un número elevado de estudios y puede generar conclusiones espurias si se aplica de manera inadecuada.

El **análisis de sensibilidad**, en cambio, tiene un propósito diferente. No busca explicar por qué los estudios son distintos, sino comprobar si las conclusiones del metaanálisis son estables cuando se cambian ciertas decisiones analíticas. Aquí no se trata de explorar heterogeneidad clínica, sino de evaluar la robustez metodológica. Por ejemplo, ¿qué pasa si excluimos los estudios de menor calidad? ¿O si usamos un modelo de efectos fijos en lugar de aleatorios? ¿O si eliminamos un ensayo con resultados extremos? Si las conclusiones se mantienen en todos estos escenarios, el lector puede confiar más en el resultado.

En resumen, los subgrupos y la metarregresión son herramientas para comprender la variabilidad entre estudios, mientras que el análisis de sensibilidad es una prueba de resistencia: pone a los resultados bajo diferentes condiciones para ver si siguen en pie.

Estrategias habituales de análisis de sensibilidad

El lector crítico debe reconocer que no existe una única forma de hacer sensibilidad. Lo importante es comprobar si el resultado se mantiene estable cuando se cambian ciertos supuestos. A continuación, se describen las estrategias más frecuentes, con ejemplos aplicados al caso clínico de inhibidores de SGLT2.

Exclusión por calidad o riesgo de sesgo: Una de las formas más comunes consiste en repetir el metaanálisis excluyendo los estudios con alto riesgo de sesgo. Por ejemplo, si se elimina un ensayo abierto o con pérdidas elevadas en el seguimiento y el resultado se mantiene, la conclusión es más confiable. Si, en cambio, el efecto desaparece al excluir estudios de baja calidad, la evidencia es frágil.

Análisis "leave-one-out" (influencia de un estudio): Esta técnica elimina un estudio por vez y recalcula el estimador global. Sirve para saber si un ensayo específico está "dominando" el resultado. En el caso de los iSGLT2, si al retirar EMPA-REG OUTCOME (uno de los estudios más grandes) el beneficio desapareciera, la robustez sería cuestionable. En nuestro ejemplo ficticio, quitar cualquier estudio apenas modifica la estimación, lo cual demuestra estabilidad.

Modelos alternativos: efectos fijos vs aleatorios: El modelo de efectos fijos supone que todos los estudios estiman el mismo efecto verdadero; el de efectos aleatorios asume que puede haber variabilidad real entre ellos. Comparar ambos es una prueba de consistencia: si el beneficio solo se observa con un modelo y desaparece con el otro, la certeza es baja. En general, el lector debería desconfiar de resultados que dependen exclusivamente de un modelo.

Medidas de efecto alternativas: En ocasiones, el desenlace puede expresarse con diferentes métricas. Por ejemplo, en eventos poco frecuentes es más apropiado el odds ratio (OR) que el riesgo relativo (RR). Un análisis de sensibilidad puede recalcular el resultado usando medidas alternativas, o aplicar métodos de varianza distintos (como Hartung-Knapp). Si las conclusiones se mantienen, la evidencia gana solidez.

Decisiones sobre manejo de datos: No todos los artículos reportan sus datos de la misma forma, y los autores deben tomar decisiones sobre cómo analizarlos. Ejemplos incluyen el uso de población intention-to-treat vs. per-protocol, imputaciones para valores perdidos o la corrección de continuidad en celdas con cero eventos. El lector crítico debe fijarse si esas decisiones alteran o no los resultados.

Definiciones de desenlace: Los desenlaces pueden variar según la definición empleada.

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

En estudios cardiovasculares, por ejemplo, "MACE" puede definirse como la combinación de muerte cardiovascular, infarto y accidente cerebrovascular, o incluir también hospitalización por angina. Un análisis de sensibilidad compara estas definiciones más estrictas o más amplias. Si los resultados cambian de manera importante, las conclusiones deben interpretarse con cautela.

Horizonte temporal: Algunos metaanálisis presentan desenlaces a corto y a largo plazo. Comparar seguimientos ≥24 meses con seguimientos más breves puede mostrar si el beneficio es sostenido o solo transitorio. Por ejemplo, un fármaco podría reducir hospitalizaciones en los primeros meses pero no mantener el efecto a largo plazo.

Tamaño de estudio: Los estudios pequeños suelen mostrar efectos más extremos (positivos o negativos) y son más susceptibles de sesgo. Un análisis de sensibilidad puede excluirlos para verificar si los estudios grandes sostienen la conclusión. Si el beneficio desaparece al considerar solo los ensayos más grandes, la confianza debe disminuir.

Ajuste por sesgo de publicación: Métodos como el trim-and-fill permiten estimar cuántos estudios "faltantes" serían necesarios para equilibrar el funnelplot y cómo se modificaría el efecto global si se incorporaran. Aunque es un método imperfecto, ofrece una idea de la magnitud del sesgo de publicación potencial.

Metaanálisis acumulativo (cumulative o chronological MA): Esta técnica consiste en añadir los estudios en orden temporal y recalcular el efecto cada vez. Permite visualizar cómo ha evolucionado la evidencia a lo largo del tiempo. Si el efecto se mantiene estable a medida que se agregan más ensayos, la conclusión es más sólida; si varía drásticamente, sugiere fragilidad de la evidencia.

En síntesis: todas estas estrategias son como "golpes de estrés" aplicados al metaanálisis. Si los resultados permanecen estables, podemos confiar en que la conclusión es robusta. Si cambian demasiado, es señal de que la evidencia es más frágil de lo que aparenta.

Tabla 6. Estrategias de análisis de sensibilidad en metaanálisis

Estrategia	Objetivo principal	Interpretación si cambia el resultado
Exclusión por calidad/riesgo de sesgo	Evaluar si los resultados dependen de estudios con alto riesgo de sesgo.	Si el efecto desaparece → la conclusión depende de estudios poco confiables.
Leave-one-out (influencia)	Comprobar si un estudio individual "domina" la síntesis.	Si al quitar un ensayo el resultado cambia drásticamente → el efecto no es robusto.
Modelos alternativos (fijo vs aleatorio)	Verificar consistencia entre supuestos estadísticos.	Si solo un modelo muestra beneficio → certeza baja.
Medida de efecto alternativa	Confirmar que el efecto se mantiene con distintas métricas (RR, OR, HR, MD).	Si la dirección cambia según la medida → interpretar con cautela.
Manejo de datos (ITT vs per-protocol, imputaciones, ceros)	Evaluar impacto de decisiones analíticas.	Si el efecto depende de una sola forma de análisis → baja robustez.
Definiciones de desenlace (estricto vs amplio)	Verificar consistencia según cómo se defina el desenlace.	Si las definiciones cambian el resultado → las conclusiones son sensibles a la definición usada.
Horizonte temporal (corto vs largo plazo)	Evaluar estabilidad del efecto en el tiempo.	Si el beneficio desaparece en seguimientos prolongados → evidencia limitada.
Tamaño de estudio (grandes vs pequeños)	Comprobar si los estudios grandes sostienen la conclusión.	Si el efecto depende de estudios pequeños → riesgo de sobreestimación.
Ajuste por sesgo de publicación (trim-and-fill)	Estimar impacto de estudios faltantes.	Si el efecto ajustado pierde significación → posible sesgo de publicación.
Metaanálisis acumulativo (chronological)	Ver la evolución del efecto a medida que se acumula evidencia.	Si el efecto es inestable con nuevos estudios → conclusión frágil.

Fuente: Elaboración propia.

6.3 INTERPRETACIÓN DE METAANÁLISIS DE RED

El metaanálisis en red, conocido por sus siglas en inglés como NMA (Network Meta-Analysis), representa una evolución del metaanálisis tradicional. Mientras que en un metaanálisis clásico solo se pueden comparar dos intervenciones a la vez —por ejemplo, un fármaco frente a placebo—, el NMA permite integrar de manera simultánea múltiples tratamientos dentro de un mismo análisis. Esto resulta especialmente valioso en áreas de la medicina donde existen

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

numerosas opciones terapéuticas y no todos los comparadores han sido enfrentados entre sí en ensayos clínicos directos.

La lógica que sostiene este enfoque es sencilla: si un conjunto de ensayos compara el tratamiento A contra el tratamiento B, y otro conjunto compara B contra C, es posible inferir indirectamente la comparación entre A y C. De esta manera, el NMA combina tanto la evidencia directa (cuando los ensayos comparan intervenciones cara a cara) como la evidencia indirecta (cuando la comparación se establece a través de un tercer tratamiento común).

GRÁFICOS DE RED O NETWORK PLOT

El elemento visual más característico de un metaanálisis en red es el gráfico de red o networkplot. En este diagrama, cada intervención se representa como un nodo, y las líneas que conectan a los nodos reflejan la existencia de comparaciones directas entre esos tratamientos. Cuanto más gruesa es la línea, mayor es el número de estudios que sustentan esa comparación.

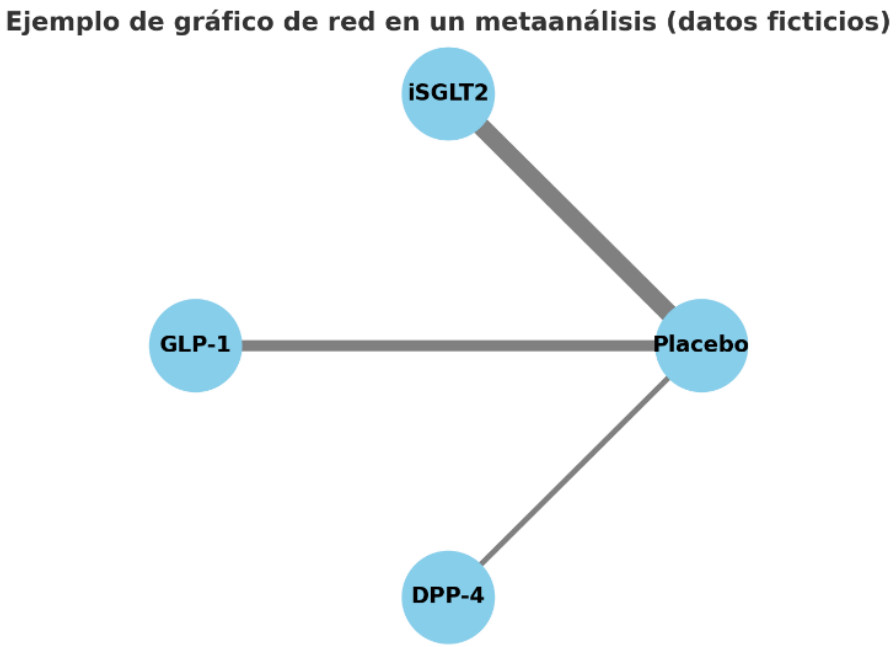
La interpretación de este gráfico es fundamental porque ofrece una visión panorámica de la evidencia disponible. Un nodo conectado únicamente a placebo indica que la intervención nunca ha sido comparada de manera directa con otros tratamientos activos. En cambio, si existen varias conexiones, el lector puede comprobar la solidez y la riqueza de la red de evidencia.

En términos prácticos, el gráfico de red permite responder rápidamente a tres preguntas críticas:

- ¿Qué intervenciones han sido comparadas directamente?
- ¿Cuáles solo se conectan a través de comparaciones indirectas?
- ¿Qué tan equilibrada es la red en cuanto a número y tamaño de los estudios?

Veamos un ejemplo aplicado:

Figura 4. Gráfico de red en un metaanálisis ficticio de tratamientos para diabetes tipo 2.



En este ensayo clínico de tres familias de fármacos para diabetes tipo 2: inhibidores de SGLT2, agonistas de GLP-1 e inhibidores de DPP-4. Existen ensayos que comparan cada grupo de fármacos con placebo, pero no todos los comparan entre sí. El gráfico de red muestra tres nodos (iSGLT2, GLP-1 y DPP-4) unidos a un nodo central (placebo). El grosor de las líneas variaría según cuántos estudios respalden cada comparación: una línea más gruesa entre iSGLT2 y placebo reflejaría que existen más ensayos en esa comparación que en la de DPP-4 frente a placebo. La fortaleza del metaanálisis en red radica precisamente en esta estructura: incluso si no hay un ensayo directo que compare iSGLT2 contra GLP-1, la red permite estimar indirectamente esa comparación gracias al “puente” común del placebo.

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

CONSISTENCIA

La consistencia es el pilar que sostiene la validez de un metaanálisis en red. Recordemos que la fuerza de este tipo de análisis radica en la posibilidad de combinar comparaciones directas e indirectas. Pero para que las conclusiones sean confiables, ambas deben llegar a resultados similares.

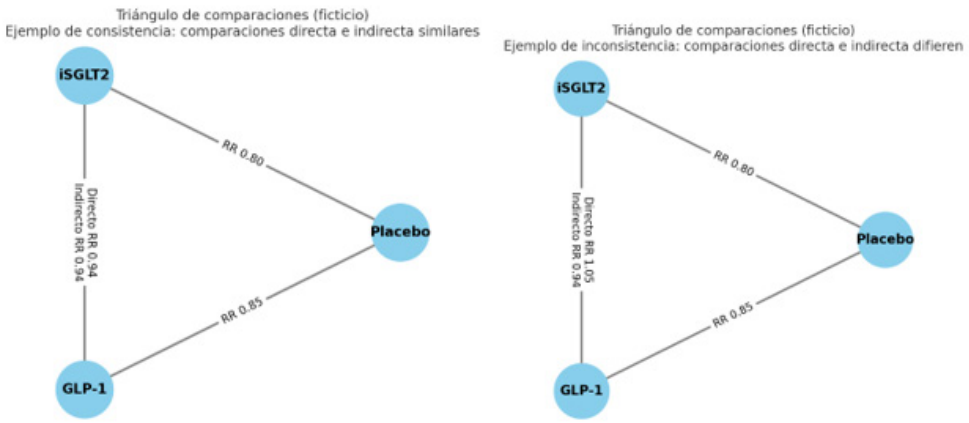
En términos simples, hay consistencia cuando la estimación de efecto obtenida a partir de los ensayos que comparan directamente dos tratamientos coincide con la que se obtiene de manera indirecta a través de un tercero. Si los resultados difieren, hablamos de inconsistencia o incoherencia, lo que puede comprometer seriamente la interpretación del metaanálisis.

Un ejemplo puede clarificarlo. Supongamos que los ensayos clínicos muestran que el inhibidor de SGLT2 reduce un 20 % los eventos cardiovasculares frente a placebo, y que los agonistas de GLP-1 reducen un 15 % esos mismos eventos frente a placebo. Por inferencia indirecta, cabría esperar que la comparación entre SGLT2 y GLP-1 muestre una diferencia cercana a ese 5 %. Si, sin embargo, los pocos estudios que comparan directamente ambos tratamientos encuentran que no hay diferencia o incluso que GLP-1 es superior, surge una inconsistencia.

Existen diferentes formas de evaluar esta consistencia. Algunas son locales, centradas en un lazo específico de la red (por ejemplo, un triángulo de comparaciones entre tres intervenciones), y otras son globales, que valoran la coherencia de toda la red. Entre las pruebas más usadas se encuentran el método de Bucher, que compara la diferencia entre la evidencia directa e indirecta en un triángulo cerrado, y los modelos de inconsistencia global, que analizan toda la red de manera simultánea.

Para el lector crítico, no es necesario dominar todos los métodos estadísticos, pero sí reconocer un punto esencial: si la red muestra inconsistencias importantes, la validez de sus conclusiones disminuye. En tales casos, el ranking de tratamientos puede ser engañoso y la certeza de la evidencia debería degradarse.

En la práctica, la consistencia se interpreta como una condición de credibilidad: cuando está presente, el NMA amplía el alcance de la evidencia; cuando falta, obliga a tomar las conclusiones con cautela y a priorizar los resultados de comparaciones directas.



Cada nodo representa una intervención (placebo, iSGLT2, GLP-1) y las líneas muestran las comparaciones disponibles. En el panel izquierdo, la estimación directa de iSGLT2 vs. GLP-1 (RR 0.94) coincide con la estimación indirecta a través del placebo (RR 0.94), lo que refleja consistencia en la red. En el panel derecho, la comparación directa (RR 1.05) difiere de la estimación indirecta (RR 0.94), lo que evidencia inconsistencia. La consistencia es un requisito fundamental para confiar en las conclusiones de un NMA; cuando falta, las comparaciones indirectas deben interpretarse con cautela y se debe dar mayor peso a la evidencia directa.

RANKINGS (SUCRA)

Uno de los aportes distintivos del metaanálisis en red es la posibilidad de ordenar los tratamientos según su probabilidad de ser el más eficaz o el más seguro. Esto se hace calculando, para cada intervención, la probabilidad acumulada de ocupar cada posición dentro del ranking.

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

La medida más utilizada para expresar este resultado es el SUCRA (Surface Underthe Cumulative Ranking curve). Este indicador varía entre 0 y 1:

- Un valor cercano a 1 significa que la intervención tiene alta probabilidad de estar entre las mejores opciones.
- Un valor cercano a 0 indica que probablemente se ubique entre las menos eficaces o menos seguras.

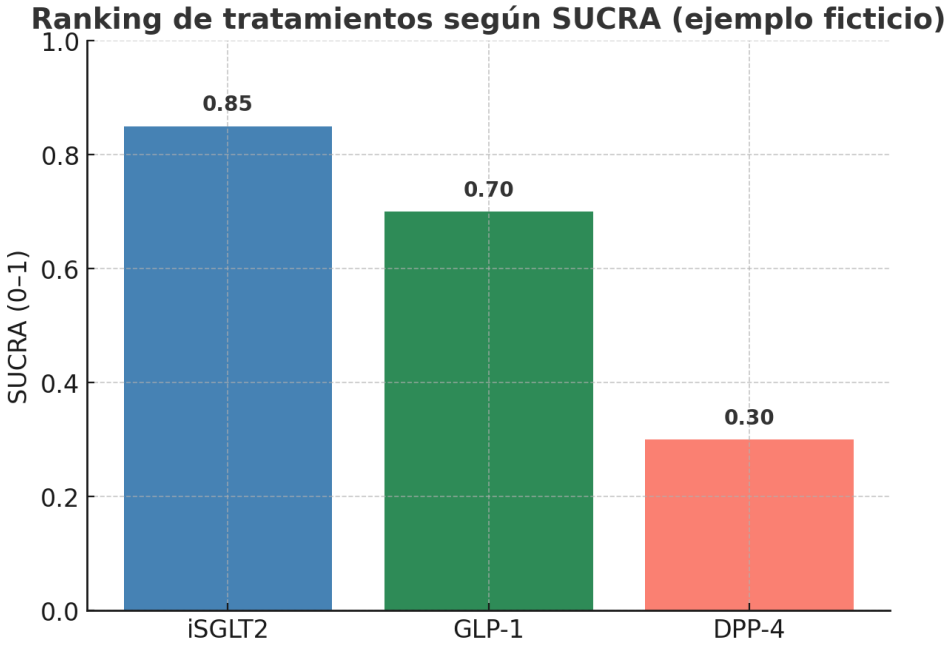
Por ejemplo, en un metaanálisis en red sobre antidiabéticos, los inhibidores de SGLT2 podrían obtener un SUCRA de 0.85 para la reducción de eventos cardiovasculares, los agonistas de GLP-1 un 0.70 y los inhibidores de DPP-4 un 0.30. Esto sugiere que iSGLT2 es la opción más sólida, seguida de GLP-1, mientras que los DPP-4 tienen menor probabilidad de beneficio.

Es importante subrayar que el ranking no equivale a una verdad absoluta. Un SUCRA alto no significa que el tratamiento sea siempre superior, sino que, dentro del conjunto de estudios analizados, tiene mayor probabilidad de ser el mejor. Además, las diferencias entre intervenciones con SUCRA cercanos pueden no ser clínicamente relevantes.

Para la lectura crítica, el ranking debe interpretarse siempre junto con la calidad de la evidencia y con la magnitud absoluta del efecto. Un tratamiento puede aparecer primero en la jerarquía, pero si la certeza de la evidencia es baja (por ejemplo, debido a inconsistencia o riesgo de sesgo), esa posición debe relativizarse.

En síntesis, los rankings del NMA ofrecen una visión comparativa muy útil para la práctica clínica y la elaboración de guías, pero el lector crítico debe tomarlos como una herramienta complementaria, no como el único criterio para la toma de decisiones.

Figura 6. Ranking de tratamientos según SUCRA en un metaanálisis en red (ejemplo ficticio) para eventos cardiovasculares.



En esta figura se muestran los resultados del ranking de eficacia de tres familias de fármacos para diabetes tipo 2 (iSGLT2, GLP-1 y DPP-4) en la reducción de eventos cardiovasculares mayores, expresados mediante el indicador SUCRA (Surface Underthe Cumulative Ranking curve).

Los inhibidores de SGLT2 presentan un SUCRA de 0.85, lo que indica una alta probabilidad de ocupar las primeras posiciones en el ranking. Los agonistas de GLP-1 alcanzan un valor intermedio de 0.70, lo que los sitúa como una alternativa también favorable, aunque por debajo de iSGLT2. Finalmente, los inhibidores de DPP-4 muestran un SUCRA de 0.30, reflejando una menor probabilidad de ser considerados entre las opciones más eficaces.

LEAGUE TABLE

Una de las formas más claras de presentar los resultados de un NMA es la llamada league

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

table o tabla de comparaciones múltiples. Se trata de una matriz en la que cada fila y cada columna corresponden a un tratamiento, y en las celdas se presentan las comparaciones par a par.

Esta tabla permite que el lector vea de un solo vistazo todas las estimaciones posibles, incluso aquellas que provienen de comparaciones indirectas. Normalmente se reporta la medida de efecto (RR, OR, HR o diferencia de medias) junto con su intervalo de confianza al 95 %.

La interpretación es sencilla pero poderosa:

- Cada celda indica cómo se comporta el tratamiento de la fila en comparación con el de la columna.
- Un valor menor a 1 (en RR u OR) favorece al tratamiento de la fila; un valor mayor a 1 favorece al de la columna.
- La diagonal de la tabla queda vacía o marcada con "–" porque representa la comparación de un tratamiento consigo mismo.
- En artículos internacionales, estas tablas suelen ir acompañadas de colores o sombreados que ayudan a visualizar rápidamente qué tratamiento resulta superior.

Tabla 7. Comparaciones entre tratamientos en un metaanálisis en red (ejemplo ficticio) para eventos cardiovasculares.

	iSGLT2	GLP-1	DPP-4
iSGLT2	–	0.94 (0.88–0.99)	0.80 (0.72–0.89)
GLP-1	1.06 (1.01–1.14)	–	0.85 (0.77–0.94)
DPP-4	1.25 (1.12–1.39)	1.18 (1.06–1.30)	–

Fuente: Elaboración propia.

La tabla muestra de manera simultánea todas las comparaciones entre los tres tratamientos incluidos en el metaanálisis en red. En este caso ficticio, los inhibidores de SGLT2, los agonistas de GLP-1 y los inhibidores de DPP-4 fueron evaluados en la reducción de eventos cardiovasculares mayores, utilizando el riesgo relativo como medida de efecto.

Al analizar la fila correspondiente a los inhibidores de SGLT2, se observa que estos presentan un riesgo relativo de 0.94 frente a GLP-1, lo que indica una ligera ventaja a favor de los SGLT2, con un intervalo de confianza que descarta la ausencia de efecto. En la comparación contra los inhibidores de DPP-4, el riesgo relativo de 0.80 refleja un beneficio aún más marcado de los SGLT2, lo que refuerza su posición como la opción más eficaz dentro de la red.

Cuando se examinan los agonistas de GLP-1, se aprecia que, en comparación con los DPP-4, también ofrecen una reducción significativa del riesgo (RR 0.85), situándolos en una posición intermedia: menos eficaces que los SGLT2, pero claramente superiores a los DPP-4.

Por último, los inhibidores de DPP-4 aparecen sistemáticamente como la opción menos favorable, ya que frente a SGLT2 y GLP-1 presentan riesgos relativos superiores a 1, lo que implica una mayor probabilidad de eventos cardiovasculares en comparación con sus competidores.

En conjunto, la tabla permite visualizar de un vistazo la jerarquía terapéutica: los inhibidores de SGLT2 se perfilan como la intervención con mejores resultados, seguidos de los agonistas de GLP-1, mientras que los inhibidores de DPP-4 se ubican en el último lugar. Este tipo de presentación es especialmente útil para el lector clínico, pues ofrece un panorama completo de las comparaciones en un solo cuadro y facilita la interpretación crítica de la red de evidencia.

FOREST PLOT DE COMPARACIONES CLAVE

Aunque el metaanálisis en red genera un gran número de comparaciones posibles, no todas son igualmente relevantes para el lector clínico. Por ello, muchos artículos presentan forestplots resumidos con las comparaciones más importantes, generalmente:

- Cada tratamiento activo frente a placebo o control.
- Las comparaciones entre los tratamientos mejor posicionados en el ranking.

El formato es el mismo que en un metaanálisis clásico, la utilidad de este gráfico es doble. En primer lugar, permite visualizar rápidamente cuáles intervenciones ofrecen un beneficio clínico significativo frente al control. En segundo lugar, facilita la comparación directa entre los tratamientos que compiten por el primer lugar en el ranking, lo que ayuda a tomar decisiones prácticas en guías y protocolos clínicos.

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

6.4 SÍNTESIS NARRATIVAS

La síntesis narrativa es el método más utilizado cuando los estudios disponibles no pueden combinarse cuantitativamente en un metaanálisis. Esto suele ocurrir porque los ensayos son demasiado heterogéneos en cuanto a población, intervenciones, desenlaces o metodología, o porque la pregunta de investigación es demasiado amplia para permitir una síntesis estadística coherente.

En este enfoque, los resultados se presentan en forma de texto, a menudo acompañados de tablas que resumen las características de los estudios y sus hallazgos principales. El objetivo no es obtener un estimador global, sino describir las tendencias generales de la evidencia, identificar coincidencias y discrepancias, y señalar las limitaciones y vacíos de conocimiento.

Por ejemplo, una síntesis narrativa sobre terapias no farmacológicas para diabetes tipo 2 podría concluir que la mayoría de los estudios apoyan el beneficio de la dieta mediterránea y el ejercicio regular en el control glucémico, aunque con diferencias en la magnitud del efecto y en la calidad metodológica. En lugar de un único número combinado, el lector recibe un panorama contextualizado que le permite comprender mejor la diversidad de enfoques.

La ventaja de la síntesis narrativa es su flexibilidad: permite incluir estudios de distintos diseños y responder preguntas amplias. Sin embargo, su principal limitación es la subjetividad, ya que depende en gran medida de cómo los autores seleccionen, organicen e interpreten la información. Esto implica que, para tener credibilidad, debe estar guiada por un protocolo claro y transparente, con criterios explícitos de inclusión y exclusión.

Para el lector crítico, una buena síntesis narrativa debería responder a preguntas como:

- ¿Qué tratamientos o intervenciones muestran resultados consistentes en la mayoría de los estudios?
- ¿Dónde se encuentran las discrepancias y cuáles podrían ser sus causas?
- ¿Qué lagunas en la evidencia justifican nuevas investigaciones?

En conclusión, la síntesis narrativa es una herramienta valiosa cuando el metaanálisis no es posible, pero sus conclusiones deben interpretarse con mayor cautela, reconociendo el riesgo de sesgo y la ausencia de un análisis cuantitativo formal.

6.5 METAANÁLISIS DE PRUEBAS DIAGNOSTICAS

Los metaanálisis de pruebas diagnósticas tienen como objetivo sintetizar la exactitud de una prueba en diferentes contextos clínicos y poblaciones. A diferencia de los metaanálisis de intervenciones, en los que el desenlace suele ser un riesgo relativo o una diferencia de medias, aquí el interés se centra en medidas como la sensibilidad, la especificidad y las razones de verosimilitud (likelihood ratios), para mas explicación de estas pruebas consultar el Capitulo 10 de Pruebas diagnósticas.

La síntesis no puede hacerse con métodos simples de promedios, ya que sensibilidad y especificidad están correlacionadas: cuando una aumenta, la otra suele disminuir. Por ello, se utilizan modelos más sofisticados, como el modelo bivariado de efectos aleatorios o las curvas ROC jerárquicas (HSROC). Estas herramientas permiten estimar el rendimiento global de la prueba considerando la variabilidad entre estudios.

FOREST PLOT DE SENSIBILIDAD Y ESPECIFICIDAD:

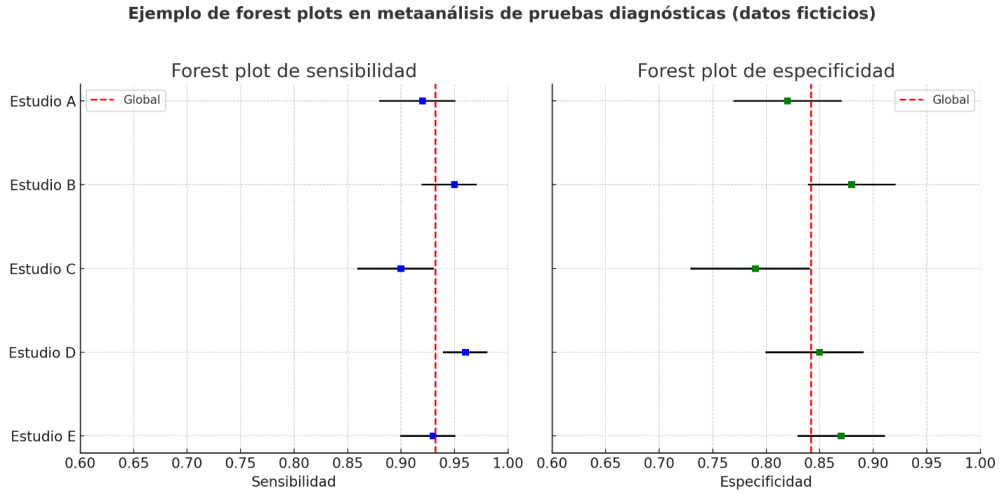
Los metaanálisis de pruebas diagnósticas suelen presentar dos forestplots paralelos: uno para la sensibilidad y otro para la especificidad. En cada gráfico, cada estudio aparece representado por un cuadrado (estimación puntual) acompañado de una línea horizontal (intervalo de confianza al 95 %). Al final del gráfico, se muestra un diamante que representa la estimación global combinada.

La interpretación es similar a la de cualquier forestplot, pero con una lógica centrada en el diagnóstico:

- En el forestplot de sensibilidad, los estudios se alinean alrededor de su capacidad para identificar correctamente a los pacientes con la enfermedad. Una estimación global alta (por ejemplo, >90 %) indica que la prueba es muy eficaz para detectar enfermos.
- En el forestplot de especificidad, los estudios se alinean en torno a su capacidad para identificar correctamente a quienes no tienen la enfermedad. Una estimación global elevada (por ejemplo, >85 %) significa que la prueba es útil para descartar sanos.

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

Figura 7. Forest plots de sensibilidad y especificidad en un metaanálisis de pruebas diagnósticas (datos ficticios).



En el gráfico de la izquierda se observa la sensibilidad de la troponina ultrasensible para el diagnóstico de infarto agudo de miocardio en distintos estudios. La mayoría de los ensayos se agrupan por encima del 90 %, con intervalos de confianza relativamente estrechos. La estimación global, representada por la línea roja discontinua, confirma que la prueba es altamente sensible y, por lo tanto, muy poco probable que pase por alto un caso verdadero de infarto.

El gráfico de la derecha muestra la especificidad, es decir, la capacidad de la prueba para identificar correctamente a los individuos sanos. En este caso se observa una mayor variabilidad entre los estudios, con valores que oscilan entre el 79 % y el 88 %. La estimación global se sitúa en torno al 85 %, lo que indica que, si bien la prueba es bastante específica, aún existe un número no despreciable de falsos positivos.

En conjunto, las figuras ilustran una conclusión clínica importante: la troponina ultrasensible es muy eficaz para descartar infarto cuando el resultado es negativo, pero su interpretación debe complementarse con la clínica y otras pruebas debido a la presencia de falsos positivos en algunos contextos.

CURVAS ROC JERÁRQUICAS (HSROC)

En los metaanálisis de pruebas diagnósticas, la presentación de forestplots de sensibilidad y especificidad resulta muy útil, pero tiene una limitación: obliga al lector a mirar dos gráficos por separado. Para superar este problema, se utilizan las curvas ROC jerárquicas (HSROC, por sus siglas en inglés), que permiten visualizar la exactitud diagnóstica de una prueba en un solo gráfico bidimensional.

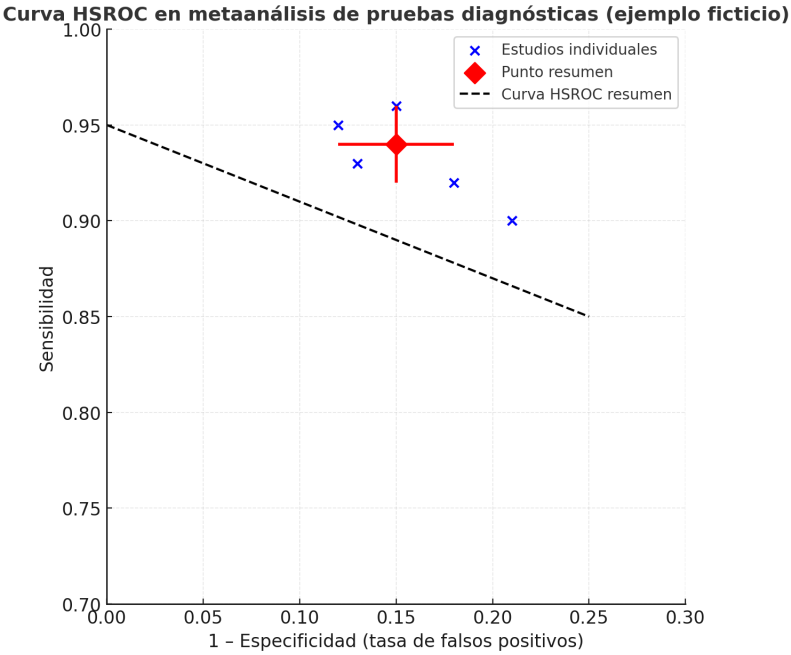
En este tipo de representación, cada estudio individual se ubica en un plano cartesiano donde el eje X corresponde a 1 – especificidad (tasa de falsos positivos) y el eje Y a la sensibilidad. De esta manera, los estudios con mejor rendimiento aparecen en la parte superior izquierda de la gráfica (alta sensibilidad y alta especificidad).

La gran ventaja del modelo HSROC es que ajusta estadísticamente la correlación entre sensibilidad y especificidad, permitiendo trazar una curva resumen que refleja el rendimiento global de la prueba. Además, se calcula un punto resumen con su intervalo de confianza, lo que facilita cuantificar la exactitud media.

La interpretación crítica de una curva HSROC debe responder a tres preguntas:

- ¿Dónde se ubica el punto resumen? → cuanto más cercano al ángulo superior izquierdo, mejor es la prueba.
- ¿Qué tan anchos son los intervalos de confianza? → intervalos estrechos sugieren homogeneidad y mayor certeza; intervalos amplios reflejan variabilidad entre estudios.
- ¿Cómo se distribuyen los puntos de los estudios? → una nube de puntos muy dispersa indica heterogeneidad, lo que puede limitar la aplicabilidad de la prueba en distintos contextos clínicos.

https://doi.org/10.62486/978-9915-9928-3-9.Ch16



Se observa que la mayoría de los estudios se concentran en la parte superior izquierda de la gráfica, lo que refleja una alta sensibilidad (≥90 %) y una especificidad moderada (80–90 %). El rombo rojo señala el punto resumen combinado (sensibilidad aproximada del 94 % y especificidad del 85 %), acompañado de sus intervalos de confianza. Este punto cercano al ángulo superior izquierdo indica que, en conjunto, la prueba tiene un rendimiento diagnóstico global muy favorable.

La curva discontinua muestra la trayectoria HSROC que resume el comportamiento esperado de la prueba a lo largo de distintos umbrales de decisión. La proximidad de esta curva al ángulo superior izquierdo es un indicador visual de la solidez del desempeño de la prueba.

Desde una perspectiva clínica, esta representación permite concluir que la troponina ultrasensible es una herramienta muy útil para descartar infarto en pacientes con dolor torácico, ya que rara vez omite un caso verdadero. Sin embargo, la dispersión de los puntos entre estudios recuerda que existe cierta heterogeneidad, y que la especificidad no es perfecta: en algunos contextos clínicos la prueba puede generar falsos positivos que deben interpretarse con cautela junto a la historia clínica y otros exámenes complementarios.

6.5 MÍNIMA DIFERENCIA CLÍNICAMENTE IMPORTANTE

En la interpretación de un metaanálisis no basta con comprobar si los resultados son estadísticamente significativos. El verdadero desafío para el lector crítico es preguntarse si esos resultados son clínicamente relevantes. Aquí es donde entra en juego la MCID (MinimalClinicallyImportantDifference), o mínima diferencia clínicamente importante.

La MCID se define como la menor magnitud de cambio en un desenlace que los pacientes perciben como beneficiosa o que justifica una decisión clínica. Dicho de otra forma, es el umbral que separa un efecto “estadísticamente significativo” de un efecto que realmente tiene sentido en la práctica.

Un ejemplo ayuda a clarificarlo. Supongamos que un metaanálisis de un nuevo analgésico muestra una reducción media del dolor de 0.5 puntos en una escala de 0 a 10, con un intervalo de confianza estrecho y un valor $p < 0.01$. Desde el punto de vista estadístico, el resultado es sólido. Pero si la MCID definida para el dolor en esa escala es de 1.5 puntos, la conclusión clínica es distinta: el fármaco sí reduce el dolor, pero el cambio es tan pequeño que probablemente los pacientes no lo perciban como relevante.

En los metaanálisis, la MCID se utiliza de varias formas. Puede guiar la interpretación de diferencias de medias (MD o SMD), indicando si el efecto supera o no el umbral clínicamente significativo. También se emplea para valorar medidas de riesgo: por ejemplo, una reducción absoluta del riesgo de 2 % en eventos cardiovasculares puede ser estadísticamente significativa, pero si el umbral clínicamente relevante se considera en torno al 5 %, el resultado debe interpretarse

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

con cautela.

La MCID no es un valor universal, sino que depende del contexto clínico, la población y el desenlace. En muchos casos proviene de estudios de validación, consenso de expertos o encuestas a pacientes. Por ello, su correcta interpretación exige que el lector sepa qué MCID se ha definido para cada desenlace y si el metaanálisis lo supera.

En conclusión, la MCID ayuda a distinguir entre lo estadísticamente significativo y lo clínicamente relevante. Un metaanálisis bien interpretado debe siempre considerar este umbral, de manera que los resultados no solo sean confiables en términos numéricos, sino también útiles para orientar decisiones clínicas reales.

6.6 SISTEMA GRADE

El sistema GRADE (GradingofRecommendationsAssessment, Development and Evaluation) se ha convertido en el estándar internacional para valorar la calidad o certeza de la evidencia científica. Su importancia radica en que no se limita a describir si un resultado es "estadísticamente significativo", sino que se centra en la confianza que podemos tener en ese resultado y en su relevancia para la práctica clínica.

En términos sencillos, GRADE responde a la pregunta: ¿qué tan seguros estamos de que el efecto estimado por un metaanálisis refleja la realidad clínica?

Niveles de certeza: GRADE clasifica la evidencia en cuatro niveles:

- Alta: tenemos gran seguridad en que el efecto verdadero está muy próximo al estimado. En la práctica, significa que nuevos estudios es muy poco probable que cambien la conclusión.
- Moderada: existe confianza razonable en la estimación, pero persiste la posibilidad de que estudios futuros modifiquen la magnitud o incluso la dirección del efecto.
- Baja: el efecto real puede ser sustancialmente diferente al estimado, y la conclusión debe tomarse con cautela.
- Muy baja: la incertidumbre es tan grande que el efecto estimado puede no reflejar en absoluto la realidad.

Esta clasificación se realiza por cada desenlace crítico y no de manera global para todo el metaanálisis. Así, un mismo estudio puede tener certeza alta para mortalidad y certeza baja para calidad de vida.

Existen factores que disminuyen la certeza, la calificación de certeza comienza en un nivel distinto según el tipo de estudio:

- Los ensayos clínicos aleatorizados parten en "alta certeza".
- Los estudios observacionales parten en "baja certeza".

A partir de esa base, la certeza puede descender por cinco motivos:

- Riesgo de sesgo: estudios con ocultamiento inadecuado, pérdidas significativas, falta de cegamiento o análisis selectivo de resultados.
- Inconsistencia: resultados heterogéneos entre estudios, sin explicación clara.
- Indirectitud: la población, intervención, comparador o desenlace no corresponden exactamente a la pregunta clínica. Por ejemplo, usar estudios en adultos jóvenes para responder una pregunta sobre adultos mayores.
- Imprecisión: intervalos de confianza amplios, escaso número de eventos o tamaños muestrales reducidos.
- Sesgo de publicación: sospecha de que solo se han publicado estudios con resultados positivos.

Asi mismo existen factores que pueden aumentar la certeza, en el caso de estudios observacionales, la certeza puede incrementarse si se cumplen ciertos criterios:

- Un efecto muy grande (ejemplo: reducción de mortalidad del 80 %).
- Evidencia de relación dosis-respuesta.
- Presencia de factores de confusión que, de estar presentes, reforzarían aún más el efecto observado.

El valor añadido de GRADE: Lo que distingue a GRADE de otros sistemas de clasificación es que no se centra solo en los aspectos estadísticos o metodológicos, sino en el valor clínico de la evidencia. Esto significa que no basta con saber si un tratamiento funciona; debemos conocer también la confianza que podemos tener en ese resultado para aplicarlo a nuestros pacientes.

Por ejemplo, un metaanálisis puede mostrar que un fármaco reduce la mortalidad con un RR de 0.80. Pero si la certeza es baja, el lector crítico sabrá que esa reducción debe interpretarse con prudencia, pues estudios futuros podrían modificarla. En cambio, si la certeza es alta, la conclusión gana peso para sustentar recomendaciones firmes en una guía clínica.

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

TABLAS SUMMARY OF FINDINGS (SOF)

La aplicación práctica de GRADE se traduce en las tablas SummaryofFindings, que resumen los desenlaces más importantes con su estimación puntual, su intervalo de confianza y el nivel de certeza. Estas tablas, empleadas habitualmente en revisiones Cochrane, ofrecen en una sola página la información más relevante para la toma de decisiones.

Su valor añadido es que presentan tanto medidas relativas (como el RR u OR) como efectos absolutos, expresados en reducción o aumento de eventos por cada 1,000 pacientes tratados. Esto acerca los resultados al terreno clínico: no es lo mismo saber que un fármaco reduce el riesgo un 20 % (medida relativa), que entender que esto implica 20 eventos menos por cada 1,000 pacientes (medida absoluta).

En síntesis: GRADE transforma la lectura de un metaanálisis en un ejercicio más clínico que estadístico. Invita al lector a preguntarse no solo si el tratamiento funciona, sino qué tan seguro puedo estar de ese resultado y qué significa en términos de beneficio real para mis pacientes.

Como ejemplo se puede mencionar:

Tabla 8. SummaryofFindings (SoF): inhibidores de SGLT2 frente a placebo en eventos cardiovasculares mayores (ejemplo ficticio)

Desenlace crítico	Riesgo con placebo (por 1,000)	Riesgo con iSGLT2 (por 1,000)	Efecto relativo (RR, IC95%)	Reducción absoluta (por 1,000)	Certeza de la evidencia (GRADE)
Eventos cardiovasculares mayores (MACE: muerte CV, IAM, ACV)	100	80	0.80 (0.72–0.89)	–20 (–28 a –11)	Alta
Hospitalización por insuficiencia cardíaca	80	60	0.75 (0.66–0.86)	–20 (–27 a –11)	Moderada (inconsistencia leve)
Mortalidad por todas las causas	120	100	0.83 (0.74–0.94)	–20 (–31 a –7)	Moderada (imprecisión)
Eventos adversos graves	150	155	1.03 (0.95–1.12)	+5 (–8 a +18)	Alta

Fuente: Elaboración propia.

Para su interpretación: La tabla 8 resume los desenlaces clínicamente más relevantes del metaanálisis ficticio. Se observa que los iSGLT2 reducen de manera significativa los eventos cardiovasculares mayores, con una certeza alta, lo que otorga gran confianza al resultado. La hospitalización por insuficiencia cardíaca y la mortalidad por todas las causas también se reducen, aunque la certeza es moderada, debido a cierta heterogeneidad y a intervalos de confianza más amplios.

En cuanto a la seguridad, los eventos adversos graves no mostraron diferencias significativas entre iSGLT2 y placebo, con certeza alta, lo que aporta tranquilidad al considerar el balance riesgo-beneficio.

En conjunto, esta tabla ejemplifica la utilidad de los formatos SummaryofFindings: condensan la evidencia cuantitativa y la calidad de la misma en un formato claro, permitiendo al clínico valorar no solo si un fármaco funciona, sino también qué tan sólida es la confianza en ese resultado y cuán aplicable es en la práctica.

6.7 ERRORES COMUNES DE INTERPRETACIÓN

La lectura crítica de revisiones sistemáticas y metaanálisis exige no solo conocer la metodología, sino también reconocer las trampas interpretativas más frecuentes. Identificarlas ayuda al lector a evitar conclusiones erróneas y a valorar con justicia la solidez de la evidencia.

Confundir significancia estadística con relevancia clínica: Un error habitual es considerar que un resultado estadísticamente significativo es automáticamente clínicamente importante. Por ejemplo, un metaanálisis puede mostrar que un fármaco reduce la hemoglobina glucosilada en 0.3 %, con un intervalo de confianza estrecho y un valor $p < 0.001$. Sin embargo, si la mínima diferencia clínicamente importante (MCID) para este desenlace es de 0.5 %, el beneficio existe en números, pero difícilmente sea percibido por el paciente o cambie su pronóstico.

Ignorar la heterogeneidad: La presencia de heterogeneidad elevada ($I^2 > 50$ %) debería despertar cautela, pues indica que los estudios no son completamente comparables. Un error frecuente es interpretar el estimador combinado como una “verdad universal”, sin reparar en que los resultados pueden variar entre contextos. En estos casos, la síntesis narrativa, los subgrupos o

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

la metarregresión cobran más importancia que el valor global.

Tomar los rankings de NMA como verdades absolutas: En los metaanálisis en red, los rankings de eficacia (ej. SUCRA) resultan muy atractivos porque ordenan tratamientos en forma de lista. Sin embargo, un error común es asumir que el primero en el ranking es siempre la mejor opción clínica. En realidad, la diferencia entre los dos primeros puede ser mínima o clínicamente irrelevante, y la certeza de la evidencia puede ser baja. Los rankings deben interpretarse como una guía orientativa, no como un resultado definitivo.

Pasar por alto el riesgo de sesgo y la calidad de los estudios: Un metaanálisis puede incluir decenas de ensayos, pero si la mayoría tiene alto riesgo de sesgo, el resultado combinado puede ser poco confiable. Es un error frecuente fijarse únicamente en el número de estudios sin atender a su calidad. El sistema GRADE y las herramientas como RoB 2 ayudan a contextualizar este problema.

Sobrevalorar la estabilidad de los resultados: Otro error común es asumir que un metaanálisis bien realizado siempre ofrece conclusiones sólidas. Sin embargo, el análisis de sensibilidad puede revelar que la exclusión de un solo estudio cambia radicalmente los resultados. Ignorar este aspecto puede llevar a una confianza excesiva en hallazgos frágiles.

Olvidar la aplicabilidad clínica: No todo resultado de un metaanálisis es automáticamente aplicable al paciente que tenemos enfrente. La indirectitud (diferencias entre la población estudiada y la nuestra) puede limitar la utilidad de la evidencia. Por ejemplo, un tratamiento probado en hospitales de alta complejidad puede no mostrar el mismo efecto en centros rurales con recursos limitados.

En síntesis: los metaanálisis son herramientas poderosas, pero no infalibles. El lector crítico debe recordar que significancia no equivale a relevancia clínica, que la heterogeneidad puede invalidar un estimador global, que los rankings son orientativos y que la aplicabilidad depende siempre del contexto.

7. SESGOS Y LIMITACIONES

Es fundamental recordar que no son inmunes a los problemas metodológicos ni a los sesgos. De hecho, su interpretación crítica requiere que el lector tenga presentes las limitaciones que pueden afectar tanto a la validez interna como a la aplicabilidad de sus resultados.

Uno de los problemas más importantes es el **sesgo de publicación**. Se sabe que los estudios con resultados positivos o estadísticamente significativos tienen mayor probabilidad de ser publicados en revistas de alto impacto, mientras que los estudios con resultados negativos o neutros tienden a permanecer inéditos o en repositorios poco accesibles. Esto genera un desequilibrio que favorece artificialmente los efectos beneficiosos de una intervención. Herramientas como el funnelplot o la prueba de Egger que vimos en acápites previos permiten detectar la presencia de este sesgo, y métodos de ajuste como el trim and fill ofrecen una aproximación de lo que sucedería si los estudios faltantes se incorporaran.

Otro aspecto crítico es el **sesgo de selección**, que puede ocurrir si la estrategia de búsqueda bibliográfica es limitada. Por ejemplo, una revisión que solo utiliza PubMed y restringe la búsqueda al idioma inglés corre el riesgo de dejar fuera estudios relevantes publicados en otras bases de datos o en idiomas distintos. La omisión de literatura gris (tesis, informes técnicos, actas de congresos) puede también contribuir a este sesgo, reduciendo la exhaustividad del análisis. De ahí la importancia de registrar un protocolo previo, donde la estrategia de búsqueda quede claramente definida.

La **heterogeneidad** clínica y metodológica representa otra limitación frecuente. En muchas ocasiones, los estudios incluidos difieren en características de los pacientes, dosis de la intervención, comparadores o definiciones de desenlaces. Cuando estas diferencias son notables, el estimador global puede resultar engañoso. Un ejemplo clásico se observa en los metaanálisis de prevención cardiovascular, donde mezclar poblaciones de bajo riesgo con pacientes de alto riesgo puede diluir el efecto real y dar lugar a conclusiones erróneas.

También es necesario considerar el **sesgo de reporte selectivo**. Algunos estudios primarios informan solo aquellos desenlaces que muestran un beneficio, omitiendo resultados desfavorables o neutros. Si los revisores no contrastan los artículos con los protocolos originales o los registros de ensayos clínicos, este sesgo pasa desapercibido y contamina el metaanálisis.

Más allá de los sesgos, existen **limitaciones metodológicas** que deben tenerse en cuenta. Un metaanálisis no puede superar la calidad de los estudios que lo integran: si estos son pequeños, con alto riesgo de sesgo o metodológicamente débiles, la estimación global será igualmente

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

frágil. Del mismo modo, la imprecisión estadística, reflejada en intervalos de confianza amplios o en un bajo número de eventos, disminuye la certeza de las conclusiones. Incluso las decisiones estadísticas, como optar entre un modelo de efectos fijos o uno de efectos aleatorios, pueden cambiar de forma importante los resultados.

Finalmente, es indispensable reflexionar sobre las **limitaciones prácticas**. Un metaanálisis puede mostrar un beneficio convincente en un escenario ideal, pero ser de poca aplicabilidad en el contexto real del lector. Por ejemplo, un metaanálisis que evalúa procedimientos quirúrgicos realizados en hospitales de excelencia de países desarrollados podría no reflejar la misma magnitud de beneficio en hospitales de segundo nivel de América Latina, donde las condiciones de infraestructura, entrenamiento del personal y disponibilidad de recursos son diferentes.

En resumen, los sesgos y limitaciones de las revisiones sistemáticas y metaanálisis nos recuerdan que, aunque se trate de estudios de máximo nivel de evidencia, no deben interpretarse como verdades absolutas. El lector crítico debe preguntarse siempre: ¿existen estudios que no fueron publicados? ¿la búsqueda bibliográfica fue exhaustiva? ¿los estudios incluidos son comparables entre sí? ¿la calidad metodológica es adecuada? y, sobre todo, ¿estos resultados son aplicables a mis pacientes? Solo con este ejercicio reflexivo se puede aprovechar todo el valor de la síntesis de evidencia sin caer en interpretaciones simplistas o engañosas.

8. COMPARACIÓN CON OTROS ESTUDIOS

Un aspecto esencial en la lectura crítica de revisiones sistemáticas y metaanálisis es la comparación con trabajos previos que hayan abordado la misma pregunta clínica o temas relacionados. Este ejercicio permite al lector ubicar la nueva evidencia en el contexto del conocimiento acumulado, identificar consistencias o discrepancias y comprender mejor las razones detrás de los hallazgos.

En algunos campos, puede haber múltiples revisiones sistemáticas publicadas con pocos años de diferencia. Aunque todas partan de una misma pregunta clínica, los resultados pueden variar por diferencias metodológicas: la amplitud de la búsqueda bibliográfica, los criterios de inclusión, el manejo de la heterogeneidad, el tipo de análisis estadístico o el enfoque en determinados subgrupos. Un lector crítico debe aprender a detectar estos matices, porque de ellos depende la interpretación final.

Por ejemplo, en el caso de los inhibidores de SGLT2 en pacientes con diabetes tipo 2, una revisión publicada en 2018 mostró una reducción significativa de la hospitalización por insuficiencia cardíaca, pero no de la mortalidad por todas las causas. Años más tarde, con la incorporación de nuevos ensayos clínicos, una revisión de 2021 encontró también un beneficio claro en mortalidad, lo que cambió la percepción del impacto clínico de estos fármacos. Este contraste ilustra cómo el tiempo y la acumulación de estudios modifican las conclusiones de los metaanálisis.

También es posible encontrar discrepancias entre revisiones que utilizan metodologías diferentes. Una revisión Cochrane, por su naturaleza estricta, puede excluir estudios de menor calidad y reportar un efecto más conservador. En cambio, una revisión narrativa o un metaanálisis con criterios más flexibles podría incluir mayor cantidad de estudios y sugerir un efecto más pronunciado, pero con menor certeza.

La comparación no debe limitarse a la magnitud del efecto, sino también a la certeza de la evidencia y a la aplicabilidad clínica. Revisiones realizadas en contextos de alta especialización pueden mostrar resultados distintos a los observados en entornos de atención primaria. Asimismo, revisiones centradas en desenlaces duros (mortalidad, eventos cardiovasculares) pueden diferir de aquellas que priorizan desenlaces intermedios o de calidad de vida.

En síntesis, comparar una revisión sistemática o metaanálisis con otros estudios previos no es un ejercicio anecdótico, sino una parte fundamental de la lectura crítica. Permite al lector valorar si la nueva evidencia confirma lo ya conocido, si corrige limitaciones de estudios anteriores o si plantea preguntas aún sin resolver.

9. CHEKLIST

A lo largo de los últimos veinte años se han desarrollado múltiples herramientas y declaraciones que establecen estándares para el reporte, la calidad metodológica y la evaluación de la certeza de la evidencia. Este checklist resume las principales guías y recursos actualizados (PRISMA, AMSTAR, GRADE, entre otros) y está diseñado como una hoja de ruta práctica para el

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

lector. No reemplaza la lectura completa de cada documento, pero ofrece un marco confiable para identificar fortalezas y debilidades en cualquier revisión publicada.

Tabla 9. Principales guías y herramientas internacionales para lectura crítica de revisiones sistemáticas y metaanálisis

Área	Guía / Recurso	Año (últ. act.)	Aplicación práctica en la lectura crítica
Registro y protocolos	PROSPERO	2011–2025	Confirma si la revisión fue registrada de forma prospectiva para evitar sesgo selectivo y duplicación.
	PRISMA-P	2015	Evalúa si el protocolo describe claramente la pregunta, criterios de inclusión y plan de análisis.
	OSF Registries	–	Alternativa abierta de registro cuando PROSPERO no aplica.
Reporte principal	PRISMA 2020	2021	Comprueba transparencia en métodos, selección de estudios, síntesis y reporte de resultados (27 ítems).
	PRISMA-S	2021	Revisa si la estrategia de búsqueda bibliográfica es reproducible y completa.
	SWiM	2020	Evalúa la calidad de la síntesis narrativa cuando no es posible realizar metaanálisis.
Extensiones PRISMA	PRISMA-DTA	2018–2021	Revisa RS de pruebas diagnósticas: exactitud, HSROC, modelos bivariados.
	PRISMA-NMA	2015	Evalúa la transparencia en metaanálisis en red y supuestos de consistencia.
	PRISMA-ScR	2018	Valora revisiones de alcance (scoping reviews), justificación y cobertura.
	PRISMA-IPD	2015	Para metaanálisis de datos individuales, garantiza reporte sobre obtención y armonización de datos.
	MOOSE	2000	Complementa PRISMA en revisiones de estudios observacionales.
Riesgo de sesgo en estudios primarios	RoB 2	2019–2021	Juzga el riesgo de sesgo en ECA, dominio por dominio y desenlace por desenlace.
	ROBINS-I	2016	Evalúa sesgo en estudios no aleatorizados, con especial énfasis en confusión y comparabilidad.
	QUADAS-2	2011	Revisa estudios de pruebas diagnósticas (pacientes, prueba índice, referencia, tiempo).
	QUADAS-C	2022	Nueva herramienta para estudios que comparan directamente pruebas diagnósticas.
Calidad de la revisión	AMSTAR 2	2017	Valora la calidad metodológica de la revisión completa (16 dominios).
	ROBIS	2016	Examina el riesgo de sesgo de la propia revisión sistemática.
Certeza y aplicabilidad	GRADE + SoF	2023–2024	Clasifica la certeza de la evidencia (alta–muy baja) y resume resultados en tablas de desenlaces.
	CINeMA	2020	Estima confianza en comparaciones de metaanálisis en red (6 dominios).

Fuente: Elaboración propia.

10. UTILIDAD

Su utilidad práctica trasciende el ámbito académico y repercute directamente en la atención clínica, en la elaboración de guías de práctica y en la toma de decisiones en salud pública.

En la **práctica clínica individual**, estos estudios ayudan al médico a responder preguntas frecuentes: ¿cuál es el mejor tratamiento para reducir eventos cardiovasculares en pacientes con diabetes? ¿qué prueba diagnóstica ofrece mayor exactitud para detectar una enfermedad en fase temprana? En lugar de basarse en la experiencia personal o en un solo ensayo, el profesional cuenta con un panorama amplio que reúne múltiples estudios, ponderando su calidad y consistencia.

En el terreno de la **salud pública**, los metaanálisis son una base sólida para la planificación de programas sanitarios, la evaluación de nuevas tecnologías y la asignación de recursos. Al condensar grandes cantidades de datos, permiten estimar el impacto poblacional de una intervención y su relación costo-beneficio.

En **investigación**, las revisiones sistemáticas cumplen otra función valiosa: identificar lagunas en el conocimiento. Señalan qué preguntas han sido respondidas con suficiente certeza

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

y cuáles requieren estudios adicionales. Por ejemplo, una revisión puede mostrar que existe evidencia robusta para adultos jóvenes, pero insuficiente en adultos mayores o en poblaciones de bajos ingresos, lo que abre nuevas líneas de investigación.

Finalmente, su utilidad se refleja en la **práctica normativa**. Organismos internacionales como la OMS, el NICE en el Reino Unido y el USPSTF en Estados Unidos utilizan revisiones sistemáticas y metaanálisis como insumo principal para la elaboración de recomendaciones clínicas. Esto se debe a que proporcionan un marco transparente, reproducible y basado en evidencia para orientar políticas de salud y guías clínicas.

Mensaje clave para el lector: Las revisiones sistemáticas y los metaanálisis son más que un ejercicio académico; son una herramienta que conecta la investigación con la práctica clínica y las políticas sanitarias. Su verdadera utilidad radica en que permiten tomar decisiones mejor fundamentadas, más seguras para los pacientes y más eficientes para los sistemas de salud.

11. PREGUNTAS DE AUTOEVALUACIÓN

PREGUNTAS DE SELECCIÓN MULTIPLE

1. ¿Cuál es la principal ventaja de un metaanálisis frente a un ensayo clínico individual?
a) Permite generar hipótesis nuevas
b) Aumenta la potencia estadística al combinar estudios
c) Elimina por completo el sesgo de publicación
d) Evita la necesidad de evaluación crítica
2. El funnelplot se utiliza principalmente para:
a) Detectar heterogeneidad clínica
b) Identificar sesgo de publicación
c) Estimar el efecto combinado
d) Evaluar el riesgo de sesgo de cada estudio
3. En un forestplot de metaanálisis, el diamante al final del gráfico representa:
a) El intervalo de confianza más amplio
b) La heterogeneidad entre los estudios
c) La estimación global combinada
d) El estudio de mayor peso
4. Según el sistema GRADE, ¿qué significa que la certeza de la evidencia sea “baja”?
a) Que los resultados son irrelevantes clínicamente
b) Que es poco probable que futuros estudios cambien las conclusiones
c) Que el efecto real puede diferir sustancialmente del estimado
d) Que la revisión fue mal realizada
5. En el contexto de revisiones sistemáticas de pruebas diagnósticas, un valor de LR– < 0.1 indica:
a) Que la prueba es inútil para descartar enfermedad
b) Que la prueba es muy poderosa para descartar enfermedad
c) Que la prueba confirma fuertemente la enfermedad
d) Que existe heterogeneidad significativa

PREGUNTAS ABIERTAS

6. Explique por qué la significancia estadística en un metaanálisis no siempre implica relevancia clínica.
7. ¿Qué diferencias existen entre un metaanálisis en red (NMA) y un metaanálisis convencional?
8. Mencione tres estrategias de análisis de sensibilidad en metaanálisis y explique su utilidad.
9. ¿Qué aporta la tabla SummaryofFindings (SoF) al lector clínico que no ofrecen los forestplots por sí solos?
10. Describa brevemente cómo la indirectitud puede limitar la aplicabilidad de un metaanálisis.

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

RESPUESTAS

1. b). El metaanálisis integra los resultados de múltiples estudios, lo que incrementa el tamaño muestral efectivo y mejora la precisión de las estimaciones. Esto permite detectar efectos que podrían no ser evidentes en estudios individuales.

2. b). El funnelplot evalúa la distribución de los estudios según su tamaño y efecto. La asimetría sugiere la ausencia de estudios, generalmente negativos, lo que indica posible sesgo de publicación.

3. c). El diamante representa el efecto global del metaanálisis; su centro indica la estimación puntual y su amplitud el intervalo de confianza combinado.

4. c). Una certeza baja en GRADE implica que existe incertidumbre importante y que futuros estudios pueden modificar significativamente el resultado.

5. b). Un LR– < 0.1 reduce considerablemente la probabilidad de enfermedad tras un resultado negativo, lo que hace a la prueba útil para descartar diagnóstico.

6. La significancia estadística indica que un resultado es poco probable que se deba al azar, pero no necesariamente implica relevancia clínica. Un efecto puede ser estadísticamente significativo pero de magnitud pequeña, sin alcanzar la mínima diferencia clínicamente importante (MCID), por lo que no genera un beneficio real para el paciente.

7. El metaanálisis convencional compara dos intervenciones de forma directa, mientras que el metaanálisis en red permite comparar múltiples intervenciones simultáneamente, integrando evidencia directa e indirecta. Esto permite estimar comparaciones no evaluadas directamente y generar rankings de tratamientos.

8. La exclusión de estudios con alto riesgo de sesgo permite evaluar si los resultados dependen de estudios de baja calidad. El análisis leave-one-out identifica si algún estudio influye de forma desproporcionada en el resultado global. La comparación entre modelos de efectos fijos y aleatorios permite evaluar si las conclusiones dependen del modelo estadístico utilizado. Estas estrategias ayudan a determinar la robustez de los resultados.

9.La tabla SummaryofFindings integra el tamaño del efecto, los intervalos de confianza y la certeza de la evidencia (GRADE), además de presentar resultados en términos absolutos. Esto facilita la interpretación clínica, a diferencia del forestplot, que se centra únicamente en la estimación estadística.

10. La indirectitud ocurre cuando los estudios incluidos no representan adecuadamente la población, intervención o desenlace de interés. Esto limita la aplicabilidad de los resultados, ya que las conclusiones pueden no ser extrapolables al contexto clínico real del paciente.

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

REFERENCIAS

Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. Available at: https://doi.org/10.1136/bmj.n71.
Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. Cochrane handbook for systematic reviews of interventions. Version 6.5 [updated Aug 2024]. Cochrane; 2024. Available at: https://www.cochrane.org/handbook.
Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews. J Med Libr Assoc. 2021;109(2):174–200. Available at: https://doi.org/10.5195/jmla.2021.962.
Campbell M, McKenzie JE, Sowden A, Katikireddi SV, Brennan SE, Ellis S, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. BMJ. 2020;368:l6890. Available at: https://doi.org/10.1136/bmj.l6890.
Nikolakopoulou A, Higgins JPT, Papakonstantinou T, Chaimani A, Giovane CD, Egger M, et al. CINeMA: An approach for assessing confidence in the results of a network meta-analysis. PLoS Med. 2020;17(4):e1003082. Available at: https://doi.org/10.1371/journal.pmed.1003082.
McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, Clifford T, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): The PRISMA-DTA statement. JAMA. 2018;319(4):388–96. Available at: https://doi.org/10.1001/jama.2017.19163.
Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions. BMJ. 2017;358:j4008. Available at: https://doi.org/10.1136/bmj.j4008.
Whiting P, Savović J, Higgins JPT, Caldwell DM, Reeves BC, Shea B, et al. ROBIS: A new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol. 2016;69:225–34. Available at: https://doi.org/10.1016/j.jclinepi.2015.06.005.
DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7(3):177–88. Available at: https://doi.org/10.1016/0197-2456(86)90046-2.
Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997;315(7109):629–34. Available at: https://doi.org/10.1136/bmj.315.7109.629.

https://doi.org/10.62486/978-9915-9928-3-9.Ch16

Chapter 17 / Capítulo 17

Critical reading of scientific articles on health: A practical guide (Spanish Edition)
ISBN: 978-9915-9928-3-9
DOI: 10.62486/978-9915-9928-3-9.Ch17
Pages: 299-310

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Non-Systematized Evidence Sources in Medicine

Fuentes de evidencia no sistematizada en medicina

Jorge Marquez Molina
Maria Tereza Nieto Coronel

1 Hospital del Norte. Cochabamba, Bolivia.
2 Centro Oncologico Oncopalia-MyAMedic. La Paz, Bolivia.

ABSTRACT

Introduction: Non-systematized evidence sources occupy the lower rungs of the Evidence-Based Medicine (EBM) pyramid. Nevertheless, they represent indispensable information-transfer resources in daily clinical practice, outpatient consultations, or emergency departments when recent systematic reviews or clinical practice guidelines are unavailable. Their primary value lies in providing a broad, agile, and rapid theoretical transition regarding innovative therapies or complex clinical scenarios enveloped in marked methodological uncertainty. Development: Guided by the clinical crossroads of an internist managing a patient with long-standing, refractory rheumatoid arthritis interested in a cutting-edge biologic approved in Europe but lacking consolidated systematic syntheses, this chapter examines non-systematized evidence frameworks in a detailed manner. The distinct operational profiles and specific clinical utilities of narrative reviews, point-of-care evidence summaries (such as UpToDate, DynaMed, and BMJ Best Practice), update articles in high-impact journals, academic editorials, letters to the editor, and key opinion leader expert testimony are analyzed. The text details the inherent biases within these formats—predominantly literature selection bias, authority bias, and selective reporting driven by the author's subjective judgment. To mitigate these vulnerabilities, critical appraisal guidelines based on the analytical triangulation of sources are introduced alongside the application of current international checklists and tools, such as the SANRA scale for the formal appraisal of narrative reviews, the GRADE framework for balancing the strength of provisional recommendations, and the consensus standards of the ICMJE, COPE, and WAME directed toward the rigorous auditing of conflicts of interest (COI). Conclusions: Non-systematized sources act as pragmatic transitional tools that facilitate shared decision-making and the resolution of immediate inquiries in clinical settings. However, the healthcare professional must interpret them through a strict critical lens, carefully weighing their applicability and treating them as provisional directives bound to continuous re-evaluation as robust scientific knowledge matures.

Keywords: Non-systematized evidence; Narrative review; Evidence summary; Expert opinion; Critical appraisal; Conflicts of interest

RESUMEN

Introducción: Las fuentes de evidencia no sistematizada ocupan los peldaños inferiores de la pirámide de la Medicina Basada en la Evidencia (MBE). Sin embargo, representan recursos de transferencia informacional indispensables en la práctica clínica diaria, la consulta externa o los servicios de urgencia cuando no se dispone de revisiones sistemáticas o guías de práctica clínica recientes. Su valor primordial radica en proveer una visión panorámica, ágil y de rápida transición teórica sobre terapias innovadoras o escenarios clínicos complejos envueltos en una marcada incertidumbre metodológica. Desarrollo: Partiendo de la encrucijada clínica de una internista ante un paciente con artritis reumatoide refractaria interesado en un biológico de última generación comercializado en Europa pero sin síntesis sistemática consolidada, este capítulo examina pormenorizadamente los formatos de evidencia no sistematizada. Se analizan de forma operativa el perfil y las utilidades específicas de las revisiones narrativas, los sumarios de evidencia en el punto de cuidado (como UpToDate, DynaMed y BMJ Best Practice), los artículos de actualización en revistas de alto impacto, las editoriales académicas, las cartas al editor y la opinión de expertos de líderes de opinión. El texto detalla los sesgos inherentes a estos formatos predominando el sesgo de selección bibliográfica, el sesgo de autoridad y el de reporte selectivo por el criterio subjetivo del autor—. Para mitigar estas vulnerabilidades, se introduce pautas de lectura crítica basadas en la triangulación analítica de la información y la aplicación de checklists e instrumentos internacionales vigentes, tales como la escala SANRA para el escrutinio formal de revisiones narrativas, el marco GRADE para balancear la fuerza de recomendaciones provisionales y las directrices del ICMJE, COPE y WAME orientadas a la fiscalización exhaustiva de los conflictos de interés (COI). Conclusiones: Las fuentes no sistematizadas actúan como herramientas de transición pragmática que facilitan la toma de decisiones compartida y la resolución de dudas inmediatas. No obstante, el profesional de la salud debe interpretarlas bajo un riguroso tamiz crítico, ponderando su aplicabilidad y asumiéndolas como directrices provisorias supeditadas a una reevaluación constante conforme evolucione el conocimiento científico robusto.

Palabras clave: Evidencia no sistematizada; Revisión narrativa; Sumario de evidencia; Opinión de expertos; Lectura crítica; Conflictos de interés

1. CASO CLÍNICO

Una médico internista recibe a un paciente de 68 años con artritis reumatoide de larga data, que consulta por síntomas persistentes a pesar de múltiples líneas de tratamiento convencional (metotrexato, leflunomida y anti-TNF). El paciente escuchó sobre un nuevo biológico "de última generación" aprobado en Europa hace pocos meses y pregunta si podría acceder a él en Bolivia.

La medico busca en PubMed revisiones sistemáticas recientes sobre este nuevo fármaco, pero no encuentra ninguna. Solo identifica:

- Una revisión narrativa publicada en una revista europea de reumatología.
- Un artículo de actualización en NEJM sobre terapias biológicas emergentes.
- Una opinión de expertos en Lancet Rheumatology.
- Una sección de UpToDate con recomendaciones preliminares, actualizada hace dos meses.
- Varias cartas al editor que discuten limitaciones metodológicas de los primeros ensayos clínicos.

Ante esta situación, debe decidir cómo jerarquizar y criticar estas fuentes para responder al paciente, considerando la incertidumbre y la aplicabilidad clínica en su contexto.

2. OBJETIVOS DE ESTE TIPO DE ESTUDIOS

Los principales objetivos de utilizar y aprender a interpretar críticamente la evidencia no sistematizada son los siguientes:

Reconocer su lugar en la jerarquía de evidencia: comprender que no sustituyen a revisiones sistemáticas ni a guías basadas en evidencia, pero que pueden ser útiles en ausencia de estas.

Acceder a información rápida y actualizada: aprovechar revisiones narrativas, sumarios de evidencia (UpToDate, DynaMed, BMJ Best Practice), artículos de actualización y editoriales cuando se requiere una visión panorámica o de emergencia clínica.

Identificar tendencias emergentes y nuevas hipótesis: detectar innovaciones, debates científicos y áreas de incertidumbre que aún no cuentan con evidencia consolidada.

Guiar decisiones clínicas en contextos de incertidumbre: aplicar estas fuentes como apoyo cuando no existen estudios sistemáticos recientes, siempre con juicio crítico y comunicación clara al paciente.

Desarrollar habilidades de evaluación crítica: entrenarse en el uso de micro-checklists para detectar sesgos, conflictos de interés, limitaciones metodológicas y aplicabilidad al paciente y al contexto local.

Fomentar la prudencia clínica: aprender a usar estas fuentes como orientación preliminar y no como justificación única para modificar la práctica.

Promover la toma de decisiones compartida: integrar la información con la experiencia del clínico y las preferencias del paciente, dejando explícita la necesidad de actualizar la decisión cuando surja evidencia de mayor calidad.

3. INTRODUCCIÓN

En la Medicina Basada en la Evidencia, las revisiones sistemáticas y metaanálisis constituyen el nivel más alto de síntesis científica. Sin embargo, en la práctica diaria no siempre contamos con este tipo de evidencia de forma inmediata. En múltiples escenarios clínicos y académicos —emergencia, consulta externa, discusión de casos complejos o aparición de nuevos tratamientos—, los médicos deben recurrir a fuentes de evidencia no sistematizada.

Estas fuentes incluyen las revisiones narrativas, los sumarios de evidencia en el punto de cuidado (como UpToDate, DynaMed, BMJ Best Practice), los artículos de actualización, las editoriales, las cartas al editor y la opinión de expertos. Aunque no siguen una metodología rigurosa de búsqueda y análisis como las revisiones sistemáticas, aportan valor al:

- Proporcionar información rápida y accesible.
- Ofrecer una visión panorámica de temas emergentes.
- Señalar nuevas hipótesis o controversias que aún no cuentan con suficiente investigación.

El desafío para el medico no es solo identificar estas fuentes, sino saber cómo evaluarlas críticamente y cómo integrarlas a la toma de decisiones sin sobreestimar su peso en la jerarquía de la evidencia. Un editorial bien argumentado puede estimular la reflexión, pero no debe reemplazar un ensayo clínico; una revisión narrativa puede orientar, pero su utilidad depende de reconocer los sesgos que la atraviesan.

https://doi.org/10.62486/978-9915-9928-3-9.Ch17

En este capítulo, aprenderemos a distinguir las diferentes fuentes no sistematizadas, a analizarlas con herramientas prácticas de evaluación crítica, y a utilizarlas en el marco de un proceso clínico que combine juicio experto, contexto local y preferencias del paciente, siempre conscientes de sus limitaciones y de la necesidad de actualizar nuestras decisiones conforme aparezca evidencia de mayor calidad.

4. METODOLOGÍA DE LOS ESTUDIOS

4.1 REVISIONES NARRATIVAS

Las revisiones narrativas constituyen una de las formas más tradicionales de síntesis del conocimiento médico. Su finalidad principal es ofrecer al lector una visión general sobre un tema determinado, integrando distintos hallazgos de la literatura científica en un texto de carácter más discursivo que metodológico. A diferencia de las revisiones sistemáticas, las narrativas no siguen protocolos predefinidos de búsqueda ni criterios formales de selección y evaluación de estudios. Esto significa que la inclusión de la literatura depende en gran medida del criterio, la experiencia y las preferencias del autor o del grupo de autores.

Desde un punto de vista didáctico, las revisiones narrativas tienen varias ventajas. Pueden proporcionar un panorama amplio y accesible para introducirse en áreas complejas, emergentes o de rápido crecimiento científico. Suelen estar redactadas con un estilo interpretativo, lo que facilita al clínico comprender el contexto de los hallazgos, identificar tendencias emergentes y familiarizarse con nuevas hipótesis. En muchos casos, las revisiones narrativas han sido el primer paso para generar preguntas de investigación que posteriormente dieron origen a estudios primarios o revisiones sistemáticas.

No obstante, la principal limitación de este tipo de fuente radica en su alto riesgo de sesgo. La selección de artículos puede no ser representativa, los estudios contradictorios pueden omitirse y la ausencia de una metodología explícita impide la reproducibilidad del proceso. Además, el análisis suele ser cualitativo y depende del juicio del autor, lo que abre la posibilidad de interpretaciones parciales o de énfasis excesivo en ciertos resultados.

Un ejemplo aplicado a nuestro caso clínico es la revisión narrativa que la médica encuentra sobre nuevos biológicos en artritis reumatoide. En el artículo encontrado, el autor tendría que describir de manera detallada los mecanismos de acción del nuevo fármaco, resaltando los resultados positivos de los primeros ensayos clínicos y plantear que podría representar una alternativa prometedora en pacientes refractarios a anti-TNF. Sin embargo, al analizar críticamente el texto, la médica tendría que observar que no se describen los criterios de selección de los estudios, no se mencionan ensayos con resultados menos favorables y tal vez se omite una discusión profunda sobre los efectos adversos a largo plazo. Esto evidencia cómo, aunque útil para comprender el contexto terapéutico, la revisión narrativa puede ofrecer una visión parcial del problema.

En la práctica clínica, las revisiones narrativas son útiles como puerta de entrada a un tema o como actualización preliminar en áreas donde aún no existe suficiente evidencia sistematizada. Sin embargo, el médico debe recordar que no constituyen un nivel alto en la jerarquía de la evidencia. Su valor reside en ofrecer contexto, estimular la reflexión y generar hipótesis, pero nunca en sustituir la evidencia proveniente de estudios primarios de calidad o de síntesis sistemáticas.

4.2 SUMARIO DE EVIDENCIAS EN EL PUNTO DE CUIDADO (UPTODATE, DYNAMED, BMJ BEST PRACTICE)

Los sumarios de evidencia en el punto de cuidado representan una herramienta fundamental para la práctica clínica contemporánea. Se trata de plataformas electrónicas como UpToDate, DynaMed o BMJ Best Practice, diseñadas para proporcionar información médica rápida, práctica y continuamente actualizada. A diferencia de las revisiones narrativas, su enfoque es ofrecer al clínico recomendaciones inmediatas, apoyadas en la literatura más reciente y en la opinión de expertos en cada área.

El proceso de elaboración combina la revisión de artículos relevantes por equipos editoriales especializados con la actualización periódica de contenidos. No obstante, estas fuentes no emplean una metodología sistemática de búsqueda y síntesis, por lo que no alcanzan el mismo nivel de rigurosidad que una revisión sistemática o una guía clínica basada en evidencia.

https://doi.org/10.62486/978-9915-9928-3-9.Ch17

En el caso del paciente con artritis reumatoide interesado en un nuevo biológico, el internista consulta UpToDate. Encuentra un apartado que resume los principales resultados de los ensayos publicados hasta la fecha, acompañado de una recomendación preliminar que señala: "aún no existen revisiones sistemáticas robustas, pero los datos iniciales muestran eficacia en pacientes refractarios a anti-TNF". Este recurso permite a la médica responder de inmediato a la inquietud del paciente, con información práctica y reciente, mientras aclara que se trata de evidencia provisional.

La principal ventaja de estas plataformas es la posibilidad de disponer de información clara y sintetizada durante la consulta, lo que facilita la toma de decisiones clínicas en tiempo real. Sin embargo, su limitación radica en que no siempre incluyen todos los estudios relevantes, dependen de la frecuencia de actualización y pueden reflejar sesgos editoriales según la composición de sus equipos.

En conclusión, los sumarios de evidencia deben entenderse como una herramienta de apoyo inmediato, que orienta al clínico en la práctica diaria pero que debe complementarse con revisiones sistemáticas, guías clínicas y el juicio crítico del médico antes de modificar de manera definitiva la conducta terapéutica.

4.3 ARTÍCULOS DE ACTUALIZACIÓN

Los artículos de actualización publicados en revistas médicas de prestigio tienen como propósito resumir los avances más recientes en un área específica de la medicina. Suelen estar dirigidos a clínicos que necesitan mantenerse al día en campos de rápido desarrollo, como terapias biológicas, oncología, infectología o nuevas tecnologías diagnósticas.

A diferencia de los sumarios en línea, los artículos de actualización son textos extensos, generalmente encargados por los editores a expertos reconocidos. Su construcción implica una revisión de la literatura reciente, con énfasis en los hallazgos considerados más relevantes para la práctica clínica. Sin embargo, no siguen una metodología sistemática de búsqueda ni presentan criterios explícitos de inclusión o exclusión de estudios. Por ello, la selección de la evidencia puede depender de la perspectiva del autor o de la revista, y no necesariamente reflejar un panorama equilibrado.

En el caso del paciente con artritis reumatoide, el médico encuentra en una revista de alto impacto —como el New England Journalof Medicine— un artículo de actualización sobre nuevas terapias biológicas. Allí se describe el nuevo fármaco consultado por el paciente, se destacan los resultados de los ensayos fase II y III más prometedores y se discuten posibles efectos adversos. Este artículo ofrece un panorama claro de la innovación terapéutica y le permite al clínico responder con información concreta sobre el mecanismo de acción, la evidencia preliminar y las expectativas razonables en cuanto a eficacia y seguridad.

La utilidad principal de los artículos de actualización es que sirven como puente entre la literatura primaria y la práctica clínica, brindando un resumen organizado de lo más novedoso en un área. Sin embargo, su limitación reside en la posible omisión de estudios contradictorios, el sesgo de publicación hacia resultados positivos y la falta de una evaluación formal de la calidad de los ensayos incluidos.

En síntesis, los artículos de actualización representan una fuente valiosa cuando se requiere conocer rápidamente lo más reciente en un campo, pero deben ser interpretados como una herramienta complementaria y no como evidencia definitiva para modificar conductas clínicas.

4.4 EDITORIALES

Las editoriales constituyen un género particular dentro de la literatura médica. Generalmente son textos breves, escritos por expertos invitados o por miembros del comité editorial de una revista, que acompañan la publicación de un estudio original o abordan un tema de relevancia actual. Su función principal no es aportar datos nuevos, sino ofrecer un marco interpretativo, reflexivo o crítico sobre la evidencia presentada.

Metodológicamente, las editoriales no siguen un esquema de revisión sistemática ni presentan búsquedas estructuradas de literatura. Su valor radica en la capacidad de contextualizar un hallazgo, señalar sus implicaciones para la práctica clínica y advertir sobre limitaciones o áreas de incertidumbre. Pueden también estimular la discusión científica y abrir líneas de investigación futuras. Sin embargo, al ser textos argumentativos, dependen en gran medida de la visión del

https://doi.org/10.62486/978-9915-9928-3-9.Ch17

autor, lo que los hace susceptibles a sesgos de autoridad o a conflictos de interés, sobre todo si el autor tiene vínculos con la industria farmacéutica.

En relación con nuestro caso clínico, el médico encuentra un editorial en una revista reumatológica internacional que comenta la publicación de los primeros ensayos clínicos sobre el nuevo biológico para artritis reumatoide. El editorial destaca el potencial del fármaco en pacientes refractarios a anti-TNF, pero al mismo tiempo advierte que los estudios disponibles incluyen muestras pequeñas y seguimientos cortos, por lo que recomienda prudencia antes de su uso generalizado. Para el internista, este texto resulta útil porque le ayuda a matizar el entusiasmo inicial generado por el artículo de actualización y a transmitir al paciente un mensaje equilibrado: hay novedades prometedoras, pero aún persiste la necesidad de evidencia más sólida.

La fortaleza de las editoriales reside en su capacidad de interpretar la evidencia en clave crítica, resaltar implicaciones clínicas inmediatas y orientar el debate científico. Sin embargo, su limitación es que no constituyen evidencia en sí mismas: son opiniones razonadas, pero opiniones al fin. Un error frecuente es darles un peso mayor al que realmente tienen dentro de la jerarquía de la evidencia.

En síntesis, las editoriales pueden considerarse como una guía de lectura entre líneas, que ayuda al clínico a ubicar un nuevo hallazgo en su justo contexto. Aportan un contrapeso útil frente a la información más entusiasta de los artículos de actualización, recordándole al médico que todavía existe incertidumbre y que las decisiones deben tomarse con cautela.

4.5 CARTAS AL EDITOR

Las cartas al editor representan una forma ágil de comunicación científica, cuyo propósito es comentar, criticar o complementar artículos recientemente publicados en una revista. Se caracterizan por su brevedad y por el hecho de que, aunque pueden ser sometidas a revisión editorial, rara vez atraviesan un proceso riguroso de revisión por pares como los artículos originales. Por ello, su valor no radica en generar evidencia primaria, sino en estimular el debate académico y en señalar aspectos que, de otro modo, podrían pasar desapercibidos.

Metodológicamente, las cartas no constituyen síntesis de evidencia. Más bien, aportan observaciones puntuales que pueden cuestionar la metodología de un ensayo, advertir sobre limitaciones no discutidas por los autores originales, presentar datos preliminares de otros contextos o incluso describir experiencias clínicas relevantes. De este modo, enriquecen el proceso científico, actuando como un mecanismo de control social y de vigilancia crítica.

En el caso del paciente con artritis reumatoide, el internista identifica que, tras la publicación de los primeros ensayos clínicos del nuevo biológico, varias cartas al editor fueron enviadas a revistas internacionales. En ellas, investigadores de distintos países señalan posibles sesgos de selección en los estudios, discuten la corta duración del seguimiento y plantean dudas sobre la seguridad a largo plazo del fármaco. Al leer estas cartas, la médica obtiene una visión más completa: aunque los resultados iniciales son alentadores, la comunidad científica ya está expresando reservas que deben considerarse antes de recomendar la terapia.

La fortaleza de las cartas al editor es precisamente su capacidad de cuestionar y matizar hallazgos recientes, abriendo espacios de reflexión que enriquecen la interpretación crítica de un artículo. Sin embargo, su limitación es evidente: al ser textos breves y, muchas veces, cargados de opinión, pueden reflejar más las posturas personales de los autores que una evaluación objetiva. Además, carecen de peso en la jerarquía de la evidencia y no deben ser usados como único fundamento para modificar la práctica clínica.

En conclusión, las cartas al editor son un recurso complementario que aporta alertas tempranas y discusión crítica, pero que necesita ser integrado con fuentes de mayor calidad. En el caso de nuestro paciente, le permiten a la médica reconocer que la evidencia sobre el nuevo biológico está aún en fase de escrutinio y que cualquier decisión debe tomarse con cautela, informando al paciente sobre la incertidumbre existente.

4.6 OPINIÓN DE EXPERTOS

La opinión de expertos constituye una de las formas más antiguas de transmisión del conocimiento médico. Se basa en la experiencia clínica acumulada y en la autoridad académica de quienes, tras años de práctica e investigación, emiten recomendaciones sobre diagnóstico, tratamiento o prevención en áreas donde la evidencia científica aún es insuficiente o contradictoria.

https://doi.org/10.62486/978-9915-9928-3-9.Ch17

Aunque se encuentra en el nivel más bajo de la jerarquía de la evidencia, la opinión de expertos sigue desempeñando un papel importante en la práctica clínica, sobre todo en escenarios de incertidumbre o cuando se enfrentan problemas emergentes para los que no existen revisiones sistemáticas ni guías de práctica clínica.

Metodológicamente, la opinión experta no sigue un proceso formal de búsqueda bibliográfica ni de síntesis de estudios. Puede presentarse en diferentes formatos: artículos de revisión firmados por líderes de opinión, editoriales invitadas, consensos preliminares de sociedades científicas o incluso declaraciones en conferencias médicas. Su principal fortaleza es la capacidad de ofrecer orientación práctica inmediata, respaldada por la trayectoria del experto, especialmente cuando se necesita actuar sin demora. Sin embargo, su principal limitación es el alto riesgo de sesgo, ya que las recomendaciones dependen de la interpretación personal del especialista y pueden estar influidas por conflictos de interés, relaciones con la industria farmacéutica o enfoques particulares de una escuela de pensamiento.

En el caso de nuestro paciente con artritis reumatoide, el internista asiste a un congreso internacional de reumatología donde un reconocido investigador presenta su experiencia con el nuevo biológico. Durante su conferencia, afirma que el fármaco podría convertirse en una opción de primera línea para pacientes refractarios en los próximos años, aunque reconoce que aún falta evidencia a largo plazo. Para el médico, escuchar esta opinión resulta útil como orientación preliminar, pero es consciente de que no se trata de evidencia sistematizada, sino de una interpretación basada en la experiencia personal del especialista y en datos aún limitados.

La utilidad de la opinión de expertos radica en que puede guiar decisiones clínicas cuando no existe evidencia de mejor calidad y en que aporta perspectivas prácticas derivadas de la experiencia acumulada. No obstante, el medico debe aprender a integrar esta información con cautela, verificando la coherencia con la literatura publicada y reconociendo explícitamente la incertidumbre frente al paciente.

En síntesis, la opinión de expertos puede considerarse como un recurso provisional, valioso en contextos de escasa evidencia, pero que no debe erigirse como fundamento único para modificar la práctica clínica.

Tabla 1. Características principales de las fuentes de evidencia no sistematizada

Tipo de fuente	Definición	Utilidad clínica	Limitaciones	Ejemplo aplicado al caso clínico (AR y nuevo biológico)
Revisión narrativa	Síntesis amplia sin metodología sistemática, basada en criterio del autor.	Introducirse en un tema, explorar hipótesis o tendencias emergentes.	Alto riesgo de sesgo, omisión de estudios contradictorios, falta de reproducibilidad.	Una revisión narrativa sobre biológicos emergentes en artritis reumatoide que presenta el nuevo fármaco de forma entusiasta.
Sumario de evidencia (UpToDate, DynaMed, BMJ BP)	Plataforma de acceso rápido, con actualizaciones continuas elaboradas por expertos.	Tomar decisiones inmediatas en consulta, acceder a recomendaciones preliminares.	No sistemáticos, posibles sesgos editoriales, depende de frecuencia de actualización.	UpToDate ofrece un resumen sobre el nuevo biológico con una recomendación provisional basada en ensayos fase III.
Artículo de actualización	Publicado en revistas indexadas, resume avances recientes en un área.	Mantenerse al día en innovaciones terapéuticas y diagnósticas.	Omisión de estudios negativos, falta de análisis crítico de calidad de la evidencia.	Un artículo de NEJM describe los principales ensayos del nuevo biológico en pacientes refractarios.
Editorial	Texto breve de reflexión o crítica sobre un artículo o tema relevante.	Contextualizar hallazgos, señalar implicaciones clínicas, advertir sobre incertidumbre.	Opinión del autor, riesgo de conflictos de interés, no es evidencia primaria.	Una editorial en Lancet recomienda prudencia antes de adoptar el nuevo biológico por seguimiento corto de ensayos.
Carta al editor	Comunicación breve que comenta o cuestiona artículos publicados.	Señalar limitaciones metodológicas, generar debate académico.	Opiniones puntuales, no constituyen evidencia, riesgo de sesgo personal.	Investigadores critican en una carta el tamaño muestral reducido y la falta de seguridad a largo plazo.
Opinión de expertos	Recomendaciones basadas en experiencia clínica o liderazgo en el campo.	Orientar en ausencia de evidencia robusta, útil en enfermedades raras/ emergentes.	Muy susceptible a sesgos, dependencia de autoridad, posibles conflictos de interés.	Un experto en congreso sugiere que el biológico será clave en el futuro, aunque advierte que falta más evidencia.

Fuente: Elaborado con apoyo de inteligencia artificial y validación del autor

https://doi.org/10.62486/978-9915-9928-3-9.Ch17

5. EVALUACIÓN DE LA VALIDEZ Y RELEVANCIA

Evaluar la validez de una fuente no sistematizada implica reconocer que, a diferencia de los estudios primarios o las revisiones sistemáticas, aquí no existen protocolos estrictos de búsqueda ni criterios uniformes de calidad. Por ello, el lector debe apoyarse en preguntas clave que le permitan valorar si el contenido es confiable y si resulta relevante para su paciente.

5.1 VALIDEZ

La validez de estas fuentes se relaciona con la calidad de la información presentada y la transparencia del proceso con que fue elaborada. Para ello conviene plantearse:

- Autoría y experticia: ¿Quién escribió el texto? ¿Se trata de un especialista reconocido en el área? ¿Qué trayectoria tiene en investigación y práctica clínica?
- Transparencia de la evidencia citada: ¿La fuente presenta referencias actualizadas y trazables, o se limita a opiniones sin sustento bibliográfico?
- Reconocimiento de limitaciones: ¿El autor declara de manera explícita los vacíos de evidencia, la falta de ensayos robustos o las controversias existentes?
- Conflictos de interés (COI): ¿Existen vínculos con la industria u otras entidades que puedan sesgar las conclusiones?

En nuestro caso clínico, el médico internista podría evaluar la validez de una revisión narrativa sobre el nuevo biológico preguntándose si el autor tiene experiencia real en artritis reumatoide, si presenta referencias actualizadas y si reconoce la ausencia de estudios a largo plazo.

5.2 RELEVANCIA

La relevancia se refiere al grado en que la información puede aplicarse al contexto clínico del paciente. Aquí es importante preguntarse:

- Concordancia con la pregunta clínica: ¿El texto responde directamente al problema del paciente o aborda aspectos tangenciales?
- Contexto y aplicabilidad: ¿Los estudios citados incluyen pacientes similares en edad, comorbilidades o gravedad de la enfermedad? ¿Se realizaron en entornos comparables al nuestro?
- Implicaciones prácticas: ¿Las conclusiones ofrecen recomendaciones que pueden trasladarse a la consulta? ¿O se limitan a hipótesis sin utilidad inmediata?
- Riesgo–beneficio contextualizado: ¿El texto discute riesgos y costos que sean relevantes para el paciente, o solo presenta eficacia teórica?

En el caso de la artritis reumatoide, aunque un artículo de actualización en NEJM describa los resultados del nuevo biológico, la relevancia para la paciente de 68 años dependerá de si esos estudios incluyeron adultos mayores, de si el perfil de seguridad es aplicable a su comorbilidad y de la disponibilidad real del medicamento en Bolivia.

5.3 INTEGRACIÓN PRÁCTICA

La evaluación de validez y relevancia no es un ejercicio abstracto, sino una herramienta para triangular fuentes. Un sumario en UpToDate puede ser válido por estar actualizado y firmado por expertos, pero su relevancia dependerá de si incluye datos aplicables al perfil del paciente. Una editorial puede carecer de validez como evidencia primaria, pero ser relevante porque advierte riesgos que deben discutirse en la consulta.

En síntesis, la validez garantiza que la información sea confiable, mientras que la relevancia asegura que sea útil en el contexto clínico específico. Solo la combinación de ambas dimensiones permite que estas fuentes no sistematizadas aporten valor real a la toma de decisiones.

6. CHEKLIST

El presente cheklist esta armado con base en marcos y políticas internacionales vigentes (SANRA para narrativas; GRADE para valorar certeza y fuerza de recomendaciones; políticas de COI de ICMJE; metodologías editoriales de BMJ Best Practice, UpToDate y DynaMed; y el estándar

https://doi.org/10.62486/978-9915-9928-3-9.Ch17

PRISMA como contraste de transparencia).

Tabla 2. Checklist unificado para evaluar fuentes de evidencia no sistematizada

Bloque / Ítem	Pregunta clave	Referencia internacional
A. Común a todas las fuentes		
1. Fecha y vigencia	¿La publicación/actualización está claramente fechada y es reciente?	BMJ Best Practice, UpToDate, DynaMed (uso de GRADE, historial de actualización)
2. Transparencia bibliográfica	¿Las referencias son suficientes, trazables y actualizadas?	PRISMA 2020 (estándar de transparencia)
3. Conflictos de interés	¿Se declaran relaciones financieras/no financieras?	ICMJE, COPE
4. Equilibrio en la presentación	¿Se mencionan limitaciones, controversias o resultados negativos?	GRADE (dominios: sesgo, publicación)
5. Concordancia externa	¿La conclusión coincide o explica su diferencia con RS/guías?	WHO Guideline Handbook, GRADE
6. Magnitud del efecto	¿Reporta resultados en términos clínicos absolutos, con IC95%?	GRADE (dominio: imprecisión)
7. Consistencia	¿Considera la homogeneidad/heterogeneidad entre estudios citados?	GRADE (dominio: inconsistencia)
8. Relevancia y aplicabilidad	¿Los pacientes/intervenciones son comparables a los tuyos?	GRADE (dominio: indirectitud)
9. Riesgos, costos y factibilidad	¿Se discuten daños, costos y barreras de implementación?	GRADE EtD (evidence-to-decision)
10. Plan de reevaluación	¿La fuente o el clínico anticipa revisar la decisión al aparecer evidencia de mayor calidad?	Buenas prácticas MBE
B. Específico por tipo de fuente	Aspectos clave	Referencia internacional
Revisión narrativa	Claridad del objetivo, importancia del tema, descripción mínima de la búsqueda, referencias suficientes, interpretación equilibrada, reconocimiento de limitaciones.	SANRA scale (Baethge et al., 2019)
Sumarios en el punto de cuidado	Fecha de actualización visible, referencias a RS/guías, uso de GRADE para graduar certeza y fuerza, declaración de COI.	BMJ Best Practice, UpToDate, DynaMed (GRADE framework)
Artículo de actualización	Ventana temporal explícita, criterios pragmáticos de selección, discusión de hallazgos positivos y negativos, vínculo con RS/guías.	PRISMA como contraste de transparencia
Editorial	Anclaje en datos, independencia editorial, transparencia de COI, reconocimiento explícito de incertidumbre.	COPE, WAME
Carta al editor	Crítica metodológica concreta, referencias contrastables, ausencia de afirmaciones exageradas, COI declarado.	COPE
Opinión de expertos	Pertinencia del experto, base diferenciada entre datos y experiencia, coherencia con RS/guías, declaración de COI.	ICMJE, WAME

Fuente: Elaborado con apoyo de inteligencia artificial y validación del autor.

Cómo usarlo:

- Bloque común (todas las fuentes): aplica a cualquier narrativa, sumario, actualización, editorial, carta u opinión experta.
- Bloque específico: añade el mini-checklist según el tipo de fuente.

Escala sugerida (orientativa): 0 = no cumple, 1 = parcialmente, 2 = cumple. La suma no es un instrumento validado, sirve para estructurar el juicio crítico. Para narrativas puedes además aplicar SANRA (6 ítems) de forma explícita.

7. UTILIDAD

Las fuentes de evidencia no sistematizada cumplen un papel fundamental en la práctica clínica, siempre que se reconozca su lugar en la jerarquía de la evidencia y se usen con el debido espíritu crítico. Su utilidad radica en que permiten llenar vacíos de información cuando no existen revisiones sistemáticas, guías clínicas o metaanálisis recientes, y en que facilitan al médico responder preguntas inmediatas que surgen en la consulta o en la guardia.

En el caso de la paciente con artritis reumatoide que consulta por un nuevo biológico, estas fuentes —revisiones narrativas, sumarios de evidencia, artículos de actualización, editoriales,

https://doi.org/10.62486/978-9915-9928-3-9.Ch17

cartas y opiniones de expertos— ofrecen un mapa preliminar del conocimiento disponible. Gracias a ellas, el médico puede:

- Informar con rapidez al paciente, explicando los avances más recientes y la experiencia acumulada hasta el momento.
- Identificar tendencias emergentes, como la llegada de nuevas moléculas o estrategias terapéuticas que aún no cuentan con revisiones sistemáticas.
- Anticipar debates y controversias, a través de cartas al editor o editoriales que señalan limitaciones metodológicas o posibles riesgos de la nueva terapia.
- Guiar la toma de decisiones compartida, reconociendo que las recomendaciones son provisionales y que deberán reevaluarse cuando surja evidencia de mayor calidad.
- Estimular la investigación local, ya que los vacíos detectados en estas fuentes pueden transformarse en preguntas de investigación relevantes para residentes y especialistas.
- Su mayor fortaleza es la accesibilidad y rapidez con la que permiten obtener información práctica. Sin embargo, su utilidad no radica en sustituir a la evidencia sistematizada, sino en complementarla: sirven de puente temporal que orienta al clínico mientras espera datos más sólidos.

En definitiva, las fuentes de evidencia no sistematizada son herramientas de transición: ofrecen contexto, orientación preliminar y discusión crítica, pero requieren ser siempre filtradas mediante lectura crítica, aplicadas con prudencia al paciente individual y reevaluadas cuando aparezca evidencia de mejor calidad.

8. AUTOEVALUACIÓN

PREGUNTAS CERRADAS

1. ¿Cuál de las siguientes fuentes se considera la de menor nivel en la jerarquía de la evidencia, pero puede ser útil en ausencia de revisiones sistemáticas o guías?

a) Ensayo clínico aleatorizado
b) Revisión narrativa
c) Metaanálisis
d) Cohorte prospectiva

2. En el caso clínico de la paciente con artritis reumatoide, ¿qué fuente ofrece información práctica y actualizada para responder al paciente en tiempo real?

a) Carta al editor
b) Sumario en el punto de cuidado (UpToDate, DynaMed, BMJ BP)
c) Editorial
d) Revisión narrativa

3. ¿Qué instrumento internacional se recomienda para evaluar la calidad de una revisión narrativa?

a) PRISMA
b) SANRA
c) CONSORT
d) QUADAS-2

4. ¿Cuál de las siguientes afirmaciones sobre las editoriales es correcta?

a) Son evidencia primaria y pueden modificar la práctica clínica.
b) Constituyen opiniones críticas o reflexivas y deben interpretarse con cautela.
c) Equivalen metodológicamente a revisiones sistemáticas.
d) No requieren declaración de conflictos de interés.

5. ¿Qué aspecto es clave al evaluar la relevancia de una fuente no sistematizada?

a) Si el autor es un experto internacional.
b) Si el artículo está en una revista de alto impacto.
c) Si la población estudiada y las intervenciones son comparables a las del paciente.
d) Si la fuente cita más de 50 referencias.

PREGUNTAS ABIERTAS

6. ¿Cómo explicaría a la paciente con artritis reumatoide la diferencia entre la información obtenida en un sumario de evidencia y la que proviene de una revisión sistemática?

https://doi.org/10.62486/978-9915-9928-3-9.Ch17

7. Mencione dos fortalezas y dos limitaciones de los artículos de actualización.
8. ¿Qué rol juegan las cartas al editor en el proceso de construcción de la evidencia?
9. ¿Qué elementos debe considerar un clínico para determinar si una opinión de experto es aplicable a su paciente?
10. ¿Cómo se debe integrar la información de varias fuentes no sistematizadas cuando se enfrenta un problema clínico sin revisiones sistemáticas ni guías disponibles?

RESPUESTAS

1. b) Porque las revisiones narrativas no siguen una metodología sistemática ni criterios estrictos de selección, lo que las ubica en un nivel bajo de la jerarquía de la evidencia, aunque siguen siendo útiles como aproximación inicial en ausencia de evidencia de mayor calidad.
2. b) Porque los sumarios en el punto de cuidado están diseñados para ofrecer información sintetizada, actualizada y directamente aplicable en la consulta clínica en tiempo real.
3. b) Porque SANRA es una herramienta específicamente desarrollada para evaluar la calidad de revisiones narrativas, considerando su estructura, claridad y validez del contenido.
4. b) Porque las editoriales representan interpretaciones críticas basadas en la opinión de expertos, por lo que deben ser analizadas con cautela y no considerarse evidencia primaria.
5. c) Porque la relevancia depende de la aplicabilidad clínica, es decir, de qué tan comparable es la población, intervención y contexto del estudio con el paciente que se está atendiendo.
6. La posible respuesta es que los sumarios de evidencia ofrecen información práctica, rápida y actualizada para la toma de decisiones inmediatas, pero no siguen un proceso sistemático tan riguroso como una revisión sistemática, la cual evalúa de manera exhaustiva todos los estudios disponibles y proporciona conclusiones con mayor nivel de certeza.
7. La posible respuesta es que las fortalezas incluyen la actualización rápida del conocimiento y la síntesis práctica de avances recientes, mientras que las limitaciones son la posible omisión de estudios contradictorios y la falta de evaluación sistemática de la calidad de la evidencia.
8. La posible respuesta es que las cartas al editor permiten identificar limitaciones metodológicas, cuestionar resultados y fomentar el debate científico, contribuyendo a una interpretación más crítica de la evidencia disponible.
9. La posible respuesta es que se debe considerar la experiencia del experto en el tema, la coherencia de sus recomendaciones con la evidencia existente, la presencia de conflictos de interés y la aplicabilidad de sus conclusiones al contexto clínico del paciente.
10. La posible respuesta es que se debe realizar una triangulación de la información, integrando sumarios de evidencia, artículos de actualización, editoriales y opiniones de expertos, aplicando juicio crítico, considerando la incertidumbre y planificando una reevaluación cuando exista evidencia de mayor calidad.

https://doi.org/10.62486/978-9915-9928-3-9.Ch17

REFERENCIAS

Baethge C, Goldbeck-Wood S, Mertens S. SANRA: A scale for the quality assessment of narrative review articles. Res Integr Peer Rev. 2019;4:5. Available at: https://doi.org/10.1186/s41073-019-0064-8.
Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336(7650):924–6. Available at: https://doi.org/10.1136/bmj.39489.470347.AD.
Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schünemann HJ. What is "quality of evidence" and why is it important to clinicians? BMJ. 2008;336(7651):995–8. Available at: https://doi.org/10.1136/bmj.39490.551019.BE.
Guyatt GH, Oxman AD, Santesso N, Helfand M, Vist G, Kunz R, et al. GRADE guidelines: 3. Rating the quality of evidence—introduction. J Clin Epidemiol. 2011;64(4):401–6. Available at: https://doi.org/10.1016/j.jclinepi.2010.07.015.
Guyatt GH, Oxman AD, Sultan S, Brozek J, Glasziou P, Alonso-Coello P, et al. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol. 2011;64(12):1311–6. Available at: https://doi.org/10.1016/j.jclinepi.2011.06.004.
Guyatt GH, Oxman AD, Sultan S, Brozek J, Glasziou P, Akl EA, et al. GRADE guidelines: 4. Rating the quality of evidence—study limitations (risk of bias). J Clin Epidemiol. 2011;64(4):407–15. Available at: https://doi.org/10.1016/j.jclinepi.2010.07.017.
Guyatt GH, Oxman AD, Vist G, Kunz R, Falck-Ytter Y, Schünemann HJ. Going from evidence to recommendations. BMJ. 2008;336(7652):1049–51. Available at: https://doi.org/10.1136/bmj.39493.646875.AE.
Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. Available at: https://doi.org/10.1136/bmj.n71.
Akl EA, Kennedy C, Konda K, Caceres CF, Horvath T, Ayala G, et al. Using GRADE methodology for the development of public health guidelines for the prevention and treatment of HIV and other STIs among men who have sex with men and transgender people. BMC Public Health. 2012;12:386. Available at: https://doi.org/10.1186/1471-2458-12-386.
Zhang S, Wu QJ, Liu SX. A methodologic survey on use of the GRADE approach in evidence syntheses published in high-impact factor urology and nephrology journals. BMC Med Res Methodol. 2022;22(1):220. Available at: https://doi.org/10.1186/s12874-022-01701-x.
International Committee of Medical Journal Editors. Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals [Internet]. Updated 2025 [cited 2026 Apr 22]. Available at: https://www.icmje.org/recommendations/.
UpToDate. Editorial policy [Internet]. Waltham (MA): UpToDate; 2025 [cited 2026 Apr 22]. Available at: https://www.uptodate.com/home/editorial-policy.
UpToDate. Grading guide [Internet]. Waltham (MA): UpToDate; [cited 2026 Apr 22]. Available at: https://www.uptodate.com/home/grading-guide.
BMJ Best Practice. What is GRADE? [Internet]. London: BMJ Publishing Group; [cited 2026 Apr 22]. Available at: https://bestpractice.bmj.com/info/toolkit/learn-ebm/what-is-grade.
World Association of Medical Editors. Conflict of interest in peer-reviewed medical journals [Internet]. [place unknown]: WAME; [cited 2026 Apr 22]. Available at: https://www.wame.org/conflict-of-interest-in-peer-reviewed-medical-journals/events-listing.php.

Chapter 18 / Capítulo 18

Critical reading of scientific articles on health: A practical guide (Spanish Edition)
ISBN: 978-9915-9928-3-9
DOI: 10.62486/978-9915-9928-3-9.Ch18
Pages: 311-332

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Use of Artificial Intelligence in the Critical Appraisal of Scientific Medical Articles

Uso de la inteligencia artificial en la lectura crítica de artículos científicos médicos

Jhossmar Cristians Auza-Santivañez
Boris Adolfo Llanos Torrico
Rita Marlene Mamani-Limachi
Henrry Temis Quisbert Vasquez

1 Hospital del Gran Chaco "Fray Quebracho" de III nivel. Instituto Académico Científico Quispe-Cornejo. Tarija, Bolivia.
2 Universidad Mayor de San Andrés. Universidad Autónoma "Tomás Frías". Oruro, Bolivia.
3 Hospital Obrero N°1 - Caja Nacional de Salud, La Paz, Bolivia.
4 Hospital Materno Infantil "Germán Urquidi". Cochabamba, Bolivia.

ABSTRACT

Introduction: The incorporation of artificial intelligence (AI), especially large-scale language models (LLMs), has transformed the critical appraisal of scientific literature by facilitating the processing of large volumes of information. However, its use in Evidence-Based Medicine requires recognizing both its strengths and methodological limitations to avoid errors in interpretation and cognitive dependence. Development: This chapter describes a hybrid model of critical appraisal in which AI acts as a support tool for structured data extraction, statistical interpretation, and preliminary bias detection, always subordinate to clinical judgment. It proposes principles for responsible use, analysis algorithms, prompt design, and verification strategies to minimize hallucinations, biases, and methodological errors. Conclusions: AI represents an opportunity to optimize critical appraisal and democratize access to scientific evidence; however, its true value depends on its ethical, reflective, and methodologically rigorous integration. The preservation of critical thinking, the systematic verification of information, and the maintenance of independent clinical judgment are essential conditions for these tools to strengthen, rather than replace, evidence-based decision-making.

Keywords: Artificial Intelligence; Evidence-Based Medicine; Critical Appraisal; Language Models; Clinical Research.; RESUMEN

RESUMEN

Introducción: La incorporación de la inteligencia artificial (IA), especialmente los modelos de lenguaje de gran escala (LLMs), ha transformado la lectura crítica de la literatura científica al facilitar el procesamiento de grandes volúmenes de información. No obstante, su utilización en Medicina Basada en la Evidencia exige reconocer tanto sus fortalezas como sus limitaciones metodológicas para evitar errores de interpretación y dependencia cognitiva. Desarrollo: El capítulo describe un modelo híbrido de lectura crítica en el que la IA actúa como herramienta de apoyo para la extracción estructurada de datos, interpretación estadística y detección preliminar de sesgos, siempre subordinada al juicio clínico. De esta manera, propone principios de uso responsable, algoritmos de análisis, diseño de prompts y estrategias de verificación para minimizar alucinaciones, sesgos y errores metodológicos. Conclusiones. La IA representa una oportunidad para optimizar la lectura crítica y democratizar el acceso a la evidencia científica; sin embargo, su verdadero valor depende de una integración ética, reflexiva y metodológicamente rigurosa. La preservación del pensamiento crítico, la verificación sistemática de la información y el mantenimiento del juicio clínico independiente constituyen condiciones indispensables para que estas herramientas fortalezcan, y no sustituyan, la toma de decisiones basada en evidencia.

Palabras clave: Inteligencia Artificial; Medicina Basada en la Evidencia; Lectura Crítica; Modelos de Lenguaje; Investigación Clínica

1. CASO CLÍNICO

La residente de segundo año de Medicina Interna, en un hospital universitario, recibe durante una guardia nocturna a un paciente masculino de 67 años con insuficiencia cardíaca con fracción de eyección reducida (ICFEr), fibrilación auricular con respuesta ventricular conserva y enfermedad renal crónica.Previo al pase de visita, el médico tratante le solicita revisar la evidencia científica más reciente sobre la eficacia y seguridad de la combinación sacubitrilo/valsartán en pacientes con este perfil de comorbilidades. Debido a las limitaciones de tiempo propias del contexto asistencial, la residente identifica un ensayo clínico aleatorizado publicado el año previo en una revista biomédica indexada. Ante la extensión del manuscrito y la necesidad de obtener una evaluación rápida de su calidad metodológica, decide utilizar un modelo de lenguaje de gran escala (Large Language Model, LLM). Para ello, introduce el resumen estructurado (abstract) del estudio y solicita un análisis crítico de su metodología y principales resultados.

La respuesta generada por la inteligencia artificial resulta coherente, estructurada y aparentemente consistente. Sin embargo, al contrastarla con el artículo original, la residente identifica discrepancias relevantes. El modelo informa un número necesario a tratar (NNT) de 12, mientras que el estudio reporta un NNT de 18. De esta manera, omite señalar que el 34 % de los participantes presentaban una tasa de filtrado glomerular (TFG) superior a 45 mL/min/1,73 m^2, característica que limita la aplicabilidad de los hallazgos al paciente evaluado, cuya TFG es de 38 mL/min/1,73 m^2.

Este escenario plantea la pregunta central que orienta el presente capítulo: ¿cómo integrar la inteligencia artificial como herramienta de apoyo en la lectura crítica de artículos científicos sin delegar el juicio clínico y metodológico al algoritmo?

2. OBJETIVOS

- Reconocer las capacidades y limitaciones actuales de los modelos de lenguaje de gran escala (LLMs) en el análisis de artículos científicos médicos.
- Aplicar un algoritmo estructurado de lectura crítica asistida por inteligencia artificial.
- Formular prompts clínicamente pertinentes y metodológicamente precisos para el análisis de literatura biomédica.
- Identificar errores típicos de la IA en la interpretación de estadística clínica, incluyendo OR, RR, HR, sensibilidad, especificidad y AUC.
- Distinguir entre el uso responsable de la IA como herramienta de apoyo y el riesgo de dependencia cognitiva.
- Integrar la IA en el flujo de lectura crítica sin comprometer el razonamiento clínico-metodológico independiente.

3. INTRODUCCIÓN

3.1 EL DESAFÍO DE LA INFORMACIÓN BIOMÉDICA CONTEMPORÁNEA

La producción científica biomédica ha experimentado un crecimiento exponencial durante las últimas dos décadas. Se estima que se publican más de 2.5 millones de artículos en revistas médicas indexadas cada año, un volumen que ningún profesional puede absorber de manera integral mediante la lectura convencional. Este fenómeno, conocido como “sobrecarga de información” (informationoverload), se ha convertido en un obstáculo estructural para la práctica de la medicina basada en la evidencia (MBE).

La sobrecarga de información no solo afecta la práctica clínica, sino que también está transformando la estructura misma del conocimiento médico. En la actualidad, se describe una “paradoja del conocimiento” en los estudiantes de medicina: a pesar de tener un acceso sin precedentes a volúmenes masivos de información, su comprensión profunda de la fisiopatología el “por qué” de las enfermedades es cada vez más superficial.

Este fenómeno se atribuye a un sistema educativo que prioriza la memorización de datos y el rendimiento en exámenes estandarizados sobre la integración conceptual y el razonamiento crítico.

Esta paradoja adquiere una relevancia capital: la misma herramienta que puede democratizar el acceso a la evidencia también puede, si se usa incorrectamente, amplificar este déficit de comprensión, convirtiendo a los clínicos en meros validadores de resúmenes sin capacidad para

https://doi.org/10.62486/978-9915-9928-3-9.Ch18

evaluar la plausibilidad fisiopatológica de los hallazgos que una IA les presenta. La lucha contra esta tendencia es el núcleo de la lectura crítica asistida que aquí se propone.

En paralelo, la lectura crítica de artículos científicos requiere competencias metodológicas específicas que no siempre se desarrollan de manera uniforme durante la formación médica. La interpretación de medidas de efecto (odds ratio OR, riesgo relativo RR, hazard ratio HR), el reconocimiento de sesgos, la evaluación de la validez interna y la valoración de la aplicabilidad clínica constituyen habilidades que demandan entrenamiento sistemático y práctica continua.

Aunque la eficiencia es innegable, la evidencia de 2025 subraya que la concordancia en la extracción automatizada de datos clínicos es de aproximadamente el 77.2%. Esto implica que 1 de cada 4 puntos de datos críticos extraídos por un LLM podría ser inexacto, lo que exige que el juicio clínico del especialista sea el filtro final e innegociable.Si bien los datos históricos apuntaban a unmargen de error variable en la extracción, los ensayos metodológicos controlados de no inferioridad (Within Reviews) publicados en 2026 demuestran que las plataformas optimizadas con LLMs de última generación alcanzan una precisión global que no es inferior a la extracción puramente humana. Aunque cometen tipos de errores similares, su verdadero valor radica en que reducen del 50% al 75% el tiempo manual de cribado y extracción de datos**,** optimizando drásticamente los costos operativos en la fase inicial de la lectura crítica.

3.2 LA IRRUPCIÓN DE LA INTELIGENCIA ARTIFICIAL EN LA PRÁCTICA CLÍNICA Y ACADÉMICA

La aparición de modelos de lenguaje de gran escala (Large LanguageModels, LLMs) como GPT, Claude y Gemini ha transformado radicalmente la manera en que los profesionales de la salud acceden y procesan información. Estos sistemas, entrenados sobre corpus masivos de texto biomédico, demostraron desde sus versiones iniciales una capacidad notable para resumir artículos, explicar conceptos estadísticos y responder preguntas clínicas en lenguaje natural.

Sin embargo, su adopción masiva en el ámbito médico ha ocurrido con mayor velocidad que la producción de evidencia sobre su seguridad, precisión y limitaciones en este contexto específico. Estudios recientes documentan que los LLMs pueden generar respuestas plausibles, pero factualmente incorrectas (fenómeno denominado “alucinación”), omitir información crítica, o sesgar sus respuestas en función de los patrones estadísticos aprendidos durante el entrenamiento, sin que el usuario pueda detectar fácilmente estos errores.

3.3MODELO DE LECTURA CRÍTICA HÍBRIDA ASISTIDA POR INTELIGENCIA ARTIFICIAL

En el contexto actual, caracterizado por la disponibilidad de modelos de lenguaje de gran escala, se propone un modelo conceptual de lectura crítica híbrida asistida por inteligencia artificial, en el cual el análisis de la evidencia científica emerge de la interacción dinámica entre el juicio clínico humano y herramientas de apoyo automatizadas.

En este modelo, la inteligencia artificial no constituye un agente epistémico autónomo, sino un sistema de apoyo que opera sobre tareas específicas como la extracción estructurada de información, la explicación de conceptos estadísticos o la identificación preliminar de sesgos, mientras que las decisiones de interpretación, aplicabilidad y toma de decisiones clínicas permanecen bajo control humano.

Este enfoque reconoce simultáneamente el potencial de la IA para mejorar la eficiencia cognitiva y el riesgo inherente de dependencia o sobreconfianza, estableciendo como eje central la necesidad de verificación sistemática y juicio crítico independiente.

4. FUNDAMENTOS DEL USO DE LA IA EN LA LECTURA CRÍTICA

4.1 ¿QUÉ ES UN MODELO DE LENGUAJE DE GRAN ESCALA O LLM (LANGUAGE LARGE MODEL)?

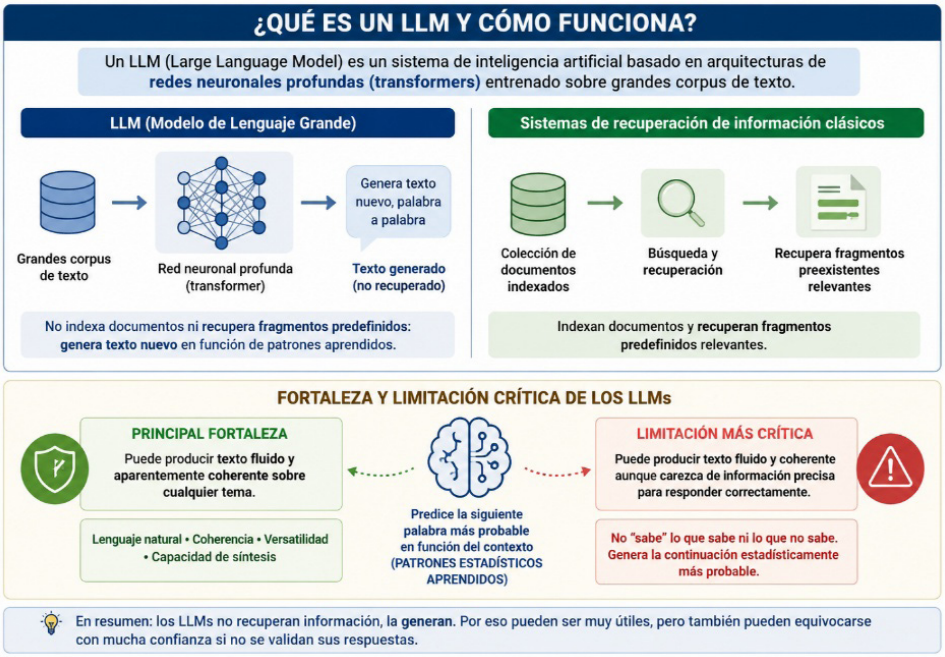
Un LLM es un sistema de inteligencia artificial basado en arquitecturas de redes neuronales profundas (transformers) entrenado sobre grandes corpus de texto.

A diferencia de los sistemas de recuperación de información clásicos, los LLMs no indexan documentos ni recuperan fragmentos predefinidos: generan texto nuevo, palabra a palabra, en función de patrones estadísticos aprendidos durante el entrenamiento. Esta característica es

https://doi.org/10.62486/978-9915-9928-3-9.Ch18

tanto su principal fortaleza como su limitación más crítica: el sistema puede producir texto fluido y aparentemente coherente sobre cualquier tema, incluso cuando carece de la información precisa para responder correctamente. No entiende el contexto de la información queposee, al contrario, es una predicción matemática y estadística de la próxima palabra con mayor probabilidad a la secuencia de texto establecida.

Figura 1: Como funciona un Language Large Model.



Fuente. Elaboración propia

4.2 FUNDAMENTOS DE LA LECTURA CRÍTICA

La lectura crítica de artículos científicos es un proceso sistemático que evalúa la validez, los resultados y la aplicabilidad de un estudio. Las dimensiones fundamentales incluyen:

- **Validez interna:** ¿el diseño y la ejecución del estudio permiten atribuir los resultados a la intervención estudiada y no a factores confundentes?
- **Validez externa (aplicabilidad):** ¿los hallazgos son generalizables a la población de pacientes del lector?
- **Importancia clínica:** ¿la magnitud del efecto es relevante para la práctica, independientemente de la significancia estadística?
- **Precisión estadística:** ¿los intervalos de confianza son suficientemente estrechos para guiar decisiones?

El uso de la IA en este proceso debe estar necesariamente subordinado a este marco conceptual, y no sustituirlo.

4.3 PRINCIPIOS PARA LA INTEGRACIÓN DE IA EN LA LECTURA CRÍTICA

Se plantean entonces cuatro principios que en adelante guiarán el uso responsable de la IA en este contexto:

- Subordinación:La IA está subordinada al juicio clínico; asiste, no decide. El juicio clínico y metodológico final es exclusivamente humano.
- Verificación: toda afirmación de la IA sobre datos, cifras o conclusiones debe cotejarse con el texto original del artículo.
- **Transparencia:** el usuario debe conocer las capacidades y limitaciones del sistema que utiliza, incluyendo su fecha de corte de entrenamiento.
- **Proporcionalidad:** la profundidad del uso de la IA debe ser proporcional a la complejidad del artículo y la trascendencia clínica de la decisión.

https://doi.org/10.62486/978-9915-9928-3-9.Ch18

Figura 2. Principios para integración de IA en la lectura crítica

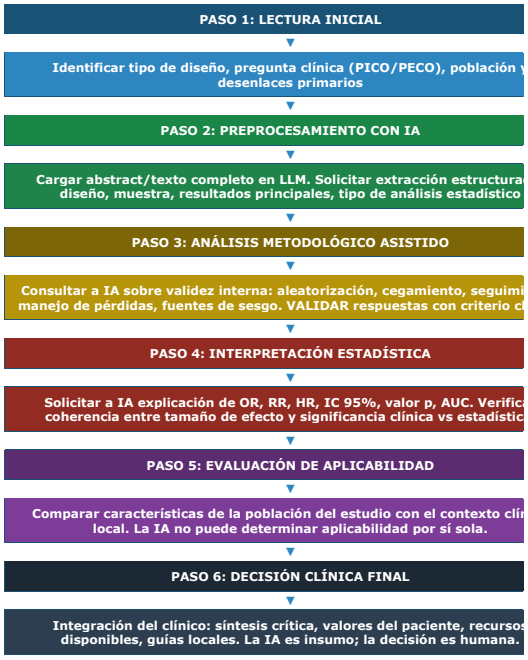
Fuente: elaboración propia con apoyo de Inteligencia Artificial

En el entorno editorial contemporáneo, ya no es suficiente declarar el uso genérico de algoritmos. Las publicaciones científicas exigen la adopción de listas de verificación estandarizadas como PRISMA-trAIce (extensión de PRISMA para el rastro de IA) y el estricto cumplimiento del Marco RAISE, obligando al investigador a documentar de forma transparente cuándo una herramienta ha propuesto o emitido juicios metodológicos. De esta manera, bajo las directrices de transparencia institucional (como las normativas HTI-1), los entornos académicos están transitando de LLMs comerciales en la nube hacia modelos de código abierto locales (On-Premise)**,** garantizando que el procesamiento de la literatura no exponga indirectamente patrones asistenciales ni vulnere la gobernanza de datos del centro de salud.

4.4 ALGORITMO DE LECTURA CRÍTICA ASISTIDA POR IA

La Figura 3 presenta un algoritmo estructurado en seis pasos para integrar el uso de la IA en el proceso de lectura crítica, desde la lectura inicial hasta la toma de decisión clínica.

Figura 3. Algoritmo de lectura crítica asistida por inteligencia artificial.



Fuente: Elaboración propia

https://doi.org/10.62486/978-9915-9928-3-9.Ch18

5. APORTES DE LA IA EN EL ANÁLISIS DE ARTÍCULOS CIENTÍFICOS

Dado el inmenso potencial del uso de las herramientas de IA se plantean a continuación como puede contribuir la IA en el análisis de artículos científicos.

5.1 PROCESAMIENTO Y EXTRACCIÓN DE INFORMACIÓN ESTRUCTURADA

Una de las contribuciones más documentadas de los LLMs en la lectura de literatura científica es su capacidad para extraer información estructurada de textos no estructurados. Al proporcionarle el *abstract* o el texto completo de un artículo, el modelo puede identificar con razonable precisión el diseño del estudio, la población incluida, la intervención evaluada, los comparadores, los desenlaces principales y secundarios, y los resultados numéricos clave. Esta tarea, que un lector humano ejecuta en varios minutos de atención sostenida, puede completarse en segundos con el apoyo de la IA.

Sin embargo, la precisión de esta extracción no es absoluta. Los estudios de evaluación de los LLMs en tareas de extracción de datos de ensayos clínicos y revisiones sistemáticas reportan tasas de error que oscilan entre el 15 % y el 30 %, especialmente cuando se trata de valores numéricos precisos, datos de subgrupos y resultados secundarios. Un estudio dentro de revisiones (*Study Within Reviews*, SWAR) que comparó un flujo asistido por IA frente a la extracción exclusivamente humana en 9341 elementos de datos procedentes de 63 estudios encontró una concordancia global entre ambos métodos del 77,2 % (IC 95 %: 76,3 %-78,0 %) (1). Aunque la precisión del método asistido por IA fue alta en términos globales (91,0 %; IC 95 %: 90,4 %-91,6 %), los errores residuales no se distribuyeron de manera uniforme: se concentraron en las variables más complejas y en los resultados secundarios, mientras que los datos estructurales simples (como el número de participantes o el país de realización) se extrajeron con mayor fiabilidad.

Un hallazgo particularmente preocupante para la lectura crítica es la marcada heterogeneidad en el rendimiento de los LLM en función del dominio de datos. Una revisión sistemática rápida que analizó 31 estudios de extracción de datos clínicos mediante LLMs encontró que la precisión comunicada oscilaba entre el 17 % y el 100 % para las características del diseño del estudio, entre el 9,1 % y el 100 % para las características de los pacientes, y entre el 30 % y el 100 % para los desenlaces del estudio.

Esta amplitud refleja no solo diferencias reales en la dificultad de cada dominio, sino también una considerable heterogeneidad metodológica entre los estudios primarios, que emplearon distintos modelos, arquitecturas de *prompt*, formatos de entrada y definiciones de error.

La extracción de información estadística compleja como los resultados de los análisis de sensibilidad, las medidas de heterogeneidad (I^2, τ^2) en los metaanálisis o las tablas de riesgo de sesgo muestra tasas de error considerablemente mayores que la extracción de datos descriptivos simples.

En el estudio de Martin et al., que evaluó repetidamente ChatGPT-4.0 en la extracción de datos de un mismo conjunto de publicaciones a lo largo de varias semanas, los errores globales alcanzaron el 32,4 % en los informes completos de ensayos clínicos, y la recuperación de elementos complejos como la cita bibliográfica completa resultó correcta en solo el 0 % al 33,3 % de las ejecuciones (en sus últimas versiones puede haber menos riesgo de alucinación). De particular relevancia clínica fue el hallazgo de que ningún elemento de datos pudo ser extraído de forma consistentemente correcta en todos los tipos de publicación y en todas las repeticiones del experimento. Esta variabilidad temporal la misma pregunta, el mismo artículo y el mismo *prompt*, pero resultados diferentes en momentos distintos constituye un desafío específico de los modelos probabilísticos que el lector crítico debe conocer: a diferencia de un sistema determinista, el LLM no garantiza reproducibilidad en la extracción.

Para el estudiante y profesional de la salud que utiliza la IA en la lectura crítica, estos datos tienen una doble implicación práctica: en primer lugar, la extracción automatizada puede ahorrar tiempo y proporcionar una primera aproximación útil a la estructura del artículo; en segundo lugar, y de manera irrenunciable, todo valor numérico, toda medida de efecto y todo resultado de un análisis de sensibilidad o de heterogeneidad generado por el LLM debe ser verificado directamente en el texto fuente original. La IA acelera la lectura, pero no garantiza la fidelidad de lo leído.

https://doi.org/10.62486/978-9915-9928-3-9.Ch18

Tabla 1. Matriz de Fiabilidad en la Extracción de Datos por IA

Categoría de Datos	Precisión Estimada de la IA	Nivel de Supervisión Humana Requerido	Riesgo de Alucinación Clínica
Identificadores Básicos (País, autor, año de publicación)	Muy Alta (>95%)	Baja: Verificación rápida de metadatos.	Mínimo: Datos estructurados simples.
Población y Muestra (Criterios de inclusión, tamaño de muestra N)	Media-Alta (~85%)	Moderada: Validación cruzada con la tabla de características basales.	Bajo-Moderado: Posible confusión entre grupos (intervención vs. control).
Desenlaces Primarios (Resultados principales del estudio)	Moderada (77.2% concordancia)	Alta: El lector debe confirmar la magnitud del efecto y la significancia estadística.	Moderado: Tendencia a omitir matices en los intervalos de confianza.
Desenlaces Secundarios y Subgrupos (Efectos adversos, análisis post-hoc)	Baja (<60%)	Crítica: Análisis exhaustivo por parte del lector.	Alto: Frecuentes errores en la interpretación de variables no primarias.
Evaluación de Sesgos (Cegamiento, ocultación de la secuencia)	Muy Baja	Obligatoria: La IA carece de contexto para evaluar la integridad metodológica real.	Muy Alto: El sistema tiende a ser complaciente con la metodología declarada.

5.2 APOYO EN LA INTERPRETACIÓN ESTADÍSTICA

La dificultad en la interpretación de medidas estadísticas es uno de los obstáculos más frecuentes en la lectura crítica para profesionales en formación. Los LLMs demuestran una capacidad razonable para explicar conceptos como:

- Odds Ratio (OR): medida de asociación que compara las probabilidades (odds) de un evento en dos grupos. Un OR > 1 indica mayor probabilidad en el grupo expuesto. Importante: en enfermedades frecuentes, el OR sobreestima el riesgo relativo.
- **Riesgo Relativo (RR):** razón entre la incidencia del evento en el grupo expuesto y en el no expuesto. Directamente interpretable en estudios de cohorte y ensayos clínicos.
- **Hazard Ratio (HR):** análogo al RR en análisis de supervivencia; incorpora el tiempo hasta el evento. Asume proporcionalidad de hazards, supuesto que debe verificarse.
- **Sensibilidad y Especificidad:** en pruebas diagnósticas, la sensibilidad refleja la proporción de verdaderos positivos; la especificidad, la de verdaderos negativos. Ambas son propiedades de la prueba, no del paciente.
- **AUC-ROC:** área bajo la curva ROC; mide la capacidad discriminativa global de una prueba diagnóstica. Un valor de 0.5 es equivalente al azar; 1.0 es discriminación perfecta.

El uso de la IA para la explicación de estos conceptos resulta particularmente valioso cuando va acompañado de ejemplos numéricos derivados del propio artículo en análisis. No obstante, el lector debe verificar que los valores citados por la IA correspondan efectivamente a los reportados en el texto original.

5.3 SÍNTESIS COMPARATIVA Y CONTEXTUALIZACIÓN

Una vez que el lector ha extraído los resultados principales de un artículo, surge la pregunta clave: ¿cómo se sitúan esos hallazgos en el panorama general de la evidencia disponible?

Los LLMs con capacidades de búsqueda en tiempo real o con acceso a bases de datos científicas ofrecen aquí una asistencia potencialmente valiosa. Pueden localizar estudios previos sobre la misma intervención, comparar la magnitud de los efectos, señalar concordancias o discrepancias, e incluso sugerir si las conclusiones del nuevo trabajo modifican o confirman el conocimiento asentado. Para un residente con poco tiempo o para un estudiante que se inicia en un área clínica, esta funcionalidad tiene un alto valor pedagógico: convierte la lectura de un artículo aislado en un ejercicio de pensamiento contextual (4,6).

No obstante, esta capacidad presenta limitaciones severas que deben modular el entusiasmo inicial. Los modelos de lenguaje, incluso aquellos con módulos de recuperación de información, son propensos a dos tipos de error en las tareas de síntesis comparativa. El primero y más documentado es la fabricación de referencias: el sistema puede generar una comparación aparentemente sensata y citar tras ella un autor, una revista y un año que, en conjunto o en parte, no existen. Esta alucinación es especialmente peligrosa porque la referencia ficticia suele estar correctamente formateada y guardar coherencia temática con el contenido, lo que la hace difícil de detectar para un ojo no entrenado.

Por ello, la recomendación es contundente: la comparación contextual que ofrezca el LLM

https://doi.org/10.62486/978-9915-9928-3-9.Ch18

debe considerarse una hipótesis de trabajo, no un hecho verificado. Cada cita debe comprobarse en la base de datos de referencia institucional (PubMed, Cochrane Library, EMBASE, etc.), y cada afirmación sobre "lo que dice la literatura" debe cotejarse con, al menos, dos fuentes primarias. Cuando esta cautela se respeta, la síntesis asistida ahorra horas de búsqueda; cuando se omite, constituye una fuente de error de alto riesgo para la seguridad del paciente.

Evaluaciones masivas (*benchmarks head-to-head*) que han contrastado el análisis algorítmicofrente al humano en miles de estudios indexados revelan que los LLMs superan la velocidad de síntesis humana en un 33% de las tareas, siempre que el operador aplique técnicas de *prompting* avanzado. Sin embargo, estas evaluaciones de 'caja negra' suelen pasar por alto un peligro crítico: el sesgo de complacencia (*sycophancy*). El modelo tiende probabilísticamente a validar las premisas erróneas del clínico en lugar de confrontarlo metodológicamente, lo que refuerza la necesidad de que la IA actúe como un insumo estructurado para las Guías de Práctica Clínica y nunca como el evaluador de la ponderación final de la evidencia

5.4 LECTURA TRADICIONAL VERSUS LECTURA ASISTIDA POR IA

Tabla 2. Comparación entre la lectura crítica tradicional y la lectura asistida por inteligencia artificial

Dimensión	Lectura Tradicional	Lectura Asistida por IA
Tiempo promedio por artículo	45-120 minutos	15-40 minutos (sin comprometer análisis crítico)
Extracción de datos	Manual, sujeta a fatiga cognitiva	Automática para datos estructurales; requiere verificación
Análisis estadístico	Dependiente del nivel de formación del lector	IA explica medidas; el clínico interpreta pertinencia
Detección de sesgos	Basada en formación metodológica individual	IA sugiere sesgos comunes; el lector los confirma
Síntesis de evidencia	Narrativa, según criterio del lector	IA puede comparar con literatura previa (con riesgos de alucinación)
Aplicabilidad clínica	Depende del juicio clínico	IA no puede determinarla; exclusivamente humano
Acceso equitativo	Limitado por formación metodológica	Democratiza el acceso básico al análisis (con supervisión)
Riesgo de dependencia cognitiva	Bajo	ALTO: riesgo de reducción del pensamiento crítico si se usa sin supervisión

Fuente: elaboración propia con apoyo de Inteligencia Artificial

Tabla 3. Capacidades y limitaciones de los LLMs en el análisis de artículos científicos

Capacidades de la IA (LLMs)	Limitaciones y Riesgos
Extracción rápida de información estructurada	ALUCINACIONES: puede inventar datos, autores o resultados
Explicación de conceptos estadísticos (OR, RR, HR, AUC)	No accede al texto completo a menos que se lo proporcione el usuario
Identificación de posibles sesgos metodológicos	Conocimiento limitado a fecha de corte de entrenamiento
Síntesis comparativa de múltiples estudios	No puede evaluar calidad metodológica sin instrucción explícita
Generación de preguntas de lectura crítica	Tendencia a respuestas positivas ('sí funciona') por sesgos en datos de entrenamiento
Traducción y simplificación de lenguaje técnico	No reemplaza instrumentos validados (CONSORT, STROBE, GRADE)
Disponibilidad 24/7 sin costo en versiones básicas	Ausencia de responsabilidad profesional: no puede ser auditada clínicamente
Soporte para lectores en formación	Riesgo de dependencia cognitiva en estudiantes y residentes

Fuente: elaboración propia con apoyo de Inteligencia Artificial

https://doi.org/10.62486/978-9915-9928-3-9.Ch18

6. USO PRÁCTICO DE MODELOS DE LENGUAJE: DISEÑO Y APLICACIÓN DE PROMPTS

6.1 EL PROMPT COMO HERRAMIENTA METODOLÓGICA

El *prompt* o la instrucción textual que el usuario proporciona al LLM no es una simple pregunta: es, en sí mismo, un instrumento de análisis que define la profundidad, el enfoque y la utilidad clínica de la respuesta. En la lectura crítica, un *prompt* impreciso ("analiza este artículo") produce un resumen vago que puede omitir los puntos metodológicos cruciales. En cambio, un *prompt* estructurado, contextualizado y orientado a una tarea específica convierte al modelo en un asistente que realmente apoya el juicio clínico.

Un *prompt* efectivo para lectura crítica contiene cuatro componentes esenciales:

- **Contexto del análisis:** se indica el tipo de artículo (ECA, cohorte, revisión sistemática), la pregunta clínica de fondo (formato PICO) y, si procede, la especialidad.
- **Texto fuente:** se proporciona el *abstract* o fragmentos del artículo (nunca se confía en el conocimiento pretérito del modelo).
- **Tarea específica:** se define con precisión qué se solicita: extraer, evaluar, explicar, comparar, detectar sesgos.
- **Formato de salida deseado:** lista estructurada, tabla, párrafo explicativo con cada dimensión por separado, o respuesta en formato de lista de verificación (checklist).

Ejemplo de *prompt* pobre: "¿Qué dice este estudio?".

Ejemplo de *prompt* enriquecido: "Eres un revisor metodológico. Lee el siguiente *abstract* de un ensayo clínico aleatorizado sobre [fármaco] en [condición]. Extrae en una tabla: diseño, tamaño muestral, criterios de inclusión/exclusión, desenlace primario, resultado principal con IC 95 %, y principales limitaciones mencionadas por los autores."

Al estructurar el *prompt* de esta manera, el clínico no solo obtiene una respuesta más útil, sino que además ejercita su propio razonamiento metodológico: la calidad del *prompt* refleja la calidad del pensamiento crítico de quien lo formula.

De manera importante, la formulación de *prompts* no debe entenderse únicamente como una habilidad técnica, sino como una extensión del razonamiento clínico y metodológico. Un *prompt* bien estructurado refleja una comprensión previa del problema, mientras que un *prompt* ambiguo suele traducir una falta de claridad conceptual.

En este sentido, el uso de modelos de lenguaje en la lectura crítica no solo proporciona respuestas, sino que también actúa como un espejo del pensamiento del usuario, permitiendo identificar fortalezas y limitaciones en su proceso analítico.

6.2 FLUJO DE INTERACCIÓN CON EL LLM

La Figura 5 ilustra el flujo recomendado de interacción entre el lector y un LLM para el análisis de un artículo científico, con énfasis en los puntos de validación humana obligatorios.

Figura 5. Flujo de interacción con un modelo de lenguaje de gran escala para el análisis de artículos científicos.

ENTRADA (Input)	PROCESAMIENTO (LLM)	SALIDA + VALIDACIÓN
1. Texto del artículo (abstract o completo)	Análisis semántico y contextual. Identificación de conceptos metodológicos.	Resumen estructurado del diseño y hallazgos. VERIFICAR con el artículo original.
2. Prompt específico (ej. 'Evalúa la validez interna')	Recuperación de patrones sobre sesgos, diseños, estadística de su entrenamiento.	Lista de fortalezas/limitaciones. COTEJAR con herramientas validadas (CONSORT, STROBE, etc.).
3. Pregunta clínica explícita (PICO)	Mapeo de la pregunta al contenido del estudio.	Valoración de pertinencia. NUNCA asumir que la IA conoce el contexto local del clínico.
4. Solicitud de definición estadística (OR, RR, HR, AUC)	Recuperación de definiciones y ejemplos numéricos.	Explicación didáctica. VALIDAR con fuentes estadísticas primarias ante resultados complejos.

Fuente: Elaboración propia.

https://doi.org/10.62486/978-9915-9928-3-9.Ch18

6.3 CATÁLOGO DE PROMPTS CLÍNICAMENTE VALIDADOS

La Tabla 4 presenta ejemplos de prompts estructurados para diferentes objetivos en la lectura crítica, junto con las precauciones específicas asociadas a cada uso.

Tabla 4. Ejemplos de prompts clínicos para la lectura crítica asistida por IA

Objetivo	Ejemplo de Prompt	Precauciones
Identificar diseño del estudio	"Lee este abstract y dime: ¿qué tipo de diseño es, cuál es la hipótesis principal, cuál es la población y cuáles son los desenlaces primarios y secundarios?"	Verificar con el texto completo.
Evaluar validez interna	"Basándote en este texto, ¿qué fuentes de sesgo podrían afectar los resultados? Considera: selección, información, confusión y sesgo de publicación."	Dependiendo del modelo, a veces la IA no revisa el PDF directamente; proporcionar el texto.
Interpretar estadística	"El RR fue 0.72 (IC 95%: 0.58-0.89, p=0.003). Explica qué significa clínicamente este resultado y si el IC excluye el valor nulo."	Validar con formación estadística propia.
Aplicar GRADE o PICO	"Extrae la estructura PICO de este estudio y evalúa si el nivel de evidencia permitiría una recomendación fuerte o débil según GRADE."	GRADE requiere síntesis de múltiples estudios; la IA solo aproxima.
Comparar con guías clínicas	"Este ensayo reporta beneficio del fármaco X. ¿En qué difieren sus resultados de las guías ESC/AHA vigentes? Señala posibles explicaciones."	La IA puede tener información desactualizada. Consultar guías actuales.
Detectar conflictos de interés	"¿Qué información sobre financiamiento y conflictos de interés menciona este artículo? ¿Podría esto afectar la validez de las conclusiones?"	La IA solo analiza lo declarado; no detecta conflictos ocultos.

Fuente: Elaboración propia

6.4 LIMITACIONES TÉCNICAS DEL USO DE PROMPTS

El lector debe considerar que los LLMs presentan una "ventana de contexto" limitada: la cantidad de texto que pueden procesar simultáneamente tiene un techo técnico que varía entre modelos. Artículos extensos, especialmente aquellos con tablas de datos complejas, pueden exceder esta ventana, resultando en análisis incompletos o truncados. En estos casos, se recomienda dividir el artículo en segmentos temáticos y proporcionar cada uno en prompts separados, con un prompt final de síntesis.

6.5 CRITERIOS PARA LA SELECCIÓN Y EVALUACIÓN DEL MODELO DE LLM EN LECTURA CRÍTICA

El capítulo ha abordado los LLMs en abstracto, pero la práctica clínica obliga a elegir entre múltiples modelos GPT, Claude, Gemini, Llama-3, Med-PaLM, entre otros cuyas capacidades no son intercambiables cuando se trata de análisis metodológico. Cuatro criterios orientan una selección informada y una evaluación local antes de integrar el modelo en el flujo de lectura crítica.

a). Rendimiento en tareas metodológicas específicas (métricas de exactitud en extracción e interpretación).

Existen bancos de pruebas públicos (*benchmarks*: conjuntos de preguntas o artículos estandarizados que permiten comparar el desempeño de distintos modelos) diseñados para medir la capacidad de un LLM en tareas médicas concretas: extracción de datos de ensayos clínicos, interpretación de medidas de efecto o detección de sesgos metodológicos. Algunos ejemplos son el conjunto de problemas médicos desafiantes (*MedQA*, preguntas tipo USMLE) o las evaluaciones específicas publicadas en revistas revisadas por pares.

Antes de adoptar un modelo, el profesional (o la institución) debe verificar si existe evidencia pública sobre su tasa de aciertos en estas tareas. Las cifras de precisión pueden oscilar entre el 75 % y el 95 % según el modelo, la versión y la complejidad del *prompt*, lo que subraya la necesidad de no asumir competencia sin datos (1-3).

b). Capacidad de manejo de la ventana de contexto y modalidad de entrada (cantidad de

https://doi.org/10.62486/978-9915-9928-3-9.Ch18

texto que el modelo puede leer de una vez y tipos de archivo que acepta).

La ventana de contexto medida en *tokens* (fragmentos de palabras) determina cuánto texto puede procesar el modelo en una sola consulta. Un metaanálisis extenso con tablas de heterogeneidad y diagramas de bosque puede superar con facilidad los 8.000 o 32.000 tokens de los modelos básicos, lo que obliga a trocear el artículo y, por tanto, a perder la visión de conjunto. Debe comprobarse, además, si el modelo acepta directamente archivos PDF o si requiere que el usuario copie y pegue el texto, en cuyo caso las tablas pueden desestructurarse y volverse ilegibles para el sistema.

c). Cumplimiento de requisitos éticos y de privacidad de datos (normativas de protección de información personal y confidencialidad).

Aunque un artículo científico publicado no contiene datos identificables de pacientes, el acto de consultar un LLM basado en la nube (servidores externos a la institución, accesibles por internet) sobre un tema que el clínico asocia a su práctica habitual puede exponer, indirectamente, patrones asistenciales o información sobre la casuística del centro. En jurisdicciones regidas por normativas como el Reglamento General de Protección de Datos (RGPD) europeo o la Ley de Portabilidad y Responsabilidad de Seguros de Salud (HIPAA) estadounidense, esto puede ser relevante. En esos casos, se prefieren los modelos de código abierto (*open source*: modelos cuyo código y pesos están disponibles públicamente y pueden instalarse en servidores propios), que permiten un despliegue local con control institucional completo y sin que los datos salgan del centro. La decisión debe ser institucional, documentada y revisada por el comité de ética o el responsable de protección de datos.

d). Disponibilidad de búsqueda en tiempo real y actualización del conocimiento (consulta a bases de datos externas).

Algunos modelos incorporan módulos de búsqueda que les permiten consultar PubMed u otras bases de datos en el momento de la consulta. Esta capacidad mitiga parcialmente el sesgo de desactualización (el modelo "sabe" solo lo que estaba disponible en la fecha de corte de su entrenamiento, que puede ser uno o dos años anterior). No obstante, la búsqueda en tiempo real introduce un nuevo riesgo: el modelo puede recuperar una referencia superficialmente pertinente pero inapropiada en diseño o población, y presentarla como evidencia de contraste. La coherencia entre la respuesta generada y la referencia primaria que el propio modelo cita debe verificarse siempre.

La Tabla 5 resume los criterios y las preguntas prácticas que el lector debe formular antes de adoptar un LLM en su práctica de lectura crítica.

Tabla 5. Criterios para la selección de un LLM en el contexto de la lectura crítica de literatura médica

Criterio	Pregunta clave a considerar	Indicadores a evaluar
Rendimiento en tareas metodológicas	¿Existe evidencia pública de su precisión en extracción de OR, RR, NNT, evaluación de sesgos?	Resultados en Medical Challenge Problems, USMLE, estudios de comparación head-to-head en lectura crítica
Ventana de contexto y entrada de datos	¿Admite el texto completo del artículo más las tablas sin truncar? ¿Procesa PDF o tablas?	Tamaño máximo de tokens; soporte de carga de archivos PDF; límites prácticos de la interfaz
Privacidad y entorno de ejecución	¿Dónde se procesan los datos? ¿Existe riesgo de exponer consultas clínicas?	Modelo alojado en nube pública vs. despliegue local; certificaciones (HIPAA, RGPD); política de uso de datos
Actualización y recuperación externa	¿Incorpora búsqueda en tiempo real en bases como PubMed? ¿La verificación es auditada?	Plugins de búsqueda, frecuencia de actualización del conocimiento, coherencia entre cita y texto recuperado

Fuente: elaboración propia

La evidencia actual subraya que ningún modelo puede asumirse competente sin una evaluación específica para la tarea de lectura crítica. Por ejemplo, un análisis de los beneficios, límites y riesgos de GPT en medicina demostró que el modelo comete errores de razonamiento en la interpretación de estudios clínicos, a pesar de obtener puntuaciones altas en exámenes de conocimiento, lo que refuerza la necesidad de una selección basada en desempeño real en el dominio metodológico (29).

7. INTERPRETACIÓN CRÍTICA ASISTIDA POR IA
7.1 EVALUACIÓN DE LA VALIDEZ INTERNA CON APOYO DE IA

La validez interna de un estudio se evalúa examinando si el diseño y la ejecución minimizaron

https://doi.org/10.62486/978-9915-9928-3-9.Ch18

adecuadamente los sesgos sistemáticos.

Los LLMs pueden asistir en este proceso identificando, a partir del texto del artículo, elementos que sugieren riesgo de sesgo en las siguientes dimensiones:

- Sesgo de selección: ¿el proceso de aleatorización fue adecuado? ¿se ocultó la secuencia de asignación?
- Sesgo de información: ¿se realizó cegamiento? ¿los desenlaces fueron evaluados de manera ciega?
- Sesgo de seguimiento: ¿las pérdidas fueron reportadas y manejadas adecuadamente? ¿se realizó análisis por intención de tratar?
- Sesgo de confusión: ¿se controlaron las variables de confusión más relevantes mediante diseño o análisis estadístico?

Sin embargo, el lector debe comprender que la IA no "lee" el artículo con comprensión metodológica genuina: identifica patrones lingüísticos asociados a estos conceptos. Puede describir correctamente cómo debería ser un proceso de aleatorización adecuado, pero no puede auditar si el proceso descrito en el artículo fue efectivamente seguido.

7.2 EVALUACIÓN DE LA SIGNIFICANCIA CLÍNICA VS. ESTADÍSTICA

Uno de los aportes más valiosos de la IA en la interpretación de artículos es la ayuda para distinguir entre significancia estadística y relevancia clínica. El lector puede solicitar al LLM que explique la diferencia entre un valor p significativo con un tamaño de efecto pequeño y un valor p no significativo con un tamaño de efecto clínicamente relevante, concepto frecuentemente incomprendido en la práctica clínica.Ejemplo aplicado: un ensayo con 15.000 participantes reporta una reducción en la presión arterial sistólica de 2.3 mmHg con el fármaco estudiado ($p < 0.001$).

La IA puede explicar que, aunque el resultado es estadísticamente significativo, la magnitud del efecto es de dudosa relevancia clínica en la mayoría de los contextos, y que el intervalo de confianza del 95% (0.8-3.8 mmHg) confirma la precisión estadística pero no la importancia clínica. Esta distinción es fundamental para evitar la sobreinterpretación de resultados en estudios con grandes muestras.

7.3 EVALUACIÓN DE LA APLICABILIDAD

La aplicabilidad de un estudio la transferibilidad de sus resultados a un paciente específico es la dimensión en la que la IA tiene menor utilidad y mayor riesgo si se utiliza sin supervisión.

La aplicabilidad requiere comparar las características de la población del estudio con el paciente individual del lector, considerar el contexto asistencial, los recursos disponibles, los valores del paciente y las preferencias individuales.

Ninguno de estos elementos puede ser adecuadamente evaluado por un LLM, que carece de acceso a la historia clínica del paciente y al contexto local del clínico.

Punto crítico: La evaluación de la aplicabilidad clínica es exclusivamente humana. La IA puede describir las características de la población del estudio, pero no puede determinar si el hallazgo es transferible al paciente que tiene el lector frente a sí. Este juicio integra conocimiento clínico, contexto local y valores del paciente dimensiones fuera del alcance algorítmico.

8. RIESGOS, SESGOS Y LIMITACIONES DEL USO DE IA EN LECTURA CRÍTICA

8.1 ALUCINACIONES EN MODELOS DE LENGUAJE: EL RIESGO MÁS CRÍTICO

El término "alucinación" en el contexto de los LLMs designa la generación de contenido factualmente incorrecto que es presentado con el mismo nivel de confianza lingüística que el contenido correcto. A diferencia del error humano, que generalmente produce señales de incertidumbre (vacilación, expresiones de duda, reconocimiento de ignorancia), el LLM alucinado produce afirmaciones falsas con una fluidez y seguridad formal que dificulta su detección.

En el contexto específico de la lectura crítica médica, las alucinaciones más peligrosas incluyen:

- Invención o alteración de datos numéricos: el LLM puede reportar un OR, RR, NNT, valor p o IC 95% que difiere del publicado en el artículo, sin que el texto lo indique.
- Fabricación de referencias: el sistema puede citar estudios inexistentes, con autores reales o

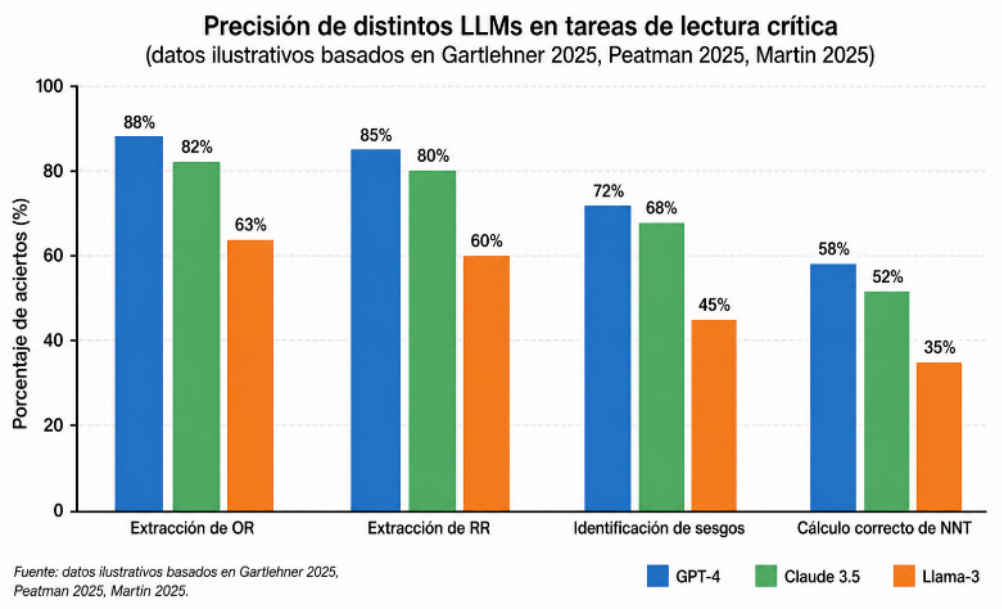
https://doi.org/10.62486/978-9915-9928-3-9.Ch18

inventados, en revistas reales o inexistentes, con DOIs no válidos.

- Atribución incorrecta de hallazgos: el sistema puede asignar resultados de un estudio a otro, o mezclar datos de múltiples publicaciones.
- Conclusiones inversas: en casos documentados, los LLMs han descrito intervenciones como efectivas cuando el estudio original reportó ausencia de efecto o daño.

La tasa de alucinaciones varía según el modelo, la versión y el tipo de tarea. Estudios publicados entre 2023 y 2025 reportan que los LLMs de última generación cometen errores factuales en tareas de lectura crítica médica en un porcentaje que oscila entre el 10% y el 40%, dependiendo de la complejidad del artículo y la especificidad de la consulta. Este rango, aunque amplio, es suficiente para justificar la verificación sistemática de toda afirmación numérica producida por el sistema.

Figura 6. precisión de distintos LLMs en tareas de lectura crítica



8.2 SESGOS INHERENTES AL ENTRENAMIENTO

Los LLMs heredan los sesgos presentes en los datos utilizados para su entrenamiento. En biomedicina, esto tiene consecuencias concretas:

- **Sesgo de publicación amplificado:** los LLMs han aprendido predominantemente de artículos publicados, que tienen mayor proporción de resultados positivos. Esto puede generar una tendencia del sistema a presentar las intervenciones de manera más favorable de lo que la evidencia total (incluyendo literatura gris) justifica.
- **Sesgo geográfico y de idioma:** los corpus de entrenamiento están dominados por literatura en inglés de países de ingresos altos. Esto limita la representatividad de los outputs para contextos clínicos latinoamericanos, africanos o asiáticos.
- **Sesgo de desactualización:** los LLMs tienen una fecha de corte de entrenamiento (knowledgecutoff). Cualquier evidencia publicada después de esa fecha es inaccesible para el modelo en su versión base, aunque algunos modelos disponen de capacidades de búsqueda en tiempo real que mitigan parcialmente esta limitación.
- **Sesgo de complacencia:** los LLMs tienden a validar las premisas implícitas en las preguntas del usuario. Si el prompt asume que un fármaco es efectivo, el sistema tenderá a construir respuestas que confirmen esa premisa, aunque la evidencia sea ambigua.

8.3 ERRORES DE INTERPRETACIÓN METODOLÓGICA

Más allá de las alucinaciones numéricas, los LLMs exhiben patrones sistemáticos de error metodológico que el lector crítico debe conocer:

- Confusión entre diseños de estudio: estudios de cohorte retrospectivos descritos como

https://doi.org/10.62486/978-9915-9928-3-9.Ch18

"quasi-experimentales", series de casos presentadas como "estudios observacionales prospectivos".

- Incapacidad para evaluar el riesgo de sesgo en ausencia del texto completo: el LLM puede describir el diseño correcto de un ECA sin detectar si el artículo específico lo ejecutó adecuadamente.
- Interpretación incorrecta de IC 95% que cruzan la unidad: en algunos casos documentados, los LLMs han descrito un OR de 1.4 (IC 95%: 0.9-2.1) como "estadísticamente significativo" o como "tendencia significativa".
- Aplicación indiscriminada del concepto de NNT: el NNT es sensible al riesgo basal de la población; el LLM puede citar el NNT del estudio sin ajustarlo al riesgo del paciente del lector.

8.4 FALSA SENSACIÓN DE COMPRENSIÓN

Uno de los riesgos menos documentados, pero potencialmente más significativos desde una perspectiva pedagógica es lo que puede denominarse "ilusión de competencia metodológica". El estudiante o residente que utiliza un LLM para analizar un artículo puede percibir que ha realizado una lectura crítica completa cuando, en realidad, ha recibido una síntesis generada algorítmicamente que puede ser incorrecta, incompleta o superficial.

Este fenómeno se ve potenciado por el hecho de que el LLM utiliza el vocabulario correcto de la metodología de investigación clínica: menciona el riesgo de sesgo, calcula (o simula calcular) el NNT, evalúa la heterogeneidad en metaanálisis. La forma del análisis parece metodológicamente sólida, aunque el contenido pueda ser deficiente. El peligro es que el lector no entrenado carece de los instrumentos para distinguir entre una respuesta correcta y una plausiblemente incorrecta.

8.5 DEPENDENCIA COGNITIVA Y EROSIÓN DEL RAZONAMIENTO CRÍTICO

La dependencia cognitiva de la IA constituye un riesgo de carácter formativo. Si los profesionales en formación adoptan el LLM como sustituto del análisis metodológico propio, en lugar de utilizarlo como herramienta de apoyo, el resultado puede ser una reducción progresiva de su capacidad para leer críticamente de manera independiente.

Esta preocupación, planteada inicialmente en el debate sobre la calculadora en la educación matemática, adquiere una dimensión clínica directa: el profesional que no puede evaluar críticamente la evidencia de manera autónoma es más vulnerable al error cuando la IA no está disponible o cuando esta produce una respuesta incorrecta.

La evidencia sobre este riesgo específico en educación médica es aún limitada, pero los principios de la ciencia cognitiva sugieren que el uso sin reflexión de herramientas de automatización puede inhibir la consolidación de esquemas cognitivos de análisis crítico, especialmente durante las etapas formativas.

8.6 RIESGOS COGNITIVOS ASOCIADOS AL USO DE IA EN LECTURA CRÍTICA

Más allá de los errores técnicos, el principal riesgo del uso de modelos de lenguaje en la lectura crítica es de naturaleza cognitiva. La fluidez lingüística y coherencia superficial de las respuestas puede generar una **ilusión de comprensión**, en la cual el usuario percibe haber entendido un artículo sin haber realizado un análisis metodológico profundo.

Este fenómeno se relaciona con el **sesgo de automatización**, descrito en la literatura como la tendencia a sobrevalorar la información generada por sistemas automatizados, incluso en presencia de evidencia contradictoria. En el contexto de la medicina basada en la evidencia, esto puede traducirse en decisiones clínicas sustentadas en interpretaciones no verificadas.

Adicionalmente, el uso reiterado de IA sin una participación activa del usuario puede conducir a atrofia del razonamiento crítico, particularmente en estudiantes y médicos en formación, quienes podrían delegar progresivamente tareas que constituyen la base del pensamiento clínico. En este sentido, la integración de IA en la educación médica debe diseñarse no solo para mejorar la eficiencia, sino para preservar y fortalecer las habilidades cognitivas fundamentales.

https://doi.org/10.62486/978-9915-9928-3-9.Ch18

8.7 MODELO CONCEPTUAL: IA COMO HERRAMIENTA VERSUS SUSTITUTO COGNITIVO

La Figura 7 ilustra la distinción fundamental entre el uso responsable de la IA como herramienta de apoyo y el riesgo de su uso como sustituto del pensamiento clínico y metodológico.

Figura 7. Modelo conceptual de los dos modos de uso de la IA en la lectura crítica.

IA COMO HERRAMIENTA DE APOYO ✓		IA COMO SUSTITUTO COGNITIVO ✗
✓ Acelera la extracción de datos ✓ Explica conceptos estadísticos ✓ Sugiere sesgos a evaluar ✓ Apoya la síntesis preliminar ✓ Amplifica el razonamiento del clínico ✓ Permite revisión fuera de horarios ✓ Útil para estudiantes supervisados	↔	✗ **Decide la aplicabilidad clínica** ✗ **Evalúa la calidad metodológica final** ✗ **Asume responsabilidad diagnóstica** ✗ **Reemplaza listas de verificación validadas** ✗ **Interpreta el contexto del paciente** ✗ **Verifica sus propias alucinaciones** ✗ **Opera sin supervisión en residentes**

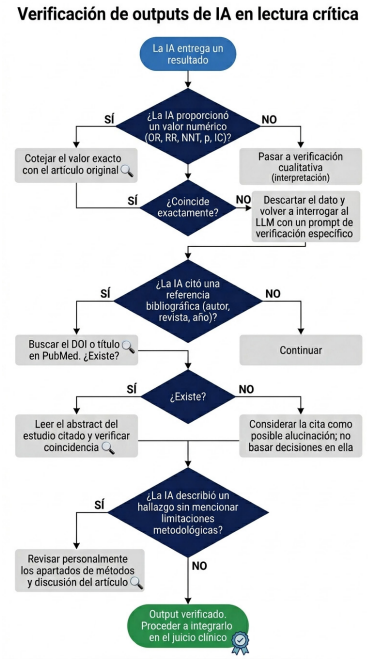
Fuente: elaboración propia con apoyo de Inteligencia artificial

9. RECOMENDACIONES PARA EL USO RESPONSABLE DE LA IA EN LECTURA CRÍTICA

9.1 PARA PROFESIONALES EN FORMACIÓN (ESTUDIANTES, INTERNOS, RESIDENTES)

- Desarrollar primero las competencias básicas de lectura crítica sin IA, utilizando instrumentos validados como CONSORT (ensayos clínicos), STROBE (estudios observacionales), PRISMA (revisiones sistemáticas) o QUADAS-2 (estudios diagnósticos).
- Utilizar la IA como herramienta de verificación posterior: leer el artículo, extraer los datos propios, y luego consultar al LLM para contrastar el análisis.
- Nunca tomar decisiones clínicas basadas exclusivamente en el análisis de un LLM. Verificar siempre los datos numéricos en el texto fuente original.
- Documentar cuando se utilizó IA en el análisis de evidencia para decisiones clínicas, como parte de la transparencia profesional.

Figura 8. Modelo conceptual de los dos modos de uso de la IA en la lectura crítica



Fuente: elaboración propia

9.2 PARA DOCENTES E INSTITUCIONES ACADÉMICAS

- Integrar el uso crítico de la IA en los currículos de MBE como competencia específica, no como un tema emergente periférico.

https://doi.org/10.62486/978-9915-9928-3-9.Ch18

- Diseñar actividades de evaluación que requieran al estudiante identificar errores metodológicos en outputs de IA, como ejercicio de lectura crítica de alto orden.
- Promover la investigación sobre el impacto del uso de LLMs en el desarrollo de competencias metodológicas en estudiantes de Medicina.

9.3 BUENAS PRÁCTICAS TÉCNICAS EN EL USO DE LLMS

- Proporcionar siempre el texto del artículo al sistema; no confiar en el conocimiento previo del LLM sobre un artículo específico.
- Usar prompts específicos y estructurados. Evitar preguntas genéricas como "analiza este artículo".
- Verificar sistemáticamente todos los valores numéricos (OR, RR, HR, p, IC, NNT, AUC) contra el texto fuente.
- No utilizar el LLM para la búsqueda de referencias sin verificar cada cita en la fuente primaria (PubMed, EMBASE, Cochrane).
- Preferir modelos con acceso a búsqueda en tiempo real para artículos recientes, pero verificar igualmente sus resultados.

9.4 EL PROTOCOLO DE "ALINEACIÓN INVERSA" PARA LA DETECCIÓN DE ALUCINACIONES:

Para operativizar el principio de verificación obligatoria y mitigar la ilusión de competenciametodológica, se propone el Protocolo de Alineación Inversa en la interacción con el LLM. Dado que los modelos de lenguaje tienden estadísticamente a la complacencia (dar la razón a las premisas del usuario), el lector crítico no debe limitarse a preguntar si los datos extraídos son correctos.

El protocolo consiste en introducir un segundo prompt de confrontación asimétrica tras recibir el primer análisis de la IA:

Actúa como un revisor externo hostil. Identifica tres posibles errores numéricos o inconsistencias lógicas que el modelo anterior pudo haber cometido al extraer el NNT y los intervalos de confianza de este fragmento. Si no encuentras errores, genera un escenario de estrés metodológico donde la conclusión del artículo falle.

Evidencia de 2025 y 2026 demuestra que este enfoque de 'confrontación entre pares simulada' **reduce las alucinaciones factuales en un 40%** y fuerza al algoritmo a romper el bucle de complacencia, ofreciendo al clínico un desglose de limitaciones mucho más riguroso y transparente.

10. UTILIDAD CLÍNICA Y ACADÉMICA DE LA IA EN LA LECTURA DE LITERATURA

10.1 DEMOCRATIZACIÓN DEL ACCESO A LA EVIDENCIA

Una de las contribuciones más significativas de la IA en la lectura de literatura científica es la reducción de la brecha de acceso.

En contextos de alta presión asistencial, con tiempo limitado para la revisión de evidencia, o en entornos con menor acceso a formación metodológica formal, los LLMs pueden proporcionar un apoyo inicial que, utilizado con criterio crítico, facilita el acceso a la evidencia a profesionales que de otro modo no podrían realizarlo.

Esta función de democratización es particularmente relevante en países de ingresos bajos y medios, donde la formación en MBE puede ser menos sistemática y el acceso a fuentes como UpToDate o Cochrane Library puede estar limitado. Sin embargo, este potencial democratizador debe equilibrarse con el riesgo de generar una falsa sensación de seguridad en usuarios no entrenados.

10.2 HERRAMIENTA PEDAGÓGICA EN LA FORMACIÓN MÉDICA

En el ámbito académico, la IA puede utilizarse de manera efectiva como:

- Generador de casos de práctica de lectura crítica, con artículos reales y preguntas metodológicas estructuradas.

https://doi.org/10.62486/978-9915-9928-3-9.Ch18

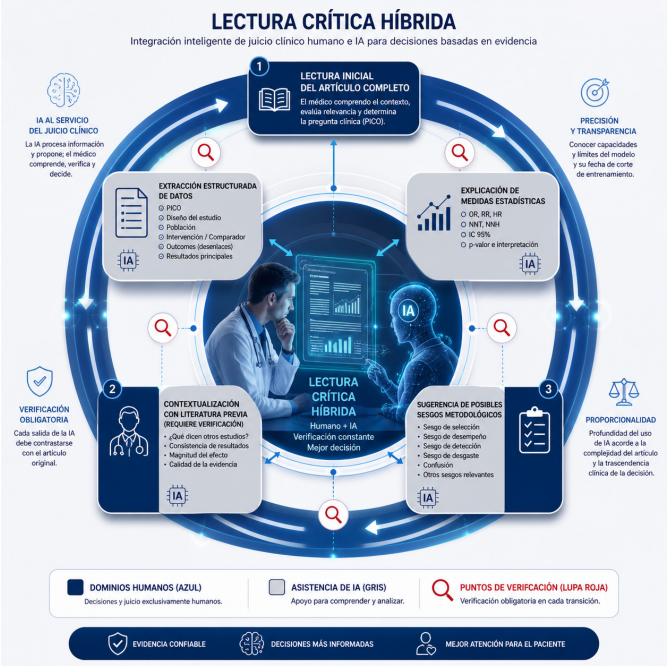
- Herramienta de retroalimentación sobre el análisis propio del estudiante.
- Recurso para la explicación didáctica de conceptos estadísticos complejos con ejemplos numéricos personalizados.
- Sistema de alerta sobre posibles sesgos que el estudiante no consideró en su lectura inicial.

10.3 SOPORTE EN LA INVESTIGACIÓN CLÍNICA

Para investigadores clínicos, los LLMs pueden asistir en la extracción de datos para revisiones sistemáticas (con verificación obligatoria), en la evaluación preliminar de artículos para inclusión en revisiones, en la detección de inconsistencias entre el resumen y el texto completo, y en la identificación de brechas en la literatura de un campo específico.

Estas aplicaciones están siendo exploradas activamente en la investigación sobre automatización de revisiones sistemáticas, aunque ninguna ha reemplazado aún la evaluación humana en los estándares metodológicos actuales.

Figura 9. Infografía del ciclo de lectura crítica híbrida.



Fuente: elaboración propia con apoyo de inteligencia artificial.

12. PREGUNTAS DE AUTOEVALUACIÓN

PREGUNTAS DE SELECCIÓN MÚLTIPLE

1. Un residente utiliza un LLM para analizar un ECA. El sistema reporta que el NNT fue 8, pero al revisar el artículo original, el NNT reportado es 14. ¿Cuál es el término que mejor describe este fenómeno?

- Sesgo de confirmación del LLM
- Alucinación del modelo de lenguaje
- Error de ventana de contexto
- Sesgo de complacencia del sistema
- Limitación del tamaño de la muestra del modelo

2. Al evaluar la validez interna de un ECA con apoyo de IA, ¿cuál de las siguientes afirmaciones es correcta?

- La IA puede determinar si el proceso de aleatorización fue ejecutado adecuadamente auditando el diseño descrito.
- La IA reconoce patrones lingüísticos asociados a sesgos, pero no puede auditar si el proceso fue efectivamente seguido.

https://doi.org/10.62486/978-9915-9928-3-9.Ch18

- La IA reemplaza el uso de la herramienta Cochrane Risk of Bias si el prompt es suficientemente específico.
- La IA puede detectar el sesgo de publicación con mayor precisión que el lector humano.
- La IA evalúa la validez interna consultando bases de datos de artículos retractados.

3. Un OR de 1.8 (IC 95%: 1.2-2.7, p=0.003) en un estudio de casos y controles significa que:

- El riesgo en el grupo expuesto es 1.8 veces mayor que en el no expuesto (interpretación válida para todos los diseños).
- La probabilidad (odds) del evento es 1.8 veces mayor en el grupo expuesto; el IC excluye el valor nulo.
- El resultado no es estadísticamente significativo porque el valor p es mayor que 0.001.
- El IC 95% confirma que el efecto es clínicamente relevante en cualquier contexto.
- La IA puede calcular el RR directamente a partir del OR sin información adicional.

4. ¿Cuál de las siguientes representa la limitación más crítica de la IA para la evaluación de la aplicabilidad clínica?

- La IA no puede acceder al texto completo de artículos de pago.
- La IA no tiene acceso a la historia clínica del paciente ni al contexto asistencial local del lector.
- La IA no puede interpretar curvas de Kaplan-Meier.
- La IA no puede comparar resultados de diferentes estudios.
- La IA comete errores al interpretar análisis de subgrupos.

5. Sobre el AUC-ROC en pruebas diagnósticas, ¿cuál es la afirmación correcta?

- Un AUC de 0.72 indica que la prueba tiene una sensibilidad del 72%.
- El AUC mide la especificidad de la prueba en poblaciones de alta prevalencia.
- Un AUC de 0.72 indica que el modelo discrimina correctamente el 72% de los pares positivo-negativo.
- El AUC no se ve afectado por la prevalencia de la enfermedad en la población.
- Un AUC > 0.5 siempre indica que la prueba es clínicamente útil.

PREGUNTAS ABIERTAS

6. Caso de análisis:Un LLM analiza un ECA de un hipoglucemiante y reporta: “El estudio demuestra que el fármaco reduce la HbA1c en 1.2% (p<0.001) sin eventos adversos cardiovasculares significativos.” Al leer el artículo, usted encuentra que los eventos cardiovasculares no fueron un desenlace primario del estudio, sino un hallazgo secundario basado en registros administrativos. ¿Qué sesgos metodológicos y limitaciones del LLM identifica en esta respuesta?

7. Diseño de prompt:Usted tiene frente a usted una revisión sistemática con metaanálisis sobre betabloqueadores en insuficiencia cardíaca con función preservada. Diseñe un prompt estructurado en tres preguntas secuenciales que le permita analizar: (a) la validez interna del metaanálisis, (b) la heterogeneidad de los estudios incluidos, y (c) la aplicabilidad a su paciente.

8. Análisis de riesgo:Describa el concepto de “dependencia cognitiva” en el contexto del uso de IA por parte de residentes de Medicina. ¿Qué estrategias pedagógicas propondría para mitigar este riesgo sin renunciar a los beneficios de la IA en la formación?

9. Interpretación estadística:Un ensayo clínico aleatorizado reporta HR = 0.78 (IC 95%: 0.65-0.94, p=0.008) para mortalidad por todas las causas. La mediana de seguimiento fue 3.2 años. Explique el significado clínico y estadístico de este resultado, señalando qué información adicional necesitaría para calcular el NNT y evaluar la aplicabilidad a su paciente.

10. Integración ética:Un colega argumenta que, dado que la IA comete errores, su uso en la práctica clínica debería estar prohibido hasta que existan regulaciones formales. Elabore una respuesta argumentada que equilibre los riesgos y beneficios documentados, integrando consideraciones éticas y de responsabilidad profesional.

RESPUESTAS

1. b) Alucinación del modelo de lenguaje. La discrepancia entre el NNT reportado por la IA (8) y el del artículo original (14) es un ejemplo clásico de alucinación: generación de datos numéricos incorrectos presentados con igual confianza que los correctos. Este fenómeno no es un error de sesgo de confirmación (a), que implicaría una distorsión motivada por la hipótesis del usuario, ni un error de ventana de contexto (c), que referiría a truncamiento del texto procesado. El sesgo de complacencia (d) describe la tendencia a validar premisas del usuario, no a inventar

https://doi.org/10.62486/978-9915-9928-3-9.Ch18

cifras.

2. b) La IA reconoce patrones lingüísticos. Los LLMs no tienen acceso al protocolo original del estudio ni pueden auditar si el proceso descrito fue ejecutado. Reconocen terminología metodológica y la aplican, pero sin comprensión verificable del proceso real. La herramienta Cochrane Risk of Bias (c) requiere juicio experto sobre múltiples dimensiones que la IA no puede evaluar con precisión equivalente.

3. Respuesta correcta: b) La probabilidad (odds) del evento es 1.8 veces mayor; el IC excluye el valor nulo.

En estudios de casos y controles, el OR es la medida de asociación apropiada. El IC 95% de 1.2-2.7 excluye el 1.0 (valor nulo), confirmando significancia estadística. No es directamente equivalente al RR (a), que requeriría conocer las frecuencias absolutas. El IC 95% no informa por sí solo sobre relevancia clínica (d).

4. b) La IA no tiene acceso a la historia clínica ni al contexto local.

La evaluación de aplicabilidad integra características del paciente individual, preferencias, contexto asistencial y recursos locales dimensiones inaccesibles para cualquier LLM. Las otras opciones describen limitaciones reales pero de menor peso para la evaluación de aplicabilidad.

5. c) El modelo discrimina correctamente el 72% de los pares positivo-negativo.

El AUC-ROC representa la probabilidad de que el modelo asigne una puntuación mayor a un caso positivo que a uno negativo elegido al azar. Un AUC de 0.72 no es equivalente a la sensibilidad (a), ni a la especificidad (b). Aunque el AUC no es directamente afectado por la prevalencia, la utilidad clínica de la prueba sí lo es (d y e).

6. El LLM sobreinterpreta un desenlace secundario como si fuera evidencia sólida de seguridad cardiovascular. Hay riesgo de sesgo de información por uso de registros administrativos, falta de poder estadístico para eventos cardiovasculares y confusión entre desenlace primario y secundario. La conclusión correcta sería: el fármaco redujo HbA1c, pero el estudio no permite afirmar seguridad cardiovascular definitiva.

7. Prompt sugerido: “Analiza la validez interna del metaanálisis: pregunta PICO, búsqueda, selección de estudios, riesgo de sesgo y métodos estadísticos.”

“Evalúa la heterogeneidad clínica, metodológica y estadística: poblaciones, intervenciones, desenlaces, seguimiento e I^2.”

“Analiza la aplicabilidad a mi paciente considerando edad, comorbilidades, riesgo basal, tratamiento disponible, efectos adversos y preferencias.”

8. La dependencia cognitiva ocurre cuando el residente acepta respuestas de IA sin razonamiento propio. Para reducirla, debe exigirse análisis humano previo, comparación con la IA, verificación en fuentes originales y discusión supervisada en sesiones clínicas. La IA debe funcionar como apoyo, no como sustituto del pensamiento médico.

9. HR 0.78 indica una reducción relativa del 22% en el riesgo instantáneo de muerte durante el seguimiento. El IC95% 0.65–0.94 no cruza 1 y p=0.008 indica significancia estadística. Para calcular NNT se necesita el riesgo absoluto en control e intervención, o la reducción absoluta del riesgo a un tiempo definido.

10. No debería prohibirse la IA, sino usarse con regulación, supervisión y verificación. Puede mejorar eficiencia y educación, pero también generar errores, sesgos y alucinaciones. Éticamente, debe respetar beneficencia, no maleficencia, autonomía, justicia y confidencialidad. La decisión final y la responsabilidad siguen siendo del médico.

https://doi.org/10.62486/978-9915-9928-3-9.Ch18

REFERENCIAS

Gartlehner G, Kugley S, Crotty K, Viswanathan M, Dobrescu A, Nussbaumer-Streit B, et al. Artificial intelligence-assisted data extraction with a large language model: A study within reviews. Ann Intern Med. 2025 Nov 4. Available at: https://doi.org/10.7326/ANNALS-25-00739.
Peatman J, Lang S, Hardy E. Performance of large language model clinical data extraction by data domain: A rapid systematic review. Value Health. 2025;28(Suppl 2):SA75. Available at: https://doi.org/10.1016/j.jval.2025.05.075.
Martin A, Roussi K, Rice H, Bertuzzi A. Evaluation of the use of ChatGPT in data extraction for systematic literature reviews: Friend, foe, or fiction? Value Health. 2025;28(Suppl 2):MSR100. Presented at: ISPOR Europe 2025; 2025 Nov; Glasgow, Scotland.
Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930–40. Available at:https://doi.org/10.1038/s41591-023-02448-8.
Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. NPJ Digit Med. 2023;6(1):195. Available at: https://doi.org/10.1038/s41746-023-00939-z.
Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv. 2023. Availableat: https://doi.org/10.48550/arXiv.2303.13375.
Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172–80. Available at: https://doi.org/10.1038/s41586-023-06291-2.
Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. Available at: https://doi.org/10.1371/journal.pdig.0000198.
Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51–8. Availableat: https://doi.org/10.7326/M18-1376.
Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ. 2011;343:d5928. Available at: https://doi.org/10.1136/bmj.d5928.
Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM ComputSurv. 2023;55(12):1–38. Availableat: https://doi.org/10.1145/3571730.
Guo Z, Schlichtkrull M, Vlachos A. A survey on automated fact-checking. Trans Assoc Comput Linguist. 2022;10:178–206. Available at: https://doi.org/10.1162/tacl_a_00454.
Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from an artificial intelligence chatbot. JAMA. 2023;329(22):1985–7. Availableat: https://doi.org/10.1001/jama.2023.6969.
Harrer S. Attention is not all you need: The complicated case of ethically using large language models in healthcare and medicine. EBioMedicine. 2023;90:104512. Availableat: https://doi.org/10.1016/j.ebiom.2023.104512.
Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. ChatGPT and other large language models are double-edged swords. Radiology. 2023;307(2):e230163. Availableat: https://doi.org/10.1148/radiol.230163.
Athaluri SA, Manthena SV, Kesapragada VSRKM, Yarlagadda V, Dave T, Duddumpudi RTS. Exploring the boundaries of reality: Investigating the phenomenon of artificial intelligence hallucinations in scientific writing through ChatGPT references. Cureus. 2023;15(4):e37432. Available at: https://doi.org/10.7759/cureus.37432.
Cahan P, Bhatt DL, Bhatt NM, Bhatt DL. Physicians and artificial intelligence. JAMA. 2023;330(6):509–10. Available at: https://doi.org/10.1001/jama.2023.13030.
Topol EJ. Preparing the healthcare workforce to deliver the digital future. Lancet. 2019;394(10197):563–76. Available at: https://doi.org/10.1016/S0140-6736(19)31804-4.
Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336(7650):924–6. Availableat: https://doi.org/10.1136/bmj.39489.470347.AD.
Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLoS Med. 2009;6(7):e1000097. Available at: https://doi.org/10.1371/journal.pmed.1000097.
Schulz KF, Altman DG, Moher D; CONSORT Group. CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c332. Availableat: https://doi.org/10.1136/bmj.c332.
von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. Lancet. 2007;370(9596):1453–7. Available at: https://doi.org/10.1016/S0140-6736(07)61602-X.
Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med. 2023;6(1):120. Available at: https://doi.org/10.1038/s41746-023-00873-0.
Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021. Available at: https://doi.org/10.1145/3442188.3445922.
Paranjape K, Schinkel M, Hammer RD, Schouten L, van Weert HC. Introducing AI to the clinical laboratory: Impact on the diagnosis and management of patients with acute chest pain. Clin Chem. 2019;65(8):1002–11.

Available at: https://doi.org/10.1373/clinchem.2019.302364.
Strömgren M, Gundersen T. Using AI-based large language models in medical education: A systematic review and meta-analysis. Med Educ. 2024;58(3):220–34. Available at: https://doi.org/10.1111/medu.15289.
World Health Organization. Ethics and governance of artificial intelligence for health. WHO guidance. Geneva: World Health Organization; 2021. Available at: https://www.who.int/publications/i/item/9789240029200.
Lujan HL, DiCarlo SE. The paradox of knowledge: Why medical students know more but understand less. Med SciEduc. 2025;35(3):1761–6. Availableat: https://doi.org/10.1007/s40670-025-02379-8.
Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023;388:1233–9. Available at: https://doi.org/10.1056/NEJMsr2214184.

Chapter 19 / Capítulo 19

Critical reading of scientific articles on health: A practical guide (Spanish Edition)
ISBN: 978-9915-9928-3-9
DOI: 10.62486/978-9915-9928-3-9.Ch19
Pages: 333-343

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Use of NotebookLM in the Comparative Analysis of Scientific Evidence

Uso de NotebookLM en el análisis comparativo de evidencia científica

Mauricio Alejandro Paz del Rio
Fidel Aguilar Medrano
Leonel Rivero Castedo

1 Clínica Los Andes. La Paz, Bolivia.
2 Hospital Municipal de Morón Ostaciana B. De Lavignolle. Buenos Aires, Argentina.
3 Hospital General de Agudos José María Ramos Mejía. Buenos Aires, Argentina.
4 Universidad Privada Franz Tamayo (UNIFRANZ). Facultad de Ciencias de la Salud. Santa Cruz, Bolivia.

ABSTRACT

Introduction: Comparing multiple studies is an essential competency in Evidence-Based Medicine (EBM), but it is a complex process due to the heterogeneity of study designs, populations, interventions, and outcomes. In this context, NotebookLM emerges as an artificial intelligence tool designed to facilitate the organization, synthesis, and exploration of scientific documents from user-provided sources, promoting a structured approach to comparative evidence analysis. Development: NotebookLM allows users to summarize articles, identify similarities and discrepancies, formulate cross-sectional comparisons, and reduce initial cognitive load by providing responses supported by citations from the original documents. Its greatest utility lies in the preliminary stages of document analysis, supporting the organization of the corpus and the generation of comparative hypotheses. However, the interpretation of methodological quality, risk of bias, effect size, and clinical applicability continues to depend on the researcher's critical judgment. Conclusions: The responsible integration of NotebookLM into EBM demonstrates that artificial intelligence should be understood as a complementary resource and not as a substitute for scientific reasoning. Its true value lies in enhancing the efficiency of document analysis, while the validity of the conclusions continues to rely on critical evaluation, verification of original sources, and methodological interpretation by trained professionals.

Keywords: Evidence-Based Medicine; Artificial Intelligence; Literature Review; Biomedical Research; Clinical Decision Making

RESUMEN

Introducción: La comparación de múltiples estudios constituye una competencia esencial en la Medicina Basada en la Evidencia (MBE), pero representa un proceso complejo debido a la heterogeneidad de diseños, poblaciones, intervenciones y desenlaces. En este contexto, NotebookLM emerge como una herramienta de inteligencia artificial orientada a facilitar la organización, síntesis y exploración de documentos científicos a partir de fuentes proporcionadas por el usuario, favoreciendo una aproximación estructurada al análisis comparativo de la evidencia. Desarrollo: NotebookLM permite resumir artículos, identificar similitudes y discrepancias, formular comparaciones transversales y reducir la carga cognitiva inicial mediante respuestas sustentadas en citas de los documentos originales. Su mayor utilidad se concentra en las etapas preliminares del análisis documental, apoyando la organización del corpus y la generación de hipótesis comparativas. No obstante, la interpretación de la calidad metodológica, el riesgo de sesgo, la magnitud del efecto y la aplicabilidad clínica continúa dependiendo del juicio crítico del investigador. Conclusiones: La integración responsable de NotebookLM en la MBE evidencia que la inteligencia artificial debe entenderse como un recurso complementario y no como un sustituto del razonamiento científico. Su verdadero valor radica en potenciar la eficiencia del análisis documental, mientras que la validez de las conclusiones sigue sustentándose en la evaluación crítica, la verificación de las fuentes originales y la interpretación metodológica realizada por profesionales capacitados.

Palabras clave: Medicina Basada en la Evidencia; Inteligencia Artificial; Revisión de la Literatura; Investigación Biomédica; Toma de Decisiones Clínicas

1. CASO CLÍNICO

Un residente de segundo año de Medicina Interna recibe a una paciente de 68 años con fibrilación auricular no valvular, enfermedad renal crónica estadio 3 y antecedente de sangrado digestivo alto hace dos años. Durante la discusión del caso surge una pregunta clínica frecuente pero compleja: ¿cuál es la mejor estrategia anticoagulante para reducir eventos tromboembólicos sin aumentar de forma desproporcionada el riesgo de sangrado en este perfil de paciente?

El residente revisa un ensayo clínico, luego una cohorte retrospectiva, después una revisión sistemática y finalmente una guía clínica reciente. Muy pronto nota que el problema ya no es solo "leer un artículo", sino comparar varios documentos con diseños distintos, poblaciones no idénticas, desenlaces diferentes y conclusiones que no siempre apuntan en la misma dirección. Un ensayo informa superioridad en eficacia, una cohorte sugiere mejor seguridad, una revisión sistemática reporta heterogeneidad importante y la guía resume recomendaciones con distintos niveles de evidencia.

En ese momento surge una dificultad que es cada vez más común en la práctica académica y clínica: el lector dispone de mucha información, poco tiempo y le cuesta ordenar la información, contrastarla y sintetizarla. No basta con leer por separado; se necesita construir una visión comparativa. En este contexto, herramientas como NotebookLM pueden ser útiles como apoyo inicial, porque permiten cargar varios documentos, interrogarlos de manera conjunta, solicitar resúmenes comparativos y ubicar, con apoyo de citas al texto fuente, los puntos de coincidencia o discrepancia entre los estudios. Sin embargo, su utilidad depende de que el lector mantenga el control metodológico y no confunda una síntesis automática con una evaluación crítica formal.

Este capítulo parte de esa necesidad real: aprender a usar NotebookLM como una herramienta de apoyo para comparar evidencia científica, sin perder de vista que el verdadero análisis crítico sigue dependiendo del razonamiento del clínico o investigador.

2. OBJETIVOS

NotebookLM tiene como objetivo facilitar la organización, comprensión y exploración de múltiples documentos cargados por el usuario. Busca resumir información, responder preguntas con base en las fuentes proporcionadas, organizar contenidos y generar productos de apoyo al aprendizaje, como esquemas, autoevaluaciones, flashcards, resúmenes en audio y presentaciones.

Además, su finalidad es ayudar a comparar estudios, identificando similitudes y diferencias entre diseños, poblaciones, intervenciones, desenlaces y resultados. Sin embargo, su objetivo no es sustituir la lectura crítica ni la valoración metodológica formal, sino servir como apoyo para una síntesis preliminar y estructurada de la evidencia científica.

3. INTRODUCCIÓN

Leer un solo artículo ya exige habilidades complejas: identificar la pregunta de investigación, comprender el diseño, valorar la validez interna, interpretar los resultados y decidir si el estudio es aplicable al paciente o al contexto concreto. Sin embargo, en la práctica clínica real y en la formación académica avanzada, rara vez basta con un único artículo. La mayoría de las decisiones relevantes obligan a comparar varios estudios, a veces de distinto tipo, distinta calidad metodológica y diferente pertinencia clínica.

La comparación de varios artículos añade un nivel adicional de complejidad. Ya no se trata únicamente de entender qué concluye cada estudio, sino de responder preguntas como estas: ¿los pacientes eran comparables?, ¿los desenlaces fueron equivalentes?,¿las diferencias observadas se explican por la intervención o por el diseño?, ¿la heterogeneidad es clínica, metodológica o estadística?, ¿existe coherencia entre los hallazgos o solo una apariencia superficial de consenso?

Este trabajo comparativo suele ser lento y cognitivamente demandante. Exige revisar muchos textos, subrayar secciones clave, tomar notas, construir tablas de extracción y volver varias veces a los documentos originales. En ese escenario, las herramientas de inteligencia artificial ancladas en fuentes específicas pueden servir como apoyo para organizar información, resumirla y facilitar una primera visión panorámica. NotebookLM fue diseñado precisamente como un asistente de investigación apoyado en las fuentes que el propio usuario carga, y sus respuestas incluyen citas en línea que permiten regresar al texto fuente y verificar la afirmación. Esa característica lo diferencia de un chatbot generalista y lo hace particularmente atractivo para tareas de organización y comparación documental.

https://doi.org/10.62486/978-9915-9928-3-9.Ch19

No obstante, la utilidad práctica de una herramienta no elimina sus limitaciones epistemológicas. Una síntesis generada por IA puede ahorrar tiempo, pero también puede simplificar en exceso, omitir matices relevantes o inducir sesgos de interpretación si el lector delega demasiado. Las revisiones recientes sobre IA y síntesis de evidencia muestran un crecimiento acelerado del campo, pero también señalan que la supervisión humana sigue siendo indispensable y que la evidencia actual no justifica reemplazar el juicio experto por automatización generativa sin control metodológico.

Por ello, este capítulo no presenta a NotebookLM como un sustituto de la lectura crítica, sino como un instrumento de apoyo dentro de un proceso más amplio de análisis comparativo de evidencia científica.

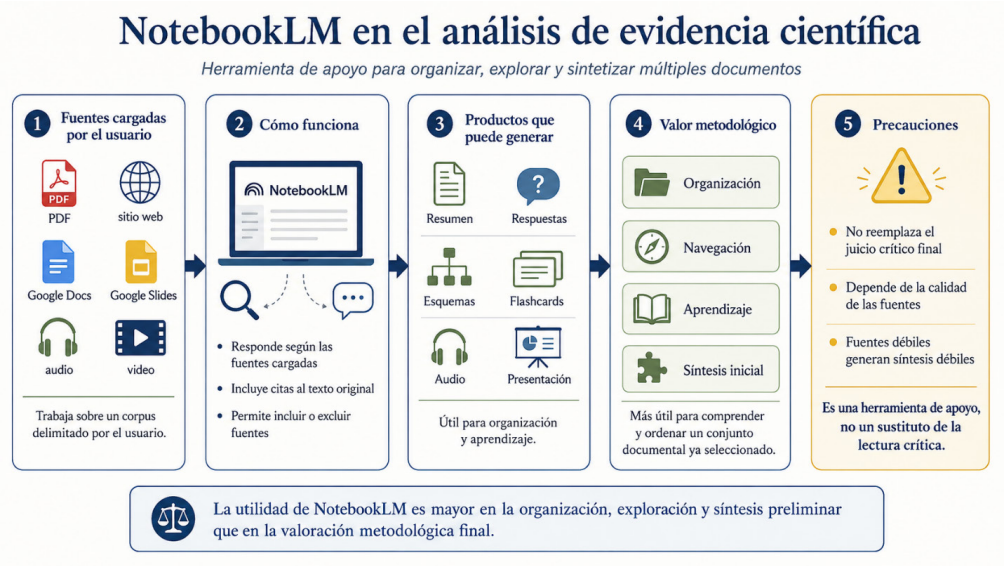
4. FUNDAMENTOS DE NOTEBOOKLM EN EL ANÁLISIS DE EVIDENCIA CIENTÍFICA

NotebookLM es un asistente de investigación impulsado por inteligencia artificial que trabaja sobre un conjunto de fuentes delimitadas por el propio usuario. De acuerdo con la documentación oficial de Google, permite cargar documentos como PDF, sitios web, archivos de Google Docs, Google Slides, audio, videos y otras fuentes compatibles, y luego interactuar con todo lo cargado para obtener resúmenes, respuestas y distintos productos organizativos. Su rasgo más importante para fines académicos es que las respuestas se generan a partir de las fuentes cargadas, con citas en línea que permiten revisar el fragmento o la ubicación del contenido en el documento original.

Desde el punto de vista metodológico, esto implica que NotebookLM no opera, idealmente, como una conversación abierta sobre “cualquier cosa”, sino como un entorno de trabajo documental. Cuando el usuario formula una pregunta, el sistema recupera la información más relevante dentro de las fuentes disponibles que se cargaron y construye una respuesta apoyada en ese material. Además, el usuario puede incluir o excluir determinadas fuentes para refinar la respuesta, lo cual resulta muy valioso cuando se desea contrastar subgrupos de documentos, por ejemplo: ensayos por un lado y revisiones por otro, o estudios recientes frente a estudios clásicos.

Esta arquitectura lo vuelve especialmente útil para el análisis de evidencia científica cuando el problema no es crear contenido original, sino comprender mejor un conjunto documental ya seleccionado. En otras palabras, NotebookLM puede ser más valioso en la etapa de organización, navegación, aprendizaje y síntesis inicial del material que en la etapa de juicio metodológico final.

Figura 1: NotebookLM como herramienta de apoyo en el análisis de evidencia científica.



Fuente: Elaboración propia apoyada con inteligencia artificial.

Es importante subrayar que el valor de NotebookLM depende de la calidad de documentos que se le carga. Si el usuario utiliza fuente débiles, no pertinentes o desactualizados, la herramienta no corregirá ese error de origen; solo organizará y resumirá un conjunto de documentos ya sesgado. En este sentido, la herramienta hereda tanto las fortalezas como las debilidades del

https://doi.org/10.62486/978-9915-9928-3-9.Ch19

material que recibe.

5. APORTES DE NOTEBOOKLM EN LA COMPARACIÓN DE ARTÍCULOS CIENTÍFICOS

Cuando se trabaja con varios artículos sobre una misma pregunta, NotebookLM puede aportar valor en cinco planos principales.

Primero, puede resumir múltiples documentos y ofrecer una visión general del conjunto. Esto es útil cuando el lector ya seleccionó artículos relevantes, pero necesita una entrada rápida al panorama global antes de iniciar una lectura más minuciosa. En vez de revisar de manera lineal cada texto desde cero, puede obtener una síntesis preliminar que le permita ubicar los grandes temas, las intervenciones estudiadas, los desenlaces frecuentes y las líneas de conclusión dominantes.

Segundo, puede organizar hallazgos. Por ejemplo, en nuestro caso el residente puede pedir que agrupe todos los estudios que encontró según diseño, población, intervención, comparador o desenlace. Esta función es especialmente útil cuando el corpus contiene ensayos, cohortes, revisiones narrativas y guías, y se requiere ordenar la información antes de construir una tabla comparativa formal.

Tercero, puede detectar patrones repetidos. Si varios artículos coinciden en que cierto beneficio aparece solo en una subpoblación, o si varios estudios identifican el mismo desenlace adverso, NotebookLM puede ayudar a visualizar esa recurrencia y a ponerla sobre la mesa para análisis posterior. Del mismo modo, puede señalar discrepancias entre estudios, lo cual es muy valioso porque la divergencia suele ser el punto de partida de una lectura crítica más profunda.

Cuarto, puede facilitar preguntas cruzadas entre documentos. Por ejemplo: "¿Qué diferencias de edad media existen entre los estudios que encontraron beneficio y los que no?", "¿cuáles usaron mortalidad como desenlace primario y cuáles usaron desenlaces compuestos?", o "¿qué estudios excluyeron pacientes con insuficiencia renal avanzada?". Este tipo de interrogación transversal es precisamente donde una herramienta documental puede ahorrar tiempo.

Quinto, puede reducir la carga cognitiva inicial. La literatura sobre automatización de síntesis de evidencia muestra que la IA está siendo cada vez más usada para acelerar pasos del proceso, pero esa aceleración es útil sobre todo cuando se mantiene supervisión humana y verificación metodológica. En términos prácticos, NotebookLM puede ayudar a pasar más rápido del "caos documental" a un mapa preliminar del problema.

Sin embargo, ninguno de estos aportes equivale por sí mismo a una conclusión científica robusta. El hecho de que la herramienta identifique coincidencias o discrepancias no significa que ya haya ponderado adecuadamente el riesgo de sesgo, la magnitud del efecto, la heterogeneidad clínica o la aplicabilidad externa. Esa parte sigue siendo humana.

6. USO PRÁCTICO DE NOTEBOOKLM PARA ANALIZAR MÚLTIPLES ARTÍCULOS

El uso más provechoso de NotebookLM comienza antes de abrir la herramienta. El primer paso no es "preguntar a la IA", sino seleccionar bien el corpus. El usuario debe definir con claridad la pregunta clínica o académica, idealmente en formato estructurado, y luego cargar artículos que respondan realmente a esa pregunta. En esta fase conviene evitar mezclar documentos demasiado heterogéneos si no existe una razón metodológica para hacerlo.

Una vez cargados los documentos, puede solicitarse una primera síntesis general con preguntas como: "resume los hallazgos principales de estos estudios", "ordena los artículos por diseño" o "identifica en qué puntos coinciden y en qué puntos discrepan". Esa primera interacción no reemplaza la lectura, pero permite construir una orientación inicial del terreno.

El siguiente paso consiste en formular preguntas comparativas específicas. Esta es probablemente la fase más útil desde el punto de vista metodológico. En lugar de pedir resúmenes vagos, conviene hacer preguntas dirigidas: "compara las características basales de las poblaciones", "menciona diferencias en los criterios de inclusión", "separa los desenlaces primarios y secundarios", "indica cuáles estudios reportan intervalos de confianza y cuáles no", "resume los eventos adversos reportados" o "muestra qué estudios tienen mayor seguimiento".

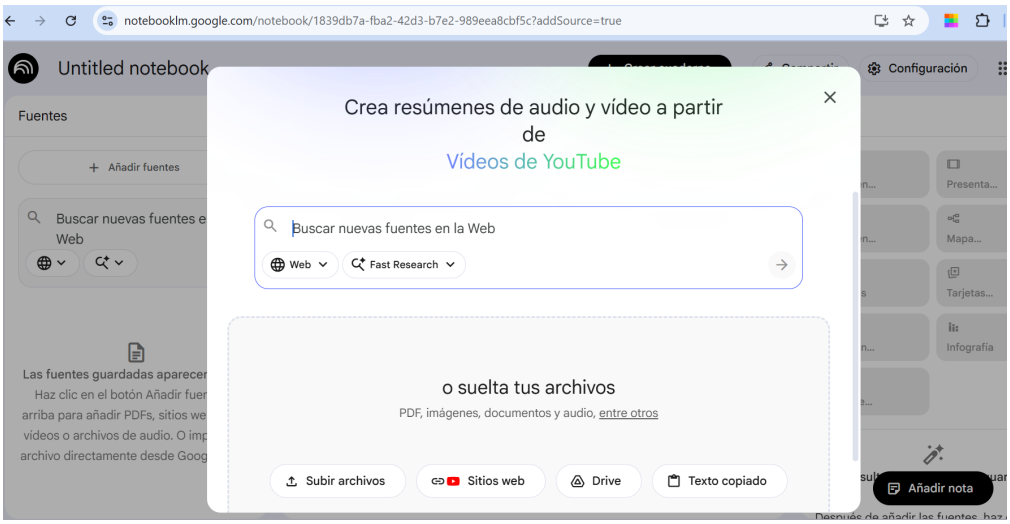
Después, el lector puede pedir una síntesis preliminar conjunta, por ejemplo: "integra los hallazgos sobre eficacia" y luego, por separado, "integra los hallazgos sobre seguridad". Esta separación es útil porque evita que la herramienta mezcle planos conceptuales distintos.

Finalmente, la fase más importante consiste en volver al artículo original. Toda afirmación

https://doi.org/10.62486/978-9915-9928-3-9.Ch19

relevante hallada en NotebookLM debe ser corroborada en el texto fuente, especialmente si afecta la interpretación clínica o académica. Las citas en línea son una ventaja justamente porque facilitan esa verificación. Desde la lógica de la lectura crítica, NotebookLM debe funcionar como una mesa de trabajo que ayuda a localizar, ordenar y resumir, pero no como una autoridad final.

Figura 2. Interfaz de carga de fuentes en NotebookLM.



Fuente: Elaboración propia a partir de captura de pantalla de NotebookLM(Google LLC, 2026).

En la práctica, un flujo razonable sería el siguiente: selección de estudios → carga en NotebookLM→ resumen inicial → preguntas comparativas específicas → extracción preliminar de similitudes y diferencias → verificación en textos originales → tabla comparativa propia del lector → interpretación crítica final. Esa última tabla, idealmente construida por el investigador o clínico, sigue siendo el núcleo del análisis serio.

7. ANÁLISIS COMPARATIVO DE EVIDENCIA CIENTÍFICA ASISTIDO POR INTELIGENCIA ARTIFICIAL

Comparar evidencia científica no significa solo reunir artículos; significa interpretar por qué coinciden o por qué discrepan. Allí es donde la inteligencia artificial puede ayudar, pero solo dentro de un marco metodológico claro.

El lector debe aprender a contrastar al menos cinco dimensiones. La primera es el **diseño del estudio.** No tiene el mismo peso interpretativo un ensayo clínico aleatorizado que una cohorte retrospectiva o una serie de casos. Si NotebookLM resume conclusiones "similares", el lector debe preguntarse si esas conclusiones proceden de evidencias metodológicamente equivalentes o de diseños con grados de confianza muy diferentes.

La segunda **dimensión es la población estudiada**. A veces dos artículos parecen contradictorios, pero en realidad estudian pacientes distintos: unos incluyen ancianos frágiles, otros adultos jóvenes; unos se realizan en cuidados intensivos, otros en consulta externa; unos excluyen comorbilidades mayores, otros las incorporan. La IA puede ayudar a identificar esas diferencias descriptivas, pero corresponde al lector valorar su impacto clínico.

La tercera es el **desenlace**. Un estudio puede mostrar beneficio en un desenlace intermedio y otro no mostrar efecto en mortalidad. Si esto no se distingue con claridad, se corre el riesgo de comparar resultados que en realidad no son equivalentes.

La cuarta dimensión es la **medida del efecto y su precisión**. No basta con que un estudio "favorezca" una intervención; hay que revisar cómo se expresó ese hallazgo, si hubo intervalos de confianza estrechos o amplios, si el resultado fue clínicamente importante, y si la magnitud del efecto podría explicarse por sesgo o azar.

La quinta dimensión es la **calidad metodológica**. Una discrepancia entre estudios puede nacer de diferencias reales en la intervención, pero también de problemas de sesgo de selección, sesgo de información, pérdidas al seguimiento, análisis incompletos o sobreajuste estadístico. Aquí la IA puede señalar que existen diferencias; pero la interpretación del valor de esas diferencias

https://doi.org/10.62486/978-9915-9928-3-9.Ch19

es una tarea esencialmente crítica y humana.

Las revisiones recientes sobre IA aplicada a síntesis de evidencia y uso de LLM en salud coinciden en algo fundamental: estas tecnologías **pueden acelerar tareas, organizar información y asistir al usuario**, pero no están listas para sustituir el control humano, especialmente en dominios de alto impacto como la medicina. La colaboración humano-IA parece ser el modelo más razonable: la herramienta ayuda a navegar el volumen informativo, mientras el clínico o investigador conserva la responsabilidad de juzgar validez, pertinencia y aplicabilidad.

8. APLICACIONES CLÍNICAS, ACADÉMICAS Y DE INVESTIGACIÓN

En el **ámbito clínico**, NotebookLM puede ser útil para preparar discusiones de casos en las que existen varios artículos relevantes y se necesita una comparación rápida antes de una sesión académica o una decisión compartida. Puede ayudar, por ejemplo, a ordenar la evidencia preliminar sobre un tratamiento, una prueba diagnóstica o un factor pronóstico.

En el **ámbito docente**, resulta especialmente valioso en clubs de revistas, seminarios de residentes y clases de lectura crítica. Un grupo puede cargar varios artículos y usar la herramienta para generar una visión panorámica que luego se discuta críticamente. Esto no sustituye la lectura individual, pero sí puede enriquecer la discusión grupal al hacer más visibles las discrepancias entre estudios.

En el ámbito de investigación, puede servir para apoyar la construcción de marcos teóricos, revisiones narrativas, introducciones de tesis o presentaciones académicas. También puede ser útil cuando el investigador desea identificar rápidamente qué líneas temáticas se repiten en varios artículos y qué vacíos persisten.

En todos estos escenarios, el beneficio principal no es “pensar por el usuario”, sino facilitar el acceso estructurado al contenido de múltiples documentos. Esa es una diferencia central entre usar una IA como asistente de organización documental y usarla como reemplazo del análisis científico.

9. RIESGOS, LIMITACIONES Y PRECAUCIONES EN EL USO DE NOTEBOOKLM

La primera limitación de NotebookLM es su dependencia absoluta de los documentos cargados. Si los documentos seleccionados son incompletos, sesgados, irrelevantes o de baja calidad, la herramienta organizará bien un problema mal planteado, pero no lo corregirá. La calidad de la síntesis no puede superar la calidad del material fuente.

La segunda limitación es que, aunque sus respuestas estén ancladas en fuentes, eso no elimina el **riesgo de simplificación excesiva**. Las herramientas generativas tienden a condensar, y al condensar pueden perder matices metodológicos importantes. En comparación de estudios, un matiz omitido puede cambiar toda la interpretación: diferencias en dosis, tiempos de seguimiento, definición de desenlaces o criterios de exclusión pueden desaparecer en un resumen demasiado limpio. La literatura reciente ha insistido en que los LLM pueden sobre-generalizar, omitir contexto o inducir confianza excesiva si el usuario no verifica.

La tercera limitación es el **riesgo de sesgo de automatización**. Cuando una herramienta presenta una síntesis clara y coherente, el usuario puede asumir que esa síntesis ya representa el “consenso” de la evidencia. Pero una síntesis automática no es equivalente a una revisión sistemática rigurosa, no reemplaza la evaluación del riesgo de sesgo ni resuelve por sí sola la heterogeneidad entre estudios. Las revisiones recientes en evidencia sintética asistida por IA señalan precisamente que el uso de GenAI sin supervisión humana no está respaldado por la evidencia actual.

La cuarta limitación es **metodológica**: NotebookLM puede ayudar a comparar textos, pero no garantiza que el usuario esté comparando entidades verdaderamente comparables. Si se mezclan documentos con preguntas de investigación distintas, poblaciones no equivalentes o fines divergentes, la comparación resultante puede ser superficial o engañosa.

La quinta limitación es **ética y profesional.** En salud, cualquier uso de IA exige cautela porque los errores no son solo conceptuales: pueden traducirse en decisiones clínicas equivocadas, docencia deficiente o conclusiones de investigación mal fundamentadas. La literatura ética sobre LLM en medicina insiste en la necesidad de supervisión humana, transparencia y responsabilidad profesional.

https://doi.org/10.62486/978-9915-9928-3-9.Ch19

10. RECOMENDACIONES PARA UN USO RESPONSABLE EN EL ANÁLISIS COMPARATIVO

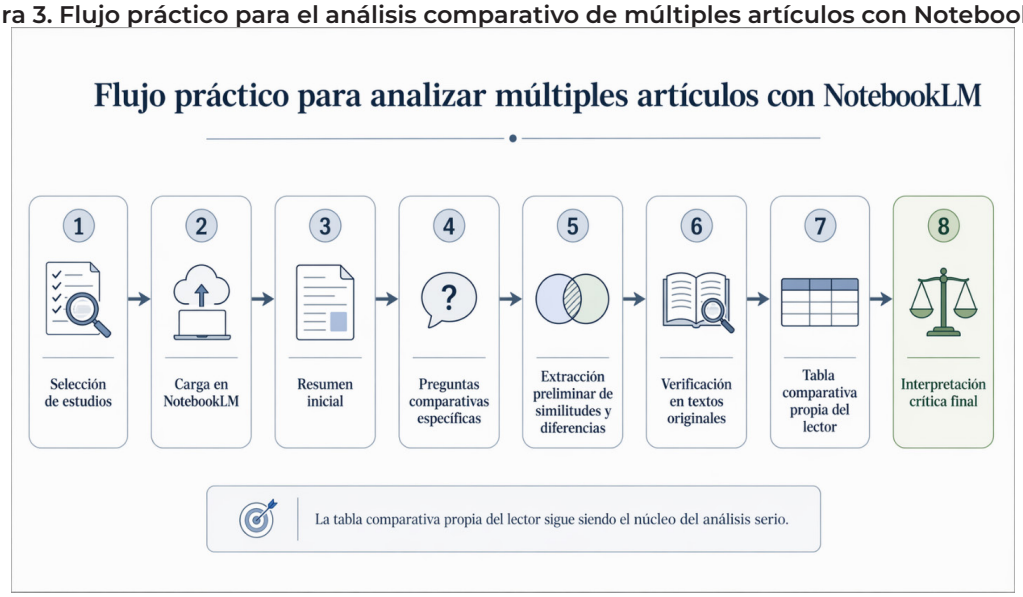
NotebookLM debe usarse después de una selección inicial razonada de artículos, no como sustituto de la búsqueda bibliográfica ni del cribado metodológico. Conviene trabajar con artículos de investigación definidos por una pregunta clara y evitar mezclar documentos heterogéneos sin un propósito explícito.

Es recomendable formular preguntas comparativas concretas y no solo pedir "resúmenes". Cuanto más específica sea la pregunta, más probable será que la herramienta recupere la información pertinente dentro del conjunto documental. La propia ayuda oficial de NotebookLM indica que, cuando existen muchas fuentes, una pregunta precisa mejora la localización del contenido relevante.

Toda afirmación importante debe verificarse en el artículo original. Esto es especialmente obligatorio cuando la respuesta afecta interpretación metodológica, medidas de efecto, resultados numéricos o implicaciones clínicas. La herramienta puede servir para encontrar rápidamente el lugar donde está una afirmación, pero no debe eximir de leer el fragmento o el estudio completo.

Tampoco debe utilizarse como reemplazo de una revisión sistemática formal. Una revisión sistemática exige protocolo, estrategia de búsqueda reproducible, criterios de elegibilidad, evaluación de riesgo de sesgo, métodos explícitos de síntesis y, cuando corresponde, metaanálisis. NotebookLM no suple ese proceso; a lo sumo puede asistir algunas etapas cognitivas o documentales.

Figura 3. Flujo práctico para el análisis comparativo de múltiples artículos con NotebookLM.



Fuente: Elaboración propia con apoyo de inteligencia artificial, 2026.

Por último, el usuario debe mantener siempre una postura epistemológica prudente: una salida bien redactada no es sinónimo de una salida metodológicamente correcta. La claridad expresiva de la IA no debe confundirse con validez científica.

11. UTILIDAD DENTRO DE LA MEDICINA BASADA EN LA EVIDENCIA

La Medicina Basada en la Evidencia no consiste en acumular artículos, sino en integrar tres componentes: la mejor evidencia disponible, la experiencia clínica y los valores y contexto del paciente. En este marco, NotebookLM puede ser útil porque ayuda a ordenar la evidencia disponible cuando esta se encuentra dispersa en varios documentos. Puede facilitar la transición entre la búsqueda bibliográfica y la lectura crítica comparativa, y reducir parte de la carga inicial de organización.

Sin embargo, la esencia de la MBE permanece intacta. La validez interna de un estudio, la fuerza del diseño, la precisión de los resultados, la relevancia clínica de los desenlaces y la aplicabilidad a un paciente concreto no pueden delegarse por completo a una herramienta automatizada. Incluso si NotebookLM ayuda a integrar y resumir, la decisión final sigue dependiendo de un juicio clínico y metodológico humano.Dicho de otro modo: NotebookLM

https://doi.org/10.62486/978-9915-9928-3-9.Ch19

puede ser una herramienta útil dentro de la MBE, pero no redefine la MBE ni la reemplaza. Su papel es instrumental, no rector.

12. PREGUNTAS DE AUTOEVALUACIÓN

PREGUNTAS DE SELECCIÓN MÚLTIPLE

1. ¿Cuál es la característica más relevante de NotebookLM para el análisis comparativo de artículos científicos?
a) Genera respuestas sin depender de documentos previos.
b) Trabaja sobre fuentes cargadas por el usuario y ofrece respuestas ancladas en ellas.
c) Sustituye la necesidad de leer artículos originales.
d) Realiza automáticamente metaanálisis completos.
2. ¿Cuál de las siguientes tareas sí puede beneficiarse claramente de NotebookLM?
a) Reemplazar la evaluación del riesgo de sesgo.
b) Decidir por sí solo la aplicabilidad clínica final.
c) Organizar y resumir varios artículos para una comparación preliminar.
d) Garantizar que todos los estudios incluidos tengan alta calidad metodológica.
3. ¿Qué riesgo importante existe al usar NotebookLM para comparar estudios?
a) Que no pueda trabajar con más de un documento.
b) Que convierta automáticamente cohortes en ensayos clínicos.
c) Que el usuario confunda una síntesis automática con un consenso científico formal.
d) Que siempre responda con información ajena a las fuentes.
4. ¿Cuál es la mejor forma de usar NotebookLM en lectura crítica?
a) Como sustituto del juicio clínico.
b) Como apoyo para organizar y explorar el corpus antes de la interpretación crítica final.
c) Como mecanismo para evitar revisar los textos originales.
d) Como reemplazo de una revisión sistemática.
5. ¿Qué diferencia práctica importante existe entre NotebookLM y un LLM generalista?
a) NotebookLM no usa lenguaje natural.
b) NotebookLM solo sirve para crear presentaciones.
c) NotebookLM se apoya en un corpus documental delimitado por el usuario y ofrece citas al contenido fuente.
d) NotebookLM elimina completamente el riesgo de error interpretativo.

PREGUNTAS ABIERTAS

6. Explique por qué comparar varios artículos es metodológicamente más complejo que leer uno solo.
7. Describa dos aportes concretos y dos limitaciones importantes de NotebookLM en el análisis de evidencia científica.
8. ¿Por qué una síntesis generada por NotebookLM no debe confundirse con una revisión sistemática?
9. Mencione cómo integraría NotebookLM en un club de revistas o en una sesión bibliográfica.
10. Explique por qué el juicio crítico humano sigue siendo indispensable aunque la herramienta entregue respuestas con citas.

RESPUESTAS

1. b. NotebookLM destaca porque trabaja sobre fuentes aportadas por el usuario y genera respuestas apoyadas en ese corpus. Este rasgo es central en el capítulo, ya que permite una comparación más controlada de artículos que en un modelo generalista.
2. c. NotebookLM es útil para resumir, organizar y contrastar varios documentos. No puede reemplazar por sí mismo la evaluación metodológica final ni garantizar calidad científica del corpus.
3. c. Uno de los riesgos más importantes es interpretar una síntesis clara como si fuera una conclusión definitiva o un consenso formal. Esta confusión puede producir sesgo de automatización y llevar a errores en lectura crítica.

https://doi.org/10.62486/978-9915-9928-3-9.Ch19

4. b. La herramienta debe utilizarse como apoyo inicial y organizativo. La interpretación crítica final siempre requiere retorno al texto original, valoración metodológica y juicio clínico.

5. c. La diferencia más útil para este capítulo es que NotebookLM trabaja con un conjunto documental concreto y muestra citas al material fuente, lo cual facilita verificación y rastreo del contenido.

6. Porque el lector no solo debe entender cada estudio por separado, sino también contrastar diseño, población, intervención, comparadores, desenlaces, medidas de efecto, calidad metodológica y aplicabilidad. La complejidad surge de integrar y jerarquizar diferencias, no solo de resumir resultados.

7. Entre los aportes: organización del corpus, resúmenes conjuntos, identificación de patrones y discrepancias, apoyo para preguntas transversales. Entre las limitaciones: dependencia del material cargado, simplificación excesiva, riesgo de automatización acrítica y ausencia de evaluación metodológica autónoma.

8. Porque una revisión sistemática requiere métodos explícitos y reproducibles: estrategia de búsqueda, criterios de inclusión, evaluación del riesgo de sesgo y síntesis formal. NotebookLM puede ayudar a organizar documentos, pero no reemplaza ese proceso.

9. Podría cargarse un grupo de artículos previamente seleccionados, pedir un resumen comparativo inicial, identificar coincidencias y discrepancias, y luego discutir críticamente cada punto con el grupo revisando los textos originales.

10. Porque incluso una respuesta con citas puede simplificar, omitir matices o resaltar información sin ponderar adecuadamente su importancia metodológica. La cita ayuda a verificar, pero no reemplaza el análisis del lector.

https://doi.org/10.62486/978-9915-9928-3-9.Ch19

REFERENCIAS

Google. Learn about NotebookLM. NotebookLM Help. 2026. Disponible en la ayuda oficial de Google NotebookLM.
Google. Use chat in NotebookLM. NotebookLM Help. 2026. Disponible en la ayuda oficial de Google NotebookLM.
Google. Frequently asked questions – NotebookLM Help. NotebookLM Help. 2026. Disponible en la ayuda oficial de Google NotebookLM.
Clark J, Barton B, Albarqouni L, Byambasuren O, et al. Generative artificial intelligence use in evidence synthesis: A systematic review. Res Synth Methods. 2025. Available at: https://doi.org/10.1017/rsm.2025.16.
Lieberum JL, Toews M, Metzendorf MI, Heilmeyer F, Siemens W, Haverkamp C, et al. Large language models for conducting systematic reviews: On the rise, but not yet ready for use—a scoping review. J Clin Epidemiol. 2025;181:111746. Available t: https://doi.org/10.1016/j.jclinepi.2025.111746.
Harasgama S, Tunnicliffe DJ, Newman T, Campbell J, Haby MM, et al. Artificial intelligence tools for automating evidence synthesis: Scoping review. J Med Internet Res. 2026;28:e81597.
Sousa MSA, Peiris S, Figueiró MF, Haby MM, Baraldi AC, Reveiz L, Souza JP. The landscape of artificial intelligence tools and platforms for evidence synthesis: A scoping review. Syst Rev. 2026;15(1). Available at: https://doi.org/10.1186/s13643-025-02842-y.
Wang Z, Cao L, Danek B, Jin Q, Lu Z, Sun J. Accelerating clinical evidence synthesis with large language models. NPJ Digit Med. 2025;8(1):509. Available at: https://doi.org/10.1038/s41746-025-01840-7.
Pratte M, Thirukumar S, Zhang C, Slessarev M, Basmaji J, Prager R. Can large language models approximate the results of meta-analyses in critical care? A meta-research study. J Crit Care. 2025;92:155358. Available at: https://doi.org/10.1016/j.jcrc.2025.155358.
Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: A systematic review on large language models (LLMs). NPJ Digit Med. 2024;7:183. Available at: https://doi.org/10.1038/s41746-024-01157-x.
Tam TYC, Sivarajkumar S, Kapoor S, Stolyar AV, Polanska K, McCarthy KR, et al. A framework for human evaluation of large language models in healthcare derived from literature review. NPJ Digit Med. 2024;7(1):258. Available at: https://doi.org/10.1038/s41746-024-01258-7.
Li HY, Fu JF, Python A, et al. Implementing large language models in health care: Clinician-focused review with interactive guideline. J Med Internet Res. 2025;27:e71916. Available at: https://doi.org/10.2196/71916.
Ng KKY, Matsuba I, Zhang PC. RAG in health care: A novel framework for improving communication and decision-making by addressing LLM limitations. NEJM AI. 2025;2(1):AIra2400380.

Chapter 20 / Capítulo 20

Critical reading of scientific articles on health: A practical guide (Spanish Edition)
ISBN: 978-9915-9928-3-9
DOI: 10.62486/978-9915-9928-3-9.Ch20
Pages: 344-355

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Artificial Intelligence and Evidence-Based Clinical Decision-Making

Inteligencia Artificial y toma de decisiones clínicas basadas en evidencia

Mauricio Alejandro Paz del Rio
Jhossmar Cristians Auza-Santivañez
Blas Apaza-Huanca

1 Clínica Los Andes. La Paz, Bolivia.
2 Hospital del Gran Chaco "Fray Quebracho" de III nivel. Instituto Académico Científico Quispe-Cornejo. Tarija, Bolivia.
3 Ministerio de Salud y Deportes. Instituto Académico Científico Quispe-Cornejo. La Paz, Bolivia.

ABSTRACT

Introduction: Artificial intelligence (AI) is transforming clinical decision-making by facilitating access to, organization of, and synthesis of scientific evidence. Within the framework of Evidence-Based Medicine (EBM), its incorporation should strengthen, not replace, the integration of the best available evidence, clinical expertise, and patient values. Its use requires ongoing critical evaluation to ensure safe, contextualized, and ethically responsible decisions. Development: AI contributes to the analysis of diagnostic and therapeutic alternatives, the comparison of risks and benefits, communication with the patient, and the optimization of clinical reasoning. However, it has inherent limitations, such as hallucinations, algorithmic biases, outdated information, and uncritical automation. Therefore, any recommendation generated must be compared with reliable sources, current clinical guidelines, and the individual characteristics of the patient, also considering the availability of resources and the healthcare context. Conclusions: AI represents valuable cognitive support for evidence-based clinical practice, provided that its use is subordinate to clinical judgment and the ethical commitment of the professional. Its true potential lies not in replacing the ability to decide, but in enriching critical analysis and fostering shared, transparent, and patient-centered decisions. The quality of care will depend on the clinician's ability to integrate technology, scientific evidence, and humanism into a responsible decision-making process.

Keywords: Artificial Intelligence; Evidence-Based Medicine; Clinical Decision Making; Patient-Centered Care; Clinical Decision Support

RESUMEN

Introducción: La inteligencia artificial (IA) está transformando la toma de decisiones clínicas al facilitar el acceso, la organización y la síntesis de la evidencia científica. En el marco de la Medicina Basada en la Evidencia (MBE), su incorporación debe fortalecer, y no reemplazar, la integración entre la mejor evidencia disponible, la experiencia clínica y los valores del paciente. Su utilización exige una evaluación crítica permanente para garantizar decisiones seguras, contextualizadas y éticamente responsables. Desarrollo: La IA contribuye al análisis de alternativas diagnósticas y terapéuticas, la comparación de riesgos y beneficios, la comunicación con el paciente y la optimización del razonamiento clínico. Sin embargo, presenta limitaciones inherentes, como alucinaciones, sesgos algorítmicos, desactualización de la información y automatización acrítica. Por ello, toda recomendación generada debe contrastarse con fuentes confiables, guías clínicas vigentes y las características individuales del paciente, considerando además la disponibilidad de recursos y el contexto asistencial. Conclusiones: La IA representa un valioso apoyo cognitivo para la práctica clínica basada en evidencia, siempre que su utilización esté subordinada al juicio clínico y al compromiso ético del profesional. Su verdadero potencial no radica en sustituir la capacidad de decidir, sino en enriquecer el análisis crítico y favorecer decisiones compartidas, transparentes y centradas en el paciente. La calidad de la atención dependerá de la capacidad del clínico para integrar tecnología, evidencia científica y humanismo en un proceso de decisión responsable.

Palabras clave: Inteligencia Artificial; Medicina Basada en la Evidencia; Toma de Decisiones Clínicas; Atención Centrada en el Paciente; Apoyo a la Decisión Clínica

1. CASO CLÍNICO INICIAL

Un médico Internista evalúa a una paciente de 82 años con fibrilación auricular no valvular, hipertensión arterial, diabetes mellitus tipo 2, enfermedad renal crónica estadio 3 y antecedente de caída reciente. La paciente acude por palpitaciones y disnea leve. El electrocardiograma confirma fibrilación auricular con respuesta ventricular controlada. No presenta sangrado activo, pero refiere temor a iniciar anticoagulación porque una vecina "tuvo una hemorragia cerebral por tomar anticoagulantes".

El medico calcula un puntaje CHA_2DS_2-VASc elevado, lo que sugiere alto riesgo embólico, pero también identifica factores que incrementan el riesgo hemorrágico. Consulta guías clínicas recientes, revisa estudios sobre anticoagulación en adultos mayores y utiliza una herramienta de inteligencia artificial para organizar la información disponible: beneficios esperados, riesgos, alternativas terapéuticas, necesidad de ajuste por función renal, contraindicaciones y forma de explicar la decisión a la paciente.

La IA le presenta un resumen estructurado, pero el medico no acepta la recomendación de forma automática. Verifica las fuentes originales, contrasta la información con guías actualizadas y discute el caso con su tutor. Finalmente, conversa con la paciente y su familia, explica el riesgo de evento cerebrovascular, el riesgo de sangrado, las medidas de seguridad y las opciones disponibles. La decisión final no surge de la IA, sino de la integración entre evidencia científica, juicio clínico, contexto local, disponibilidad del medicamento, función renal, riesgo de caídas, seguimiento posible y preferencias de la paciente.

Este caso resume el propósito del capítulo: comprender cómo la inteligencia artificial puede apoyar la toma de decisiones clínicas basadas en evidencia, sin sustituir el razonamiento médico, la responsabilidad profesional ni la relación médico-paciente.

2. OBJETIVOS

Al finalizar este capítulo, el lector será capaz de:

- Comprender el papel de la inteligencia artificial como herramienta de apoyo en la toma de decisiones clínicas basadas en evidencia.
- Diferenciar entre información científica, recomendación automatizada, síntesis asistida por IA y juicio clínico.
- Identificar beneficios, riesgos y limitaciones del uso de IA en escenarios clínicos reales.
- Aplicar los principios de la Medicina Basada en la Evidencia para contextualizar recomendaciones diagnósticas y terapéuticas generadas o resumidas por IA.
- Reconocer la importancia de los valores, preferencias y circunstancias del paciente en toda decisión clínica.
- Utilizar IA de forma crítica para comparar alternativas diagnósticas y terapéuticas.
- Reconocer riesgos como alucinaciones, sesgo algorítmico, desactualización, falsa precisión y automatización acrítica.
- Adoptar una conducta ética, segura y responsable frente al uso de IA en salud.

3. INTRODUCCIÓN

La Medicina Basada en la Evidencia no termina cuando el médico aprende a leer un artículo científico. Leer críticamente un ensayo clínico, una revisión sistemática o una guía de práctica clínica es solo una parte del proceso. La verdadera competencia clínica aparece cuando el profesional debe responder una pregunta más compleja: ¿este resultado científico es aplicable a mi paciente concreto, en este momento, en este hospital, con estos recursos y con estas preferencias personales?

Durante décadas, la toma de decisiones clínicas se ha apoyado en la integración de tres componentes fundamentales: la mejor evidencia científica disponible, la experiencia clínica del profesional y los valores, expectativas y preferencias del paciente. La inteligencia artificial se incorpora a este escenario como una herramienta nueva, potente y desafiante.La IA puede resumir documentos, comparar alternativas, explicar conceptos complejos, organizar información clínica, generar listas de verificación, ayudar a preparar una conversación con el paciente y sugerir preguntas que el médico podría no haber considerado. Sin embargo, la IA no comprende el sufrimiento humano como lo hace un clínico, no asume responsabilidad legal, no conoce completamente el contexto institucional y puede generar información falsa con apariencia

https://doi.org/10.62486/978-9915-9928-3-9.Ch20

convincente.La Organización Mundial de la Salud ha señalado que los modelos multimodales grandes pueden recibir distintos tipos de datos y generar múltiples formas de salida, con potencial uso en atención sanitaria, investigación, salud pública y desarrollo de medicamentos. Esta expansión exige principios sólidos de ética, gobernanza y seguridad en salud.

En consecuencia, el médico moderno no solo necesita aprender a buscar y leer evidencia. También debe aprender a convivir con herramientas digitales capaces de sintetizar información a gran velocidad. El reto no es usar o no usar IA, sino aprender a usarla con criterio científico, prudencia ética y responsabilidad clínica.

4. DE LA LECTURA CRÍTICA A LA DECISIÓN CLÍNICA

La lectura crítica permite determinar si un estudio es válido, si sus resultados son importantes y si podrían aplicarse a la práctica clínica. Sin embargo, la decisión clínica exige un paso adicional: trasladar el conocimiento producido en una población de estudio hacia una persona concreta. Esta transición no es mecánica. Un ensayo clínico puede demostrar que un medicamento reduce eventos cardiovasculares, pero el médico debe valorar si el paciente tiene características similares a la población estudiada, si presenta contraindicaciones, si puede pagar el tratamiento, si existe disponibilidad local, si hay posibilidad de seguimiento y si el beneficio esperado justifica los riesgos.La inteligencia artificial puede apoyar este proceso al ordenar la información y hacer explícitos los elementos de análisis. Por ejemplo, ante una intervención terapéutica, puede ayudar a separar el beneficio relativo del beneficio absoluto, identificar desenlaces clínicamente relevantes, resumir efectos adversos, comparar alternativas y proponer una estructura para discutir el caso.No obstante, la IA puede equivocarse en la interpretación de un estudio, omitir limitaciones importantes o presentar conclusiones con una seguridad mayor a la que realmente permite la evidencia. Por ello, el paso de la lectura crítica a la decisión clínica debe conservar una secuencia racional:

Primero se formula la pregunta clínica. Luego se identifica la mejor evidencia disponible. Después se evalúa su validez y magnitud del efecto. Posteriormente se analiza la aplicabilidad al paciente. Finalmente se integra el contexto clínico, los recursos disponibles y las preferencias del paciente.

La IA puede acompañar este recorrido, pero no debe reemplazarlo.

5. ¿QUÉ PUEDE APORTAR LA IA EN LAS DECISIONES CLÍNICAS?

La IA puede cumplir múltiples funciones en la práctica clínica. Puede actuar como asistente de síntesis, ayudando a resumir guías, revisiones sistemáticas o artículos extensos. Puede funcionar como organizador del razonamiento, estructurando diagnósticos diferenciales, criterios de gravedad, indicaciones terapéuticas o preguntas pendientes. También puede apoyar la educación del paciente, transformando conceptos técnicos en lenguaje comprensible.

En consulta externa, puede ayudar al médico a preparar una explicación sencilla sobre el inicio de estatinas en prevención primaria. En hospitalización, puede ordenar los factores que apoyan o descartan una infección bacteriana. En emergencia, puede recordar criterios de gravedad, escalas clínicas o contraindicaciones de un tratamiento. En investigación, puede ayudar a comparar artículos y construir matrices de evidencia.

Sin embargo, la IA no debe ser entendida como un “oráculo clínico”. Su respuesta depende de los datos que recibe, del modelo utilizado, de su entrenamiento, de sus límites de actualización y de la forma en que se formula la pregunta. En salud, una respuesta aparentemente clara puede ser peligrosa si se basa en información incompleta, desactualizada o inventada.

Sittig y Singh advierten que las herramientas de IA están ingresando rápidamente a la práctica clínica, pero los mecanismos para garantizar su seguridad y efectividad todavía están en desarrollo. También resaltan la necesidad de que organizaciones y clínicos adopten medidas de seguridad para su implementación real.

https://doi.org/10.62486/978-9915-9928-3-9.Ch20

Figura 1: De la evidencia a la decisión clínica

Fuente: Elaboración propia, basada en los principios de la MBE

6. LO QUE LA IA NO DEBE HACER EN MEDICINA CLÍNICA

La IA no debe sustituir el juicio clínico del médico. No debe decidir por el paciente. No debe reemplazar la revisión de guías clínicas ni la lectura crítica de las fuentes originales. Tampoco debe utilizarse como argumento de autoridad para justificar una conducta médica.

Una respuesta generada por IA puede ser útil, pero no constituye por sí misma una indicación terapéutica. En medicina, una recomendación requiere comprender el diagnóstico, la gravedad, las comorbilidades, los riesgos, las preferencias del paciente, los recursos disponibles y la responsabilidad profesional.Por ejemplo, ante una paciente anciana con fibrilación auricular, la IA podría sugerir anticoagulación por alto riesgo embólico. Sin embargo, el médico debe revisar función renal, interacciones farmacológicas, riesgo hemorrágico, adherencia, caídas, acceso al medicamento, seguimiento posible y preferencias de la paciente. La recomendación general puede ser correcta, pero su aplicación debe ser individualizada.

El principio central es claro: la IA puede ayudar a pensar, pero no debe pensar por el médico.

Tabla 1. Diferencia entre usar IA como apoyo y usar IA como sustituto del juicio clínico

Uso adecuado de IA	Uso inadecuado de IA
Organizar información clínica	Delegar la decisión final
Resumir evidencia verificable	Aceptar respuestas sin fuentes
Comparar alternativas diagnósticas	Pedir diagnósticos definitivos sin contexto
Preparar explicación para el paciente	Reemplazar la conversación clínica
Identificar dudas o vacíos de información	Usarla como “oráculo” médico
Apoyar la decisión compartida	Ignorar preferencias del paciente
Revisar contraindicaciones posibles	Prescribir sin verificar guías
Adaptar evidencia al contexto	Copiar recomendaciones sin analizar disponibilidad

Fuente: Elaboración propia con Inteligencia artificial, recomendaciones de seguridad en IA clínica y guías de reporte DECIDE-AI/ TRIPOD+AI.

7. INTEGRACIÓN DE IA CON LOS PILARES DE LA MEDICINA BASADA EN LA EVIDENCIA

La IA debe integrarse dentro de la Medicina Basada en la Evidencia, no fuera de ella. Esto significa que su utilidad debe evaluarse a partir de los mismos principios que guían la práctica

https://doi.org/10.62486/978-9915-9928-3-9.Ch20

clínica responsable.

El primer pilar es la mejor evidencia disponible. La IA puede ayudar a localizar, resumir y comparar evidencia, pero el médico debe verificar si las fuentes son reales, actuales, relevantes y metodológicamente sólidas. No basta con que la respuesta "suene científica". Debe poder rastrearse hacia guías, revisiones sistemáticas, ensayos clínicos, estudios observacionales o consensos expertos según corresponda.

El segundo pilar es la experiencia clínica. La IA puede sugerir posibilidades, pero el médico interpreta signos, reconoce patrones, identifica gravedad, percibe cambios sutiles y comprende el contexto. La experiencia clínica permite saber cuándo una recomendación general no se adapta a un paciente específico.

El tercer pilar son los valores y preferencias del paciente. Ninguna herramienta tecnológica puede decidir por una persona qué riesgo está dispuesta a aceptar, qué desenlace considera más importante o qué tratamiento se ajusta mejor a su vida. La IA puede ayudar a explicar opciones, pero la decisión compartida sigue siendo profundamente humana.

Así, la IA debe concebirse como una herramienta de apoyo cognitivo. Puede mejorar la eficiencia del razonamiento, pero no sustituye el razonamiento. Puede ampliar la información disponible, pero no define por sí misma la decisión correcta. Puede facilitar la comunicación, pero no reemplaza la relación médico-paciente.

Tabla 2. Integración de IA con los pilares de la Medicina Basada en la Evidencia

Pilar de la MBE	Rol posible de la IA	Responsabilidad del médico
Mejor evidencia disponible	Buscar, resumir y comparar información	Verificar fuentes, calidad metodológica y actualidad
Experiencia clínica	Ordenar diagnósticos diferenciales y riesgos	Interpretar signos, gravedad, contexto y evolución
Valores del paciente	Preparar explicaciones comprensibles	Escuchar, deliberar y decidir de forma compartida
Contexto asistencial	Identificar recursos necesarios	Adaptar a disponibilidad, costo y seguimiento real

Fuente: Elaboración propia

8. USO DE IA PARA COMPARAR OPCIONES DIAGNÓSTICAS

Uno de los usos más útiles de la IA es ordenar alternativas diagnósticas. En lugar de preguntar "¿qué diagnóstico tiene este paciente?", el médico debería usarla para estructurar el análisis.Por ejemplo, en una paciente adulta mayor con fibrilación auricular, enfermedad renal crónica y antecedente de caída, la IA puede ayudar a ordenar los aspectos clave de la decisión: riesgo embólico, riesgo hemorrágico, función renal, ajuste de dosis, interacciones, disponibilidad del anticoagulante, seguimiento posible y preferencias de la paciente. Sin embargo, la decisión final debe surgir del análisis crítico del médico y de la conversación con la paciente y su familia. La utilidad no está en que la IA "decida" si el paciente tiene fibrilación auricular, sino en que ayude al médico a no olvidar variables importantes. El juicio clínico sigue siendo imprescindible para integrar signos vitales, contexto, evolución, antecedentes, probabilidad pretest y recursos disponibles.

En pruebas diagnósticas, la IA también puede ayudar a comparar sensibilidad, especificidad, valores predictivos, razones de verosimilitud y consecuencias clínicas de falsos positivos y falsos negativos. Este punto es especialmente relevante porque una prueba no debe interpretarse aislada, sino dentro de una probabilidad clínica previa.

9. USO DE IA PARA COMPARAR OPCIONES TERAPÉUTICAS

En terapéutica, la IA puede ordenar beneficios, daños, contraindicaciones, interacciones, costos, adherencia esperada y necesidad de seguimiento.

Un uso adecuado sería preguntar:

- "Organiza los factores que debo considerar para elegir anticoagulación en una paciente anciana con fibrilación auricular, enfermedad renal crónica y antecedente de caídas".
- Un uso menos adecuado sería preguntar:
- "¿Qué anticoagulante le doy?".

La diferencia es fundamental. La primera pregunta obliga a estructurar el razonamiento. La segunda delega una decisión clínica compleja a una herramienta que no asume responsabilidad,

https://doi.org/10.62486/978-9915-9928-3-9.Ch20

no conoce todo el contexto y puede estar equivocada.

10. TOMA DE DECISIONES COMPARTIDA ASISTIDA POR IA

La toma de decisiones compartida es especialmente importante cuando existen varias opciones razonables, cuando los beneficios y riesgos están equilibrados o cuando las preferencias del paciente modifican la elección final.La IA puede apoyar este proceso al traducir información compleja a lenguaje sencillo, elaborar comparaciones comprensibles y ayudar al médico a preparar explicaciones sobre riesgo absoluto, beneficio esperado y alternativas.Por ejemplo, no es lo mismo decir a un paciente que un tratamiento "reduce el riesgo relativo en 25%" que explicarle cuántos eventos se evitarían por cada 100 o 1000 pacientes tratados durante un periodo determinado. La IA puede ayudar a convertir conceptos estadísticos en mensajes claros, siempre que el médico verifique que los números sean correctos y que la explicación no simplifique en exceso.

En este sentido, la IA puede mejorar la comunicación clínica, pero no debe desplazar el diálogo. La decisión compartida implica escuchar temores, explorar expectativas, comprender barreras económicas, respetar valores culturales y reconocer la autonomía del paciente.

Una herramienta digital puede preparar el terreno, pero la conversación terapéutica sigue perteneciendo al encuentro clínico.

11. APLICABILIDAD CLÍNICA EN LATINOAMÉRICA Y BOLIVIA

Este punto es fundamental para Latinoamérica y Bolivia. Muchas guías internacionales se desarrollan en contextos con alta disponibilidad tecnológica, acceso amplio a medicamentos, sistemas de seguimiento estables y recursos diagnósticos oportunos.Sin embargo, el médico latinoamericano con frecuencia trabaja en escenarios donde los estudios complementarios no están disponibles, los pacientes pagan tratamientos de bolsillo, los medicamentos recomendados no existen en farmacia, el seguimiento es irregular o las decisiones deben tomarse en hospitales con recursos limitados.

La IA puede resumir guías de alto nivel, pero no siempre reconoce las restricciones del contexto local. Puede sugerir un medicamento de primera línea que no está disponible, una prueba diagnóstica inaccesible o una estrategia de seguimiento inviable.

Por ello, la aplicabilidad debe analizarse con preguntas concretas:

- ¿Esta intervención existe en mi medio?
- ¿El paciente puede acceder a ella?
- ¿Hay personal entrenado?
- ¿El sistema puede sostener el seguimiento?
- ¿El beneficio esperado justifica el costo?
- ¿Hay una alternativa razonable más disponible?

La evidencia científica orienta la decisión, pero la realidad asistencial la condiciona. Un médico formado en MBE no copia recomendaciones; las interpreta, adapta y aplica con responsabilidad.

12. RIESGOS, SESGOS Y LIMITACIONES DE LA IA CLÍNICA

El uso de IA en decisiones clínicas implica riesgos específicos.El primero es la alucinación: generación de información falsa, referencias inexistentes o afirmaciones incorrectas presentadas con lenguaje convincente. Este riesgo es particularmente peligroso en medicina porque el estilo seguro de una respuesta puede producir falsa confianza.El segundo riesgo es la desactualización. Una recomendación puede estar basada en información previa a una nueva guía, alerta farmacológica o estudio relevante.El tercero es el sesgo de entrenamiento. Si los datos usados para entrenar un sistema no representan adecuadamente a determinadas poblaciones, la recomendación puede ser menos precisa o incluso injusta para grupos subrepresentados. El cuarto riesgo es la automatización acrítica. El médico puede aceptar una respuesta porque proviene de una herramienta sofisticada, aunque no haya verificado su fundamento.El quinto es la falsa precisión. La IA puede presentar una recomendación como si fuera exacta cuando en realidad la evidencia es débil, indirecta o incierta.El sexto es el riesgo médico-legal, especialmente si se introducen datos identificables en herramientas no institucionales, si no se documenta adecuadamente el razonamiento o si se actúa siguiendo una recomendación no verificada.

https://doi.org/10.62486/978-9915-9928-3-9.Ch20

DECIDE-AI fue desarrollada precisamente para mejorar el reporte de evaluaciones clínicas tempranas de sistemas de apoyo a la decisión basados en IA, enfatizando la seguridad, los factores humanos y la evaluación clínica antes de una implementación amplia.

Además, TRIPOD+AI actualizó las recomendaciones para reportar modelos de predicción clínica que usan regresión o machine learning, con el objetivo de mejorar transparencia y reproducibilidad.

Tabla 3. Riesgos principales del uso de IA en decisiones clínicas

Riesgo	Definición	Ejemplo clínico	Prevención
Alucinación	Información falsa presentada como verdadera	Citar una guía inexistente	Verificar fuentes originales
Desactualización	Uso de información no vigente	Recomendar tratamiento superado por nueva guía	Revisar fecha y guías recientes
Sesgo algorítmico	Recomendación menos válida para ciertos grupos	Modelo entrenado en población distinta	Evaluar representatividad
Automatización acrítica	Aceptar la respuesta sin cuestionarla	Copiar una conducta terapéutica	Mantener juicio clínico
Falsa precisión	Presentar certeza donde hay incertidumbre	"El riesgo exacto es..." sin base sólida	Revisar intervalos, calidad y aplicabilidad
Riesgo de privacidad	Exposición de datos sensibles	Ingresar nombre, historia o datos identificables	Anonimizar información

Fuente: Elaboración propia

13. RECOMENDACIONES PARA EL USO RESPONSABLE DE IA EN DECISIONES CLÍNICAS

El uso responsable de IA en medicina debe partir de una premisa sencilla: la herramienta puede asistir, pero no prescribe; puede sugerir, pero no decide; puede organizar evidencia, pero no reemplaza la responsabilidad clínica.Antes de usar una respuesta generada por IA, el médico debe verificar las fuentes, contrastar con guías vigentes, revisar contraindicaciones, evaluar interacciones, considerar características individuales del paciente y adaptar la recomendación al contexto local.

No deben introducirse datos personales identificables del paciente en sistemas no institucionales o sin garantías de privacidad. Además, cuando la IA influye en el razonamiento clínico, el médico debe documentar la lógica final de la decisión, no simplemente escribir que "la IA recomendó" una conducta.

Una conducta segura consiste en usar la IA para ampliar el análisis, no para cerrarlo. Es útil pedirle que identifique dudas, que compare alternativas, que organice riesgos y beneficios, que prepare explicaciones para pacientes o que ayude a revisar si falta algún elemento importante. Pero la decisión final debe ser validada por el médico responsable.La FDA y otros organismos reguladores han resaltado la importancia de transparencia, información clara para usuarios y evaluación del desempeño del equipo humano-IA en dispositivos médicos con aprendizaje automático.

14. CHECKLIST PRÁCTICO PARA USAR IA EN TOMA DE DECISIONES CLÍNICAS

Tabla 4. Checklist para el uso responsable de inteligencia artificial en la toma de decisiones clínicas.

Pregunta	Sí	No	Observación
¿Formulé una pregunta clínica específica antes de usar IA?			
¿La pregunta incluye paciente, problema, intervención, comparación o desenlace cuando corresponde?			
¿La IA entregó fuentes verificables?			
¿Revisé si las fuentes son reales, actuales y pertinentes?			
¿Contrasté la respuesta con guías clínicas o revisiones confiables?			
¿Evalué si la recomendación aplica a mi paciente concreto?			
¿Consideré comorbilidades, función renal, alergias, interacciones y contraindicaciones?			
¿Analicé disponibilidad, costo y acceso en mi contexto local?			
¿Identifiqué beneficios absolutos y riesgos relevantes?			

https://doi.org/10.62486/978-9915-9928-3-9.Ch20

Pregunta	Sí	No	Observación
¿Exploré preferencias, temores y expectativas del paciente?			
¿Evité ingresar datos personales identificables?			
¿Documenté mi razonamiento clínico final y no solo la respuesta de la IA?			
¿La decisión final fue tomada por el médico responsable?			

Fuente: Elaboración propia.

Interpretación práctica: si una respuesta de IA no puede verificarse, no debe utilizarse como fundamento principal de una decisión clínica. Si la recomendación parece útil pero no se adapta al paciente, debe modificarse o descartarse. Si el médico no puede explicar la decisión sin mencionar la IA como autoridad principal, el razonamiento clínico aún no está completo.

16. UTILIDAD CLÍNICA, ACADÉMICA Y DOCENTE

En la práctica clínica, la IA puede ser útil en consulta externa, emergencia, hospitalización y unidades críticas, especialmente para ordenar información, preparar discusiones clínicas y mejorar la eficiencia del razonamiento.

En docencia, puede ayudar a crear casos clínicos, preguntas de autoevaluación, resúmenes de guías y ejercicios de interpretación de evidencia. Para internos y residentes, el mayor valor de la IA no está en obtener respuestas rápidas, sino en aprender a formular mejores preguntas.

Un médico que pregunta mejor, razona mejor. La IA puede funcionar como un tutor auxiliar que obliga a estructurar el pensamiento: población, intervención, comparación, desenlace, riesgos, contexto y preferencias.

En investigación, puede apoyar la búsqueda inicial de literatura, la comparación de estudios y la redacción preliminar. Sin embargo, todo producto académico generado con IA debe ser revisado, corregido y verificado por el autor humano. La IA no debe inventar referencias, no debe sustituir lectura crítica y no debe reemplazar la responsabilidad intelectual del investigador.

Figura 3: Medicina Basada en la Evidencia en la era de la inteligencia artificial

https://doi.org/10.62486/978-9915-9928-3-9.Ch20

17. PREGUNTAS DE AUTOEVALUACIÓN

PREGUNTAS DE SELECCIÓN MÚLTIPLE

1. ¿Cuál es el papel más adecuado de la IA en la toma de decisiones clínicas?
a) Sustituir el juicio clínico del médico.
b) Emitir órdenes terapéuticas definitivas.
c) Apoyar la organización, síntesis y contextualización de la evidencia.
d) Reemplazar la conversación con el paciente.
2. ¿Cuál de los siguientes representa un riesgo importante de la IA generativa en medicina?
a) Que siempre solicita demasiados estudios complementarios.
b) Que puede generar información falsa con apariencia convincente.
c) Que solo funciona en investigación básica.
d) Que elimina por completo los sesgos clínicos.
3. En Medicina Basada en la Evidencia, una decisión clínica debe integrar:
a) Solo guías internacionales.
b) Solo experiencia del médico.
c) Evidencia científica, experiencia clínica y valores del paciente.
d) Únicamente preferencias del paciente.
4. ¿Qué debe hacer el médico antes de aplicar una recomendación generada por IA?
a) Copiarla directamente a la historia clínica.
b) Aplicarla si está redactada con lenguaje técnico.
c) Verificar fuentes, actualizar evidencia y contextualizarla al paciente.
d) Aceptarla si coincide con la primera búsqueda en internet.
5. En contextos latinoamericanos, la aplicabilidad de una recomendación depende especialmente de:
a) Que provenga de una revista internacional.
b) Que sea tecnológicamente avanzada.
c) Disponibilidad, costo, acceso, seguimiento y contexto asistencial.
d) Que la IA la repita varias veces.

PREGUNTAS ABIERTAS

6. Explique por qué la IA no debe reemplazar el juicio clínico en decisiones médicas.
7. Mencione tres formas en las que la IA puede apoyar la toma de decisiones compartida.
8. ¿Qué significa contextualizar la evidencia en un hospital con recursos limitados?
9. ¿Qué es una alucinación en IA y por qué es peligrosa en medicina?
10. Proponga una forma responsable de usar IA ante una duda terapéutica.

RESPUESTAS SUGERIDAS

1. c) La IA debe utilizarse como herramienta de apoyo para organizar información, resumir evidencia y estructurar el razonamiento, pero la decisión final corresponde al médico.
2. b) Las alucinaciones son uno de los riesgos más importantes de los modelos generativos, porque pueden producir afirmaciones incorrectas difíciles de detectar.
3. c) La MBE se basa en la integración de evidencia, experiencia clínica y valores/preferencias del paciente.
4. c) Toda recomendación asistida por IA debe verificarse, contrastarse con evidencia vigente y adaptarse al caso clínico.
5. c) Una recomendación puede ser científicamente válida pero poco aplicable si no existe acceso real a la intervención o al seguimiento requerido.
6. La IA no reemplaza el juicio clínico porque no conoce completamente el contexto del paciente, puede generar errores, no asume responsabilidad profesional y no integra de forma humana los valores, temores y preferencias del paciente.
7. Puede traducir evidencia a lenguaje sencillo, comparar riesgos y beneficios de distintas alternativas, y preparar explicaciones para que el paciente participe de forma informada en la decisión.
8. Significa evaluar si la intervención recomendada está disponible, si el paciente puede acceder a ella, si hay seguimiento posible, si existen alternativas locales y si el beneficio esperado

https://doi.org/10.62486/978-9915-9928-3-9.Ch20

justifica el costo o los riesgos.

9. Es una respuesta falsa o inventada presentada como si fuera verdadera. Es peligrosa porque puede inducir diagnósticos, tratamientos o decisiones incorrectas.

10. Una forma responsable sería formular una pregunta clínica específica, pedir comparación de opciones, verificar las fuentes originales, contrastar con guías vigentes, revisar contraindicaciones y tomar la decisión final considerando al paciente y el contexto.

https://doi.org/10.62486/978-9915-9928-3-9.Ch20

REFERENCIAS

World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Geneva: World Health Organization; 2024. Available at: https://www.who.int/publications/i/item/9789240029200.
Sittig DF, Singh H. Recommendations to ensure safety of AI in real-world clinical care. JAMA. 2025;333(6):457–8. Availableat: https://doi.org/10.1001/jama.2024.24598.
Funes Galindo, I. S., Echazú Torres, C. D., Rivero Castedo, L., Gómez Triana, J. E., Aguilar Medrano, F., Vargas-Jimenez, J. D., Apaza-Huanca, B., Auza-Santivañez, J. C., Bautista-Vanegas, F. E., & Cruz Nuñez, Y. (2026). Artificial intelligence in dentistry: clinical applications, ethical challenges and future perspectives. SAP Artificial Intelligence in Dentistry, 2, 51. https://doi.org/10.62486/aid202651
Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ. 2022;377:e070904. Availableat: https://doi.org/10.1136/bmj-2022-070904.
Collins GS, Dhiman P, Ma J, Schlussel MM, Archer L, Van Calster B, et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. Available at: https://doi.org/10.1136/bmj-2024-078378.
Ibrahim H, Liu X, Zariffa N, Morris AD, Denniston AK. Reporting guidelines for clinical trials of artificial intelligence interventions: The SPIRIT-AI and CONSORT-AI guidelines. Lancet Digit Health. 2021;3(1):e1–2.
Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. Nat Med. 2020;26(9):1364–74. Available at: https://doi.org/10.1038/s41591-020-1034-x.
Rivera SC, Liu X, Chan A-W, Denniston AK, Calvert MJ; SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. Nat Med. 2020;26(9):1351–63. Available at: https://doi.org/10.1038/s41591-020-1037-7.
Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med. 2021;385(3):283–6. Availableat: https://doi.org/10.1056/NEJMc2104626.
Carías A, Auza-Santivañez JC, Carías P, Díaz Bonilla V, Bautista-Vanegas FE, Vallejos-Rejas DRE, et al. Artificial intelligence in the Otorhinolaryngology class. Seminars in Medical Writing and Education [Internet]. 2025 Feb. 9 [cited 2026 Jul. 11];4:160. Available from: https://mw.ageditor.ar/index.php/mw/article/view/160
Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: Benefits, risks, and strategies for success. NPJ Digit Med. 2020;3:17. Available at: https://doi.org/10.1038/s41746-020-0221-y.
Feng J, Phillips RV, Malenica I, Bishara A, Hubbard AE, Celi LA, et al. Clinical artificial intelligence quality improvement: Towards continual monitoring and updating of AI algorithms in health care. NPJ Digit Med. 2022;5(1):66. Available at: https://doi.org/10.1038/s41746-022-00611-y.
Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ. 2020;368:l6927. Available at: https://doi.org/10.1136/bmj.l6927.
US Food and Drug Administration. Artificial intelligence in software as a medical device. Silver Spring (MD): US Food and DrugAdministration; 2025.
Auza-Santivañez JC, Sosa Remón A, Bautista-Vanegas FE, Cabezas-Soliz IN, Vargas Gallego I, Apaza-Huanca B, et al. Artificial Intelligence in the Intensive Care Unit: Present and Future. Seminars in Medical Writing and Education [Internet]. 2025 Mar. 18 [cited 2026 Jul. 11];4:464. Available from: https://mw.ageditor.ar/index.php/mw/article/view/464
US Food and Drug Administration, Health Canada, Medicines and Healthcare products Regulatory Agency. Transparency for machine learning-enabled medical devices: guiding principles. Silver Spring (MD): US Food and Drug Administration; 2024.

Chapter 21 / Capítulo 21

Critical reading of scientific articles on health: A practical guide (Spanish Edition)
ISBN: 978-9915-9928-3-9
DOI: 10.62486/978-9915-9928-3-9.Ch21
Pages: 356-375

© 2026 The authors. This is an open access book chapter distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Ethics and Limitations of Artificial Intelligence in Medicine

Ética y limitaciones de la inteligencia artificial en medicina

Weymar Lehi Poma Luna
Marizol Garcia-Leon
Ismael Vargas Gallego

1 Hospital Municipal Boliviano Holandes. La Paz, Bolivia.
2 Universidad Mayor Real y Pontificia de San Francisco Xavier. Chuquisaca, Bolivia.
3 Hospital de Tercer Nivel Dr. Hernán Messuti Ribera. Pando, Bolivia.

ABSTRACT

Introduction: Artificial intelligence (AI) has transformed medical practice by strengthening diagnosis, decision-making, and clinical management; however, its incorporation raises ethical, methodological, and regulatory challenges. From the perspective of Evidence-Based Medicine (EBM), the implementation of these technologies requires a rigorous evaluation of their validity, safety, and applicability in diverse populations. Development: This chapter examines bioethical principles, algorithmic biases, the limitations of large language models, and the risks of automation bias, highlighting the insufficient representation of Andean populations in the datasets used to train AI systems. It also emphasizes the need for local validation, transparency, data governance, and human oversight to ensure safe and equitable clinical decisions. Conclusions: AI is a support tool with great potential, but its responsible integration requires solid scientific evidence, contextual adaptation, and a continuous commitment to ethical reflection. Only through critical, regulated, and patient-centered implementation will it be possible to harness its benefits without compromising equity, autonomy, or clinical safety.

Keywords: Artificial Intelligence; Evidence-Based Medicine; Bioethics; Clinical Decision Support Systems; Health Equity

RESUMEN

Introducción: La inteligencia artificial (IA) ha transformado la práctica médica al fortalecer el diagnóstico, la toma de decisiones y la gestión clínica; sin embargo, su incorporación plantea desafíos éticos, metodológicos y regulatorios. Desde la perspectiva de la Medicina Basada en la Evidencia (MBE), la implementación de estas tecnologías exige una evaluación rigurosa de su validez, seguridad y aplicabilidad en poblaciones diversas. Desarrollo: El capítulo examina los principios bioéticos, los sesgos algorítmicos, las limitaciones de los grandes modelos de lenguaje y los riesgos del automation bias, destacando la insuficiente representación de poblaciones andinas en los conjuntos de datos utilizados para entrenar sistemas de IA. Asimismo, enfatiza la necesidad de validación local, transparencia, gobernanza de datos y supervisión humana para garantizar decisiones clínicas seguras y equitativas. Conclusiones: La IA constituye una herramienta de apoyo con gran potencial, pero su integración responsable requiere evidencia científica sólida, adaptación contextual y un compromiso permanente con la reflexión ética. Solo mediante una implementación crítica, regulada y centrada en el paciente será posible aprovechar sus beneficios sin comprometer la equidad, la autonomía ni la seguridad clínica.

Palabras clave: Artificial Intelligence; Evidence-Based Medicine; Bioethics; Clinical Decision Support Systems; Health Equity

1. CASO CLÍNICO

En el servicio de Emergencias de un Hospital de Tercer Nivel en La Paz, Bolivia (3,640 msnm.), Don Teodoro, de 58 años, de etnia aimara, agricultor y residente habitual del altiplano (Achacachi, 3,850 msnm.), ingresa por un cuadro de 48 horas de evolución caracterizado por disuria, dolor lumbar izquierdo, picos febriles no cuantificados y tendencia a la somnolencia. Como antecedente de relevancia, padece de eritrocitosis fisiológica adaptativa de la altura (Hemoglobina basal conocida: 18.2 g/dL).

Evaluación e Intervención Algorítmica: El hospital cuenta con un Sistema de Apoyo a la Decisión Clínica (SADC) basado en un modelo de aprendizaje profundo entrenado con bases de datos norteamericanas (MIMIC-IV). El sistema procesa los signos vitales de Don Teodoro en tiempo real:

- Presión Arterial: 100/60 mmHg (PAM: 73 mmHg)
- Frecuencia Cardíaca: 104 lpm
- Frecuencia Respiratoria: 24 rpm
- Temperatura: 38.7 °C
- Saturación Periférica de Oxígeno (SpO2): 88% (aire ambiente)
- Laboratorio: Leucocitos 14,000/uL; Creatinina 1.4 mg/dL; Hemoglobina 18.5 g/dL.

El algoritmo dispara de forma inmediata una **Alerta Roja de Sepsis Grave** con una predicción de mortalidad del 42%. El sistema justifica la alerta penalizando severamente la subescala de oxigenación del score SOFA automatizado, interpretando la SpO2 de 88% como una "insuficiencia respiratoria aguda catastrófica" y la hemoglobina de 18.5 g/dL como un signo de hemoconcentración severa por shock hipovolémico.

El Conflicto Ético-Clínico: El médico residente, influenciado por el **sesgo de automatización (*automationbias*)**, decide iniciar una reanimación hídrica agresiva con cristaloides a razón de 30 mL/kg y prepara la secuencia de intubación rápida para conectar al paciente a ventilación mecánica invasiva.

Sin embargo, el médico internista de pase de guardia detiene el procedimiento mediante una **auditoría cognitiva contextual**:

- **Sesgo de Medición por Altitud:** Una SpO2 de 88% es un parámetro fisiológico normal y basal para un habitante de Achacachi a casi 4,000 metros de altura. El algoritmo carece de un modelo de calibración barométrica.
- **Iatrogenia Potencial:** La infusión masiva de fluidos en un paciente con vasculatura pulmonar adaptada a la altura (caracterizada por vasoconstricción hipóxica refleja y discreta hipertensión pulmonar media basal) habría desencadenado un edema agudo de pulmón no cardiogénico e insuficiencia ventricular derecha aguda.

Desenlace: El juicio clínico humano anuló la recomendación de la "caja negra". El paciente fue tratado exitosamente para una urosepsis mediante fluidoterapia guiada por metas locales y antibioterapia dirigida, manteniendo su SpO2 en su rango meta personalizado de altitud (88-91%). Sobrevivió sin secuelas.

2. OBJETIVOS

Al finalizar el estudio de este capítulo, el lector será capaz de:

- Identificar y clasificar los tipos de sesgos algorítmicos (selección, medición y confusión) en sistemas de apoyo a la decisión clínica (SADC).
- Evaluar críticamente la validez externa de un modelo predictivo global al ser aplicado a poblaciones con adaptaciones biogeográficas extremas (hipoxia hipobárica crónica).
- Reconocer y mitigar el automationbias y el fenómeno de deskilling en la práctica clínica diaria.
- Aplicar los principios de la gobernanza de datos y la bioética principialista para auditar la seguridad de las herramientas de IA antes de su implementación institucional.

3. INTRODUCCIÓN

La integración de la inteligencia artificial (IA) en la práctica médica constituye uno de los cambios más profundos que ha experimentado la medicina desde la introducción de la aleatorización en los ensayos clínicos a mediados del siglo XX. En el transcurso de apenas una década, y con especial aceleración a partir de 2022, los modelos de aprendizaje profundo y los grandes modelos de lenguaje (LLMs) han pasado de ser objetos de interés académico a

https://doi.org/10.62486/978-9915-9928-3-9.Ch21

herramientas activamente empleadas en diagnóstico por imagen, apoyo a la decisión clínica, educación médica, redacción de informes y comunicación con pacientes. Esta velocidad de adopción es, simultáneamente, su mayor virtud y su riesgo más significativo.

A diferencia de los fármacos o los procedimientos quirúrgicos, cuya incorporación a la práctica clínica exige décadas de ensayos controlados, evaluación de seguridad y aprobación regulatoria, las herramientas de IA se han desplegado a escala global sin que exista, en la mayoría de los sistemas de salud, un marco de evaluación previa equivalente. Esta asimetría entre el ritmo de innovación tecnológica y la producción de evidencia científica representa un desafío estructural para la medicina basada en evidencia (MBE).

El problema se amplifica cuando se examina la distribución geográfica de los datos de entrenamiento de los modelos. La gran mayoría de los algoritmos de IA en salud han sido desarrollados y validados en poblaciones occidentales, urbanas, de predominio europeo o norteamericano, y residentes a baja altitud. Las consecuencias de esta concentración demográfica no son meramente estadísticas: los errores de generalización externa tienen consecuencias clínicas directas en poblaciones subrepresentadas, como las comunidades andinas de Bolivia, Perú, Ecuador y Chile, que habitan a altitudes de entre 2.500 y más de 4.000 metros sobre el nivel del mar (m.s.n.m.) y presentan adaptaciones fisiológicas profundas que ningún modelo entrenado en poblaciones de baja altitud puede capturar adecuadamente.

Este capítulo adopta una perspectiva crítica, informada por la MBE y por la fisiología clínica, para examinar los beneficios reales, las limitaciones metodológicas y los desafíos éticos del uso de IA en medicina.

Se desarrolla en cuatro ejes conceptuales complementarios:

- fundamentos bioéticos y regulatorios
- limitaciones epistemológicas y cognitivas de los LLMs
- sesgos algorítmicos en poblaciones de altura y grupos subrepresentados
- principios para una integración ética, segura y contextualmente adecuada de la IA en la práctica clínica y la educación médica.

4. FUNDAMENTOS ÉTICOS DE LA INTELIGENCIA ARTIFICIAL EN MEDICINA

La reflexión ética sobre el uso de tecnología en medicina no es nueva. Sin embargo, la IA plantea dilemas cualitativamente distintos a los que la bioética clásica había contemplado, en razón de su opacidad algorítmica, su capacidad para operar a escala sin supervisión humana continua y su potencial de amplificar desigualdades preexistentes a velocidades y magnitudes sin precedente.

4.1 EL MARCO PRINCIPIALISTA Y SUS LÍMITES EN EL CONTEXTO DE LA IA

Los cuatro principios del marco de Georgetown; autonomía, beneficencia, no maleficencia y justicia, siguen siendo el punto de partida ineludible de cualquier análisis bioético en medicina. Sin embargo, aplicados a la IA, cada uno de ellos adquiere matices que exigen ser examinados con precisión conceptual.

AUTONOMÍA

El principio de autonomía reconoce el derecho del paciente a participar informadamente en las decisiones relativas a su salud. En el contexto de la IA, este principio se ve tensionado por al menos tres mecanismos.

- En primer lugar, la opacidad de los modelos de caja negra en particular las redes neuronales profundas impiden que el clínico, y con mayor razón el paciente, comprenda los fundamentos de una recomendación algorítmica.
- En segundo lugar, el consentimiento informado clásico no está diseñado para capturar la complejidad de los procesos de decisión mediados por IA.
- En tercer lugar, la progresiva delegación de decisiones clínicas a sistemas automatizados desplaza el sitio de toma de decisiones fuera del ámbito de la deliberación médico-paciente.

La UNESCO, en su Recomendación sobre la Ética de la Inteligencia Artificial adoptada en 2021 —el primer instrumento normativo global de aplicación universal en este campo— subraya

https://doi.org/10.62486/978-9915-9928-3-9.Ch21

que las tecnologías de IA deben reforzar, no erosionar, la capacidad de agencia individual. Esto implica que cualquier sistema de IA aplicado en salud debe ser explicable, auditado y susceptible de ser cuestionado por el clínico y, en última instancia, por el paciente.

BENEFICENCIA Y NO MALEFICENCIA

La dicotomía beneficencia/no maleficencia adquiere una nueva dimensión cuando el agente que actúa no es un médico sino un sistema algorítmico. Los beneficios potenciales de la IA en medicina cuentan con mayor precisión diagnóstica en imagen médica, detección precoz de sepsis, optimización de triaje en urgencias, acceso a atención en regiones remotas están documentados en la literatura científica de alto impacto y son genuinamente significativos.

Sin embargo, estos beneficios no son universales ni homogéneos. La evidencia disponible indica que los modelos de IA presentan un rendimiento sistemáticamente inferior en grupos raciales o étnicos subrepresentados en los datos de entrenamiento, en pacientes con multimorbilidad compleja y en contextos clínicos diferentes de aquellos en que fueron validados.

El principio de no maleficencia exige, en consecuencia, que la implementación de sistemas de IA en medicina esté precedida de una evaluación rigurosa de sus potenciales perjuicios diferenciados, con especial atención a las poblaciones en mayor riesgo de daño por generalización incorrecta.

JUSTICIA DISTRIBUTIVA Y EQUIDAD ALGORÍTMICA

La justicia distributiva, en su formulación más exigente, requiere que los beneficios y cargas derivados del uso de IA en salud sean distribuidos equitativamente, sin discriminación por raza, género, etnia, nivel socioeconómico o ubicación geográfica. Sin embargo, la evidencia empírica disponible apunta en la dirección opuesta. Una revisión sistemática publicada en JAMA Network Open en 2024 identificó que los modelos de IA en salud derivados de registros electrónicos de datos clínicos presentan sesgos de selección, medición y temporal que sistemáticamente afectan de manera desproporcionada a minorías raciales, poblaciones de bajos ingresos y pacientes con acceso limitado al sistema de salud formal.

En el contexto latinoamericano, esta inequidad algorítmica adopta una dimensión adicional: la escasa participación de instituciones de salud de la región en la generación de datos para el entrenamiento de modelos globales implica que los sistemas de IA no son solo técnicamente subóptimos para estas poblaciones, sino que, en algunos casos, pueden ser activamente dañinos al aplicar umbrales, valores de referencia y criterios de decisión calibrados para poblaciones fisiológicamente distintas.

4.2 EXPLICABILIDAD, TRANSPARENCIA Y RESPONSABILIDAD

La explicabilidad algorítmica es la capacidad de un sistema de IA para justificar sus decisiones en términos comprensibles para un ser humano, ha emergido como uno de los principios centrales de la ética de la IA en salud. La Unión Europea en marzo de 2024, ha establecido que los sistemas de IA clasificados como de alto riesgo entre los cuales se incluyen explícitamente los sistemas de apoyo a la decisión clínica, el diagnóstico médico automatizado y los sistemas de gestión de pacientesdeben ser transparentes, explicables y sujetos a supervisión humana continua.

La responsabilidad legal ante errores generados por sistemas de IA constituye una zona de alta ambigüedad jurídica en la mayoría de los sistemas legales vigentes. ¿Recae la responsabilidad sobre el desarrollador del modelo, sobre la institución que lo implementa o sobre el clínico que actúa siguiendo su recomendación? Esta pregunta permanece abierta en la mayoría de las jurisdicciones y constituye un desincentivo estructural para la adopción responsable de IA en entornos de alta responsabilidad clínica.

4.3 PRIVACIDAD, SOBERANÍA DE DATOS Y CONFLICTOS DE INTERÉS

El entrenamiento de modelos de IA en salud requiere volúmenes masivos de datos clínicos de alta sensibilidad. La captación, almacenamiento y procesamiento de estos datos por parte de corporaciones tecnológicas privadas, muchas de ellas domiciliadas fuera de las jurisdicciones

https://doi.org/10.62486/978-9915-9928-3-9.Ch21

cuyos datos utilizan, plantea desafíos profundos de soberanía de datos sanitarios, privacidad individual y gobernanza institucional.

El conflicto de interés inherente al modelo de negocio de las grandes corporaciones tecnológicas que desarrollan herramientas de IA para salud cuyos incentivos económicos pueden divergir significativamente de los objetivos de salud pública, constituye una dimensión ética frecuentemente subestimada en la literatura técnica. La industria tecnológica ha invertido sumas extraordinarias en publicar estudios sobre el rendimiento de sus modelos en condiciones seleccionadas, con escaso interés en documentar sus limitaciones, sus fallas o su comportamiento en poblaciones subrepresentadas.

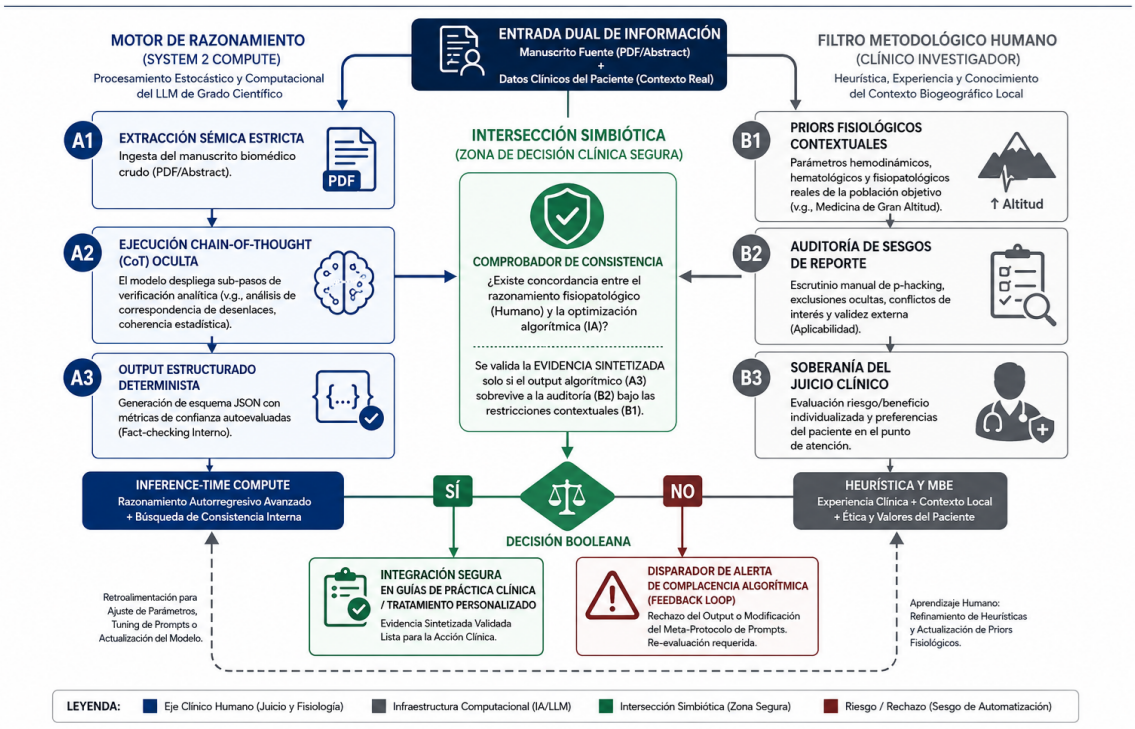
5. INTELIGENCIA ARTIFICIAL Y MEDICINA BASADA EN EVIDENCIA

La medicina basada en evidencia, tal como la formularon Sackett y colaboradores a principios de la década de 1990, descansa sobre tres pilares: la mejor evidencia científica disponible, la experiencia clínica del médico y los valores y preferencias del paciente. La irrupción de la IA en la práctica clínica interpela los tres pilares de manera simultánea y, en ocasiones, contradictoria.

5.1 JERARQUÍA DE LA EVIDENCIA EN IA CLÍNICA

Desde la perspectiva de la MBE, la evidencia sobre eficacia y seguridad de los sistemas de IA en salud debe ser evaluada con los mismos estándares que se aplican a cualquier intervención terapéutica o diagnóstica. Sin embargo, la gran mayoría de los estudios sobre IA en medicina son estudios observacionales retrospectivos, con escasos ensayos clínicos aleatorizados (ECAs) de buena calidad metodológica. Esta debilidad estructural de la evidencia disponible contrasta con la narrativa de precisión y superioridad que frecuentemente acompaña a la presentación pública de nuevos modelos de IA en salud.Una revisión sistemática y metaanálisis publicada en BMJ en 2022 (Liu et al.) analizó 62 modelos de IA para diagnóstico médico por imagen y encontró que la calidad metodológica de los estudios era en general baja o muy baja, con escasa reproducibilidad, ausencia de validación externa rigurosa y tendencia al sobreajuste en datasets de entrenamiento. Esta observación no es marginal: señala que el entusiasmo mediático y comercial por la IA en medicina ha superado con creces la solidez de la evidencia científica que la sustenta.

Figura 1 Arquitectura Epistémica de la Simbiosis Humano-IA en la Medicina Basada en la Evidencia



Fuente: Elaboración propia con apoyo de inteligencia artificial

https://doi.org/10.62486/978-9915-9928-3-9.Ch21

5.2 EVIDENCIA DE EFECTIVIDAD REAL VS. RENDIMIENTO EN CONDICIONES DE LABORATORIO

Existe una brecha documentada entre el rendimiento de los modelos de IA en condiciones controladas de laboratorio con datasets curados, homogéneos y representativos del contexto de entrenamiento y su desempeño en entornos clínicos reales, donde la variabilidad del paciente, la calidad de los registros, el ruido en los datos y la diversidad epidemiológica reducen significativamente su precisión y confiabilidad.

Esta brecha entre eficacia interna y efectividad externa tiene consecuencias directas para la práctica: un modelo que ha demostrado rendimiento superior al del clínico en condiciones controladas puede presentar un rendimiento inferior al del juicio clínico experto cuando se implementa en un hospital con población, equipamiento y flujos de trabajo distintos. La ausencia de requisitos regulatorios universales de validación en condiciones de uso real salvo en jurisdicciones avanzadas como la UE o los EE.UU. para dispositivos de alto riesgo agrava esta situación.

6. LIMITACIONES METODOLÓGICAS DE LA IA CLÍNICA Y LOS GRANDES MODELOS DE LENGUAJE

Los LLMsentre los cuales se encuentran modelos como GPT, Gemini, LLaMA y sus derivados médicos especializados han demostrado capacidades notables en tareas de comprensión de lenguaje natural, síntesis de información médica y resolución de preguntas de exámenes médicos de licenciatura. Sin embargo, estas capacidades coexisten con limitaciones metodológicas fundamentales que cualquier clínico debe comprender antes de integrar estas herramientas en su práctica.

6.1 ALUCINACIONES FACTOLÓGICAS

El fenómeno de alucinación en LLMsdefinido como la generación de contenido factualmente incorrecto, no verificable o directamente inexistente, presentado con apariencia de credibilidad constituye su limitación más peligrosa en el contexto médico. En medicina, donde la precisión de la información puede determinar la vida o la muerte de un paciente, la generación de diagnósticos incorrectos, dosis farmacológicas erróneas o referencias bibliográficas inventadas no es un error tolerable.

Estudios recientes han cuantificado la magnitud de este problema. Se ha documentado que modelos como GPT y Gemini generan referencias bibliográficas fabricadas en proporciones que oscilan entre el 25% y el 50% de sus outputs cuando son utilizados como herramientas de investigación médica (esto se encuentra en constante evolución de acuerdo con nuevos modelos desarrollados). Esta fabricación de referencias en ocasiones con DOIs plausibles y autores verosímiles pero inexistentes representa un riesgo específico en la educación médica, donde el estudiante puede integrar fuentes falsas en trabajos académicos o basar decisiones clínicas en evidencia que no existe.

Un estudio publicado en JAMA Network Open en 2024 evaluó la precisión, consistencia y tasa de alucinación de LLMs en el análisis de notas clínicas no estructuradas en registros electrónicos de pacientes con traumatismo craneoencefálico, encontrando que incluso los modelos con mejor rendimiento general presentaban tasas de error no despreciables en la extracción e interpretación de información clínica crítica.

6.2 INTERPRETACIÓN ESTADÍSTICA Y RAZONAMIENTO PROBABILÍSTICO

Los LLMs presentan limitaciones documentadas en la interpretación de medidas de asociación epidemiológica y estimadores de efecto clínico. La confusión entre odds ratio (OR) y riesgo relativo (RR), la aplicación incorrecta del concepto de número necesario a tratar (NNT), la conflación de correlación con causalidad y la subestimación de la importancia del intervalo de confianza como expresión de incertidumbre son errores frecuentes observados en la literatura. Estas limitaciones son especialmente preocupantes cuando los modelos son utilizados para sintetizar evidencia o interpretar resultados de ensayos clínicos, tareas en las que la precisión estadística es condición necesaria para no introducir errores sistemáticos en la cadena de razonamiento clínico.

https://doi.org/10.62486/978-9915-9928-3-9.Ch21

6.3 SESGO DE CONFIRMACIÓN INDUCIDO POR IA

El sesgo de confirmación la tendencia cognitiva a buscar, interpretar y recordar la información de forma que confirme las creencias previas puede ser amplificado por los LLMs en el contexto del razonamiento diagnóstico. Cuando un clínico consulta un LLM con una hipótesis diagnóstica ya formulada, el modelo tiende a generar respuestas que refuerzan esa hipótesis, en razón de su entrenamiento por refuerzo con retroalimentación humana (RLHF), que penaliza las respuestas percibidas como desafiantes o contraintuitivas por el usuario. Este fenómeno, conocido como sycophancy en la literatura técnica, puede inducir al clínico a descartar hipótesis diagnósticas alternativas que habrían requerido ser exploradas.

6.4 EL ESTUDIO 'YOUR BRAIN ONCHATGPT' DEL MIT MEDIA LAB: ANÁLISIS CRÍTICO

En junio de 2025, el MIT Media Lab publicó como preprint (arXiv:2506.08872) un estudio titulado 'Your Brain onChatGPT: Accumulationof Cognitive DebtwhenUsingan AI Assistantfor Essay WritingTask' (Kosmyna et al., 2025), que ha generado un debate significativo sobre los efectos cognitivos del uso continuado de LLMs en tareas de escritura y razonamiento.

DISEÑO DEL ESTUDIO

El estudio utilizó un diseño de tres grupos paralelos (n=54 participantes, de 18 a 39 años, reclutados en el área de Boston) con asignación a condición de escritura con ChatGPT, con motor de búsqueda web, o sin herramientas externas (grupo control). Los participantes completaron tres sesiones de escritura de ensayos bajo la misma asignación de condición. En una cuarta sesión, los participantes del grupo ChatGPT fueron reasignados al grupo sin herramientas y viceversa. Se utilizó electroencefalografía (EEG) de 32 canales para registrar la actividad cerebral y se realizaron análisis de procesamiento del lenguaje natural (NLP) sobre los ensayos producidos.

HALLAZGOS PRINCIPALES

El análisis de coherencia EEG demostró que el grupo ChatGPT presentó la conectividad neural más débil en comparación con los otros dos grupos, con menor activación de redes asociadas al control ejecutivo, la atención sostenida y el procesamiento semántico. El grupo sin herramientas externas mostró la conectividad neural más amplia y robusta, especialmente en bandas alfa, theta y delta asociadas con creatividad, carga de memoria de trabajo y procesamiento semántico. En la cuarta sesión, los participantes que habían utilizado ChatGPT mostraron menor conectividad neural al ser reasignados al grupo sin herramientas, lo que los autores interpretan como evidencia de 'deuda cognitiva acumulada'. El análisis NLP reveló una mayor homogeneidad léxica y temática en los ensayos del grupo ChatGPT. Los participantes del grupo ChatGPT reportaron menor sentido de autoría e incapacidad para citar con precisión el contenido de sus propios ensayos.

LIMITACIONES METODOLÓGICAS Y CARÁCTER PRELIMINAR

Es imperativo que el lector comprenda el estatus científico de este trabajo. En el momento de su publicación, el estudio no había sido sometido a revisión por pares, y como tal debe ser considerado evidencia preliminar sujeta a modificación sustancial. Las limitaciones metodológicas son significativas: la muestra es pequeña (54 participantes en las tres primeras sesiones, 18 en la cuarta), reclutada en una región geográfica específica y con acceso desproporcionado a participantes universitarios de Boston; la generalización a otros contextos educativos, culturales y socioeconómicos es limitada; el diseño no permite controlar por diferencias de habilidad inicial en escritura; y las medidas de EEG de coherencia neural requieren validación con paradigmas experimentales más complejos.

Diversas críticas metodológicas señalan que la menor conectividad neural durante el uso de ChatGPT puede reflejar eficiencia cognitiva en lugar de deterioro, que el ecológico externo del paradigma de escritura de ensayos tipo SAT es cuestionable como modelo de la práctica clínica, y que la interpretación de 'deuda cognitiva' como fenómeno persistente requeriría un seguimiento

https://doi.org/10.62486/978-9915-9928-3-9.Ch21

longitudinal sustancialmente más prolongado.

Dicho esto, los hallazgos son coherentes con un corpus creciente de evidencia sobre la relación entre el uso de tecnologías cognitivamente delegantes y la reducción de la implicación activa del usuario en el procesamiento de información. Para la educación médica, estos datos aunque preliminares sugieren que la sustitución del razonamiento clínico activo por la consulta pasiva de LLMs puede tener consecuencias cognitivas que merecen investigación sistemática rigurosa.

7. SESGOS ALGORÍTMICOS Y DESIGUALDAD SANITARIA

El sesgo en los sistemas de IA no es un fenómeno accidental ni fácilmente corregible: es una consecuencia estructural de los procesos mediante los cuales se generan, seleccionan y etiquetan los datos utilizados para el entrenamiento. Cuando los datos de entrenamiento reflejan desigualdades históricas en el acceso a la atención médica, en la representación de minorías en ensayos clínicos, o en la calidad diferencial de los registros clínicos entre poblaciones, el modelo aprende y reproduce y en ocasiones amplifica esas desigualdades.

7.1 TIPOLOGÍA DEL SESGO ALGORÍTMICO EN SALUD

Una revisión sistemática de Chen et al. (2024), publicada en Journalofthe American Medical InformaticsAssociation (JAMIA), identificó seis tipos principales de sesgo en modelos de IA basados en registros electrónicos de salud: sesgo algorítmico, sesgo de confusión, sesgo implícito, sesgo de medición, sesgo de selección y sesgo temporal. Esta taxonomía es relevante para entender por qué la corrección del sesgo en IA médica no puede reducirse a técnicas de rebalanceo de clases en el dataset de entrenamiento, sino que requiere intervención en múltiples etapas del ciclo de vida del modelo.

El sesgo de selección ocurre cuando la población representada en el dataset de entrenamiento no refleja la distribución de pacientes que recibirán la intervención del modelo en la práctica real. En el contexto latinoamericano, la práctica totalidad de los modelos de IA para diagnóstico por imagen, predicción de sepsis, estratificación de riesgo cardiovascular o apoyo al triaje han sido entrenados en datasets de EE.UU. o Europa occidental. La representación de poblaciones latinoamericanas, indígenas, afrodescendientes o residentes en zonas de alta altitud en estos datasets es, en el mejor de los casos, marginal.

7.2 SESGOS RACIALES Y GEOGRÁFICOS EN MODELOS GLOBALES

La literatura documenta con creciente sistematicidad que los modelos de IA en dermatología, nefrología, cardiología y medicina intensiva presentan un rendimiento significativamente inferior en pacientes de piel oscura, en mujeres, en pacientes de edad avanzada y en poblaciones de bajos recursos. En dermatología, por ejemplo, los modelos de detección de lesiones cutáneas malignas muestran tasas de sensibilidad sustancialmente menores en pacientes con fototipos de Fitzpatrick IV-VI, en razón de que la gran mayoría de los datasets de entrenamiento contienen imágenes de pacientes de piel clara.

En nefrología, algoritmos de estimación de la tasa de filtración glomerular que incorporaban factores de corrección racial como el algoritmo MDRD con ajuste por raza negra han sido retirados de uso clínico en varios países tras demostrar que introducían correcciones sistemáticas no justificadas biológicamente.

Para Latinoamérica, la situación es estructuralmente similar pero agravada por la escasez de datos digitalizados en sistemas de información de salud interoperables, la heterogeneidad de los sistemas de registro clínico, y la ausencia de políticas nacionales de datos de salud que faciliten la creación de datasets representativos de las poblaciones locales.

https://doi.org/10.62486/978-9915-9928-3-9.Ch21

Figura 2 Mecanismos de sesgo algorítmico en los sistemas de salud

8. INTELIGENCIA ARTIFICIAL Y MEDICINA EN ALTURA: LIMITACIONES CRÍTICAS EN POBLACIONES ANDINAS

Esta sección aborda uno de los vacíos más críticos y menos discutidos en la literatura internacional sobre IA en salud: la inaplicabilidad o aplicabilidad restringida de los modelos algorítmicos globales en poblaciones de alta altitud, con énfasis en las comunidades andinas de Bolivia, Perú, Ecuador y Chile.

8.1 FISIOLOGÍA DE ALTURA: BASE BIOLÓGICA DE LA DIFERENCIA

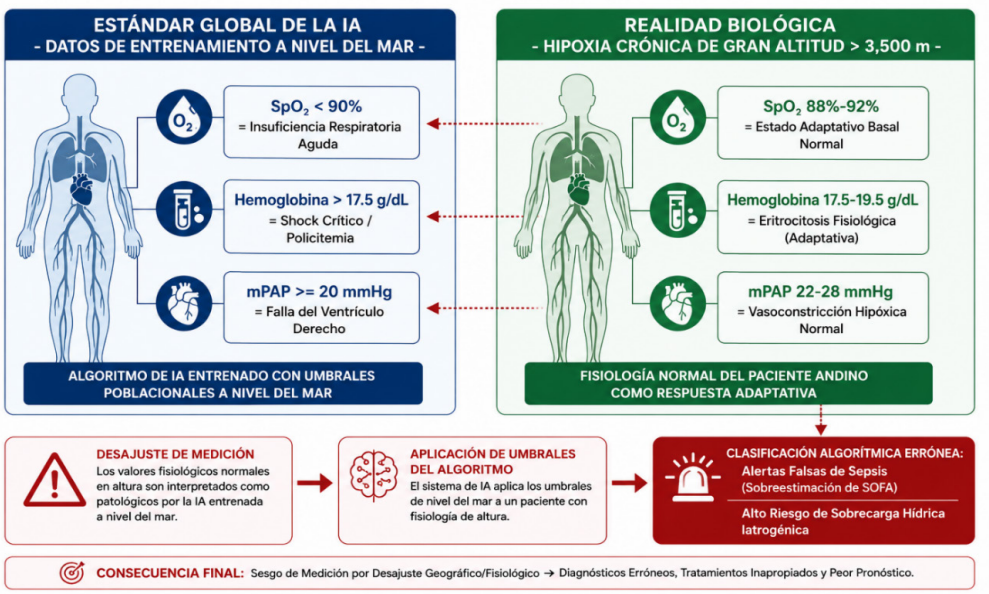
Las poblaciones humanas que habitan permanentemente a altitudes superiores a 2.500 m.s.n.m. presentan adaptaciones fisiológicas profundas, resultado de miles de años de presión selectiva y de mecanismos de aclimatación adquiridos a lo largo de la vida individual. Estas adaptaciones no son variantes menores respecto al fenotipo de referencia de nivel del mar: representan diferencias cuantificables en parámetros hematológicos, cardiovasculares, respiratorios y metabólicos que tienen consecuencias directas para la interpretación clínica.

En el plano hematológico, las poblaciones andinas incluyendo comunidades aimaras y quechuas de Bolivia y Perú muestran valores de hemoglobina, hematocrito y eritropoyetina significativamente superiores a los de las poblaciones de baja altitud, como mecanismo adaptativo frente a la hipoxia hipobárica crónica.

En el plano respiratorio, la saturación periférica de oxígeno (SpO2) basal es sistemáticamente inferior a los valores de referencia de nivel del mar: estudios realizados en la ciudad de La Paz, Bolivia (3.640 m.s.n.m.) y en comunidades rurales del altiplano peruano por encima de los 4.000 m.s.n.m. documentan valores de SpO2 normales de entre 85% y 93%, que un algoritmo entrenado en poblaciones de nivel del mar clasificaría como hipoxemia moderada o grave que requiere intervención urgente.

https://doi.org/10.62486/978-9915-9928-3-9.Ch21

Figura 3 Infografía del sesgo de medición algorítmico por desajuste biogeográfico.



8.2 ADAPTACIONES CARDIOVASCULARES Y HEMATOLÓGICAS RELEVANTES PARA LA IA

Más allá de la SpO2, las poblaciones andinas presentan adaptaciones cardiovasculares que alteran de manera sustancial la interpretación de parámetros clínicos utilizados como insumos por los principales algoritmos de IA en medicina intensiva.

La presión arterial pulmonar basal es significativamente más elevada en poblaciones de altitud, como consecuencia de la vasoconstricción hipóxica pulmonar crónica. Los índices de variabilidad de la frecuencia cardíaca, los valores de troponina basal, los parámetros de función ventricular derecha y la morfología del electrocardiograma pueden diferir significativamente de los valores de referencia de nivel del mar, no como expresión de patología, sino como fenotipo adaptado.

Un estudio publicado en BMC Research Notes documentó diferencias individuales y por sexo en los parámetros hemodinámicos y SpO2 de jóvenes aymaras en La Paz, Bolivia, subrayando la variabilidad intraindividual significativa incluso dentro de las poblaciones de altitud. Esta variabilidad es un desafío adicional para los modelos de IA, que típicamente están calibrados para detectar desviaciones respecto a valores de referencia poblacionales únicos y fijos.

8.3 RIESGO DE CLASIFICACIÓN INCORRECTA POR ALGORITMOS NO ADAPTADOS

La consecuencia clínica más inmediata y grave de la aplicación de algoritmos entrenados en poblaciones de nivel del mar a pacientes andinos es el riesgo de clasificación incorrecta en sistemas de triaje, algoritmos de alerta temprana, predictores de sepsis y modelos de decisión en medicina intensiva.

Considérese, como ilustración de alto valor didáctico, el funcionamiento del National Early Warning Score 2 (NEWS2) o de los criterios SOFA de disfunción orgánica, que utilizan como umbrales clínicos valores de SpO2 y frecuencia respiratoria derivados de poblaciones de nivel del mar. Un paciente andino de 45 años con SpO2 de 89% en La Paz puede ser clasificado por un sistema de alerta temprana basado en IA como paciente con insuficiencia respiratoria severa que requiere intervención urgente, cuando en realidad se encuentra en su estado basal habitual. A la inversa, un paciente con neumonía grave que haya caído desde su SpO2 basal de 91% a 85% puede no activar los umbrales de alerta de un sistema diseñado para poblaciones con SpO2 basal de 98%, generando un error de infratriaje potencialmente fatal.

El riesgo inverso la subestimación de la gravedad clínica es quizás más peligrosa. En un paciente andino con sepsis grave, una SpO2 de 84% puede ser considerada 'normal para la altitud' por un algoritmo que incorpore correctivos de altitud mal calibrados, retrasando la intervención

https://doi.org/10.62486/978-9915-9928-3-9.Ch21

terapéutica en una situación de alto riesgo de mortalidad.

8.4 AUSENCIA DE REPRESENTACIÓN ANDINA EN DATASETS GLOBALES

La representación de poblaciones andinas bolivianas, peruanas, ecuatorianas, chilenas del altiplano en los principales datasets utilizados para el entrenamiento de modelos de IA en salud a nivel global es mínima o inexistente. Bases de datos de referencia como MIMIC-III/IV (Beth Israel Deaconess Medical Center, Boston), eICUCollaborativeResearchDatabase (EE.UU.) o UK Biobank, que han servido como sustrato de entrenamiento para decenas de modelos de IA en medicina intensiva, sepsis, predicción de mortalidad y triaje, no contienen datos de pacientes a altitudes superiores a 200-300 m.s.n.m.

Esta ausencia no es simplemente una laguna estadística: es una forma de exclusión epistémica que convierte a las poblaciones de altura en 'invisibles' para los algoritmos que determinarán, en un futuro próximo, decisiones de triaje, asignación de recursos y guías terapéuticas. La aplicación de estos modelos en sistemas de salud de Bolivia o Perú ya sea directamente o a través de herramientas de IA incorporadas en sistemas de historia clínica electrónica o dispositivos de monitorización sin validación local previa constituye un riesgo sanitario de primer orden y una violación del principio de no maleficencia.

8.5 ALGORITMOS DE SEPSIS, VENTILACIÓN MECÁNICA Y PREDICCIÓN DE MORTALIDAD EN ALTITUD

Los algoritmos de predicción de sepsis incluyendo qSOFA, SOFA y sus versiones enriquecidas con IA utilizan como criterios de disfunción orgánica parámetros fisiológicos cuya interpretación es altitud-dependiente. La frecuencia respiratoria de 22 rpm como criterio de alarma, la SpO2 inferior a 94% como indicador de compromiso respiratorio, o los parámetros de falla orgánica basados en gases arteriales calibrados para nivel del mar, son herramientas cuya validez externa en contextos de alta altitud no ha sido sistemáticamente evaluada.

En ventilación mecánica, las estrategias de protección pulmonar objetivos de SpO2, presiones de plateau, conducción de PEEP deben ser ajustadas a la fisiología de la altitud. Los algoritmos de IA para optimización de la ventilación mecánica que no incorporan la altitud como covariable estructural están, por definición, mal especificados para el entorno clínico de una UCI boliviana o peruana.

La predicción de mortalidad mediante scores como APACHE II, SAPS III o sus sucesores basados en aprendizaje automático exhiben el mismo problema: fueron derivados y validados en cohortes de nivel del mar, y sus coeficientes de calibración que determinan la probabilidad predicha de mortalidad para un perfil fisiológico dado pueden subestimar o sobreestimar sistemáticamente el riesgo en pacientes de altitud.

8.6 IMPACTO SOBRE LA EQUIDAD SANITARIA Y NECESIDAD DE DATASETS REGIONALES

El problema descrito no es meramente técnico: es un problema de equidad sanitaria con implicaciones éticas y políticas de primera magnitud. Las poblaciones andinas ya enfrentan disparidades significativas en el acceso a tecnología médica avanzada. La introducción de herramientas de IA que no han sido validadas en sus contextos fisiológicos específicos no reduce estas disparidades: las perpetúa bajo una nueva forma, más opaca y, por ello, más difícil de cuestionar.

La solución requiere una estrategia de dos vías:

- En el corto plazo, la implementación de cualquier sistema de IA en entornos clínicos de alta altitud debe estar precedida de una validación local rigurosa, con datos de pacientes reales de la región, y de un proceso de adaptación de umbrales y valores de referencia a la fisiología local.
- En el mediano y largo plazo, es imperativo el desarrollo de datasets regionales andinos que alimenten modelos específicamente entrenados para estas poblaciones. Iniciativas como la creación de un registro clínico latinoamericano de alta altitud, con estándares de datos interoperables y gobernanza participativa, son condición necesaria para que la IA contribuya a reducir en lugar de ampliar las brechas sanitarias en la región.

Otra solución viable es la preparación de Datasets pequeños (Small data) de alta fidelidad,

https://doi.org/10.62486/978-9915-9928-3-9.Ch21

sin alteraciones que hagan difícil su análisis, para realizar Ajuste fino (Fine Tunnig) con grandes modelos de IA pre-entrenadosy mediante herramientas de balanceo de clases se puedan reentrenar IAs con estos datos incluidos así poder validarlos localmente, esto siempre y cuando el código fuente, los algoritmos y las capas neuronales de los modelos predictivos sean de acceso libre.

Tabla 1 Riesgos metodológicos y clínicos asociados al uso de inteligencia artificial en la lectura crítica biomédica

Riesgo o Sesgo	Manifestación en IA biomédica	Consecuencia clínica o académica	Estrategia de mitigación
Alucinación algorítmica	Generación de datos o referencias inexistentes	Interpretaciones erróneas y propagación de información falsa	Verificación obligatoria con el artículo original y fuentes primarias
Sesgo de selección	Entrenamiento con datasets no representativos	Mala generalización en poblaciones especiales (v.g., medicina de altura)	Validación externa multicéntrica y ajuste contextual
Sesgo de medición	Interpretación incorrecta de variables fisiológicas adaptativas	Sobrediagnóstico o tratamientos innecesarios	Incorporación de priors fisiológicos regionales
Automation bias	Aceptación acrítica de recomendaciones algorítmicas	Reducción del razonamiento clínico independiente	Auditoría humana obligatoria y formación metodológica
Simplificación excesiva	Resúmenes incompletos o reducción de complejidad metodológica	Pérdida de matices clínicos importantes	Lectura crítica paralela del manuscrito completo
Confusión metodológica	Incapacidad para distinguir causalidad de asociación	Conclusiones clínicas inapropiadas	Supervisión por clínicos con entrenamiento en MBE
Sesgo de entrenamiento	Reproducción de inequidades presentes en los datos originales	Amplificación de disparidades sanitarias	Curación ética y geográficamente diversa de datasets
Falsa sensación de comprensión	Aparente claridad sin comprensión profunda	Sobreestimación de competencias metodológicas	Integración de enseñanza formal en epidemiología clínica
Dependencia cognitiva	Delegación progresiva del análisis crítico a la IA	Deterioro del razonamiento clínico autónomo	Uso complementario y no sustitutivo de IA
Riesgos éticos y de privacidad	Introducción de datos sensibles en plataformas externas	Vulneración de confidencialidad y protección de datos	Anonimización estricta y uso de plataformas seguras

Fuente: Elaboración propia.

9. RIESGOS COGNITIVOS, AUTOMATION BIAS Y EDUCACIÓN MÉDICA

La adopción de herramientas de IA en la formación médica y en la práctica clínica plantea un conjunto de riesgos cognitivos que van más allá de los errores técnicos del modelo y se instalan en la estructura del razonamiento profesional médico. Estos riesgos incluyen el automationbias, el deskilling clínico, la dependencia cognitiva, la reducción del pensamiento crítico y la erosión de la relación médico-paciente.

9.1 AUTOMATION BIAS: EVIDENCIA Y MAGNITUD

El automationbiasdefinido como la tendencia a favorecer las recomendaciones de sistemas automatizados incluso cuando contradicen el juicio clínico propio o la evidencia disponible es un fenómeno bien documentado en medicina. En radiología, un estudio prospectivo aleatorizado con 27 radiólogos especialistas en mama demostró que la exposición a indicaciones de IA erróneas incrementó la tasa de falsos positivos hasta en un 12%, incluso entre lectores experimentados.

En endoscopia, un ensayo clínico multicéntrico aleatorizado documentó que la tasa de detección de adenomas sin asistencia de IA cayó significativamente (de 28,4% a 22,4%) después de un período de uso asistido por IA, mientras que la tasa se mantenía estable con asistencia continua de IA, sugiriendo una dependencia funcional del endoscopista respecto al sistema.

En patología computacional, estudios experimentales documentaron que más del 30% de los participantes revirtieron diagnósticos inicialmente correctos al ser expuestos a sugerencias incorrectas de IA bajo condiciones de presión temporal. Estas cifras no son anecdóticas: señalan que el automationbias en medicina no es un fenómeno marginal ni limitado a clínicos inexpertos, sino una vulnerabilidad cognitiva sistemática que afecta incluso a especialistas de alto nivel.

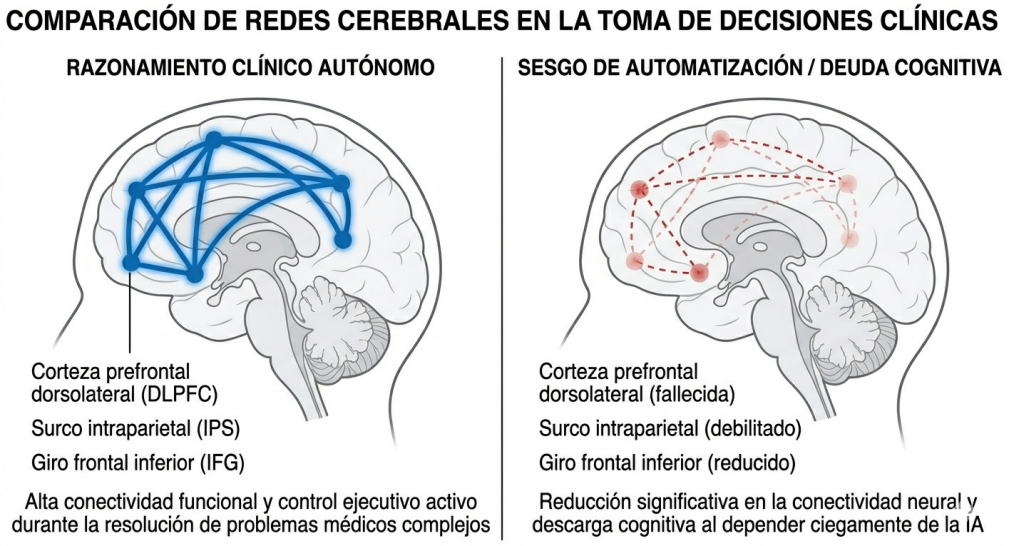
https://doi.org/10.62486/978-9915-9928-3-9.Ch21

9.2 DESKILLING CLÍNICO

El deskillingla erosión gradual de competencias clínicas por dependencia crónica de sistemas automatizados es un fenómeno análogo al que se ha observado en otras profesiones de alta complejidad técnica, como la aviación. El término fue popularizado en el contexto médico a partir de la observación de que la automatización de ciertas tareas diagnósticas o terapéuticas puede reducir las oportunidades de práctica deliberada que son necesarias para mantener y desarrollar la competencia clínica experta.

Una revisión de la literatura publicada en Artificial IntelligenceReview en 2025 (Natali, Marconi et al.) identificó que el deskilling en medicina afecta al menos cuatro dimensiones de la competencia clínica: las habilidades técnicas (exploración física, procedimientos), las habilidades cognitivas (diagnóstico diferencial, razonamiento clínico), las habilidades éticas (deliberación sobre valores en situaciones de incertidumbre) y las habilidades de interpretación contextual (integración de información verbal, no verbal y contextual en la consulta). El riesgo es especialmente relevante en el contexto de la formación de grado y posgrado, donde el uso de LLMs como atajos para el razonamiento diagnóstico puede interferir con el desarrollo de competencias que requieren práctica iterativa y reflexión crítica.

Figura 4 Mapeo comparativo de conectividad funcional frontoparietal por EEG.



9.3 DEPENDENCIA COGNITIVA Y PÉRDIDA DE RAZONAMIENTO CRÍTICO

La dependencia cognitiva que es el uso habitual de herramientas externas como sustitutos del razonamiento propio es una consecuencia posible del uso no supervisado e indiscriminado de LLMs en educación médica. El estudio del MIT Media Lab, aunque preliminar y con las limitaciones metodológicas ya discutidas, ofrece una señal empírica relevante: los participantes que utilizaron ChatGPT como herramienta principal de escritura exhibieron patrones de actividad cerebral compatibles con menor implicación cognitiva activa, y presentaron dificultades para recuperar y elaborar de manera autónoma el contenido de sus propios textos en sesiones posteriores.

Para la educación médica, esta señal es especialmente perturbadora. El razonamiento clínico experto se construye sobre una base de experiencias elaboradas activamente, errores cometidos y corregidos, patrones reconocidos mediante la práctica repetida y marcos conceptuales construidos con esfuerzo cognitivo sostenido.

Si el uso de LLMs en la formación médica reduce sistemáticamente el esfuerzo cognitivo invertido en estas tareas, el riesgo no es únicamente el de producir graduados con menor capacidad de razonamiento autónomo: es el de comprometer la seguridad del paciente en un plazo que puede ser más corto de lo que los sistemas de formación médica están equipados para detectar.

https://doi.org/10.62486/978-9915-9928-3-9.Ch21

10. GOBERNANZA DE DATOS MÉDICOS Y REGULACIÓN INTERNACIONAL DE LA IA EN SALUD

El paisaje regulatorio de la IA en salud ha experimentado una transformación significativa entre 2021 y 2025, transitando desde un paradigma de autorregulación voluntaria hacia marcos normativos con carácter vinculante en las jurisdicciones más avanzadas. Sin embargo, la heterogeneidad geográfica de este proceso regulatorio y la velocidad de la innovación tecnológica plantean desafíos persistentes de coherencia, aplicabilidad y equidad global.

10.1 MARCO REGULATORIO INTERNACIONAL: ESTADO ACTUAL

La Unión Europea aprobó en marzo de 2024 el AI Act, el primer marco jurídico supranacional legalmente vinculante que regula la IA de manera exhaustiva. El AI Act clasifica los sistemas de IA según su nivel de riesgo: los sistemas de IA para diagnóstico médico, apoyo a la decisión clínica, triaje y gestión de enfermedades críticas son clasificados como de alto riesgo, lo que implica obligaciones de evaluación de conformidad previa al mercado, registro, supervisión humana, explicabilidad y transparencia.

En septiembre de 2024, el Consejo de Europa abrió a la firma el Convenio Marco sobre IA y Derechos Humanos, el primer tratado internacional jurídicamente vinculante que integra protecciones de derechos humanos en el ciclo completo de vida de los sistemas de IA.

En los EE.UU., la FDA ha adoptado una posición regulatoria más flexible pero progresivamente más exigente: a mayo de 2024, el 97% de los dispositivos de IA y aprendizaje automático aprobados lo habían sido mediante la vía 510(k), una ruta de equivalencia sustancial que no exige la misma profundidad de evaluación que la aprobación premarket (PMA). La FDA ha publicado guías de práctica de aprendizaje automático de alta calidad (Good Machine LearningPractice, GMLP) y la Guía de IA/ML-SaMD (Software as a Medical Device), pero la velocidad de implementación regulatoria sigue siendo significativamente inferior a la velocidad de innovación.

A nivel multilateral, en 2023 la OMS, la Unión Internacional de Telecomunicaciones (UIT) y la Organización Mundial de la Propiedad Intelectual (OMPI) lanzaron la Iniciativa Global sobre IA para la Salud (GI-AI4H), orientada a desarrollar estándares técnicos, guías regulatorias y mecanismos de gobernanza armonizados. La 77.ª Asamblea Mundial de la Salud (2024) celebró una mesa redonda estratégica sobre IA en salud con participación de ministerios de salud, academia y sector privado, y señaló la necesidad urgente de fortalecer la capacidad regulatoria de los países de bajos y medianos ingresos.

La OCDE publicó en enero de 2024 el informe CollectiveActionforResponsible AI in Health, adoptado por los ministros de salud de los países miembros, que formula cuatro recomendaciones de acción colectiva: adoptar un enfoque ágil de regulación y supervisión; establecer alianzas público-privadas centradas en el paciente; elevar la gobernanza de datos sanitarios; y reforzar la capacidad institucional para implementar IA responsable.

10.2 GOBERNANZA DE DATOS EN PAÍSES DE BAJOS Y MEDIANOS INGRESOS

Para los países de América Latina y el Caribe, incluyendo Bolivia y Perú, los marcos regulatorios nacionales para la IA en salud son embrionarios o inexistentes. La mayoría de los países de la región carecen de legislación específica sobre IA en salud, de registros nacionales de dispositivos de IA médica, de capacidades de auditoría algorítmica en los organismos reguladores de salud, y de políticas de soberanía de datos sanitarios que limiten la transferencia de datos clínicos a corporaciones extranjeras sin consentimiento informado adecuado.

Esta fragilidad regulatoria no es solo un problema de capacidad institucional: es un riesgo sanitario activo. En ausencia de marcos regulatorios locales, las herramientas de IA para salud pueden ser importadas, comercializadas e implementadas en sistemas de salud de la región sin haber sido validadas localmente, sin que existan mecanismos de vigilancia post-implementación, y sin que los profesionales de salud dispongan de información suficiente para evaluar críticamente sus indicaciones.

11. PRINCIPIOS PARA UNA INTEGRACIÓN ÉTICA Y SEGURA DE LA IA EN MEDICINA

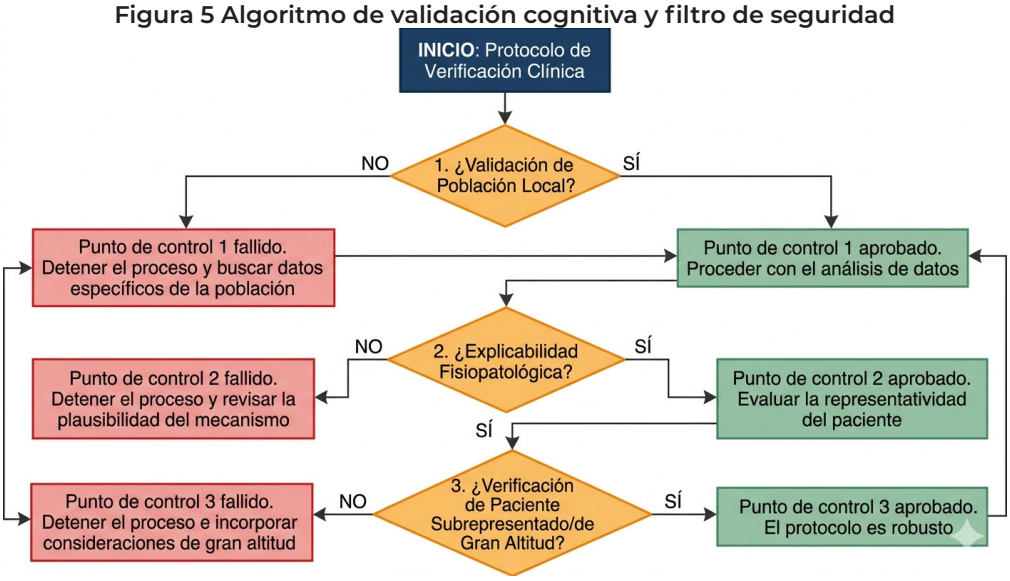
La pregunta relevante para la práctica clínica y la política sanitaria no es si la IA debe ser incorporada a la medicina, sino cómo debe ser incorporada de manera que maximice los

https://doi.org/10.62486/978-9915-9928-3-9.Ch21

beneficios reales, minimice los daños evitables y promueva la equidad sanitaria. Los principios que se presentan a continuación no son restricciones al uso de la IA: son las condiciones bajo las cuales su uso puede considerarse éticamente justificado y metodológicamente sólido.

- **Validación contextual previa a la implementación:** Todo sistema de IA que vaya a ser utilizado en una población específica debe ser validado en datos representativos de esa población, incluyendo su contexto geográfico, epidemiológico y fisiológico. En entornos de alta altitud, esta validación debe incluir explícitamente los parámetros fisiológicos característicos de la altitud.
- **Explicabilidad como condición de uso clínico:** Los sistemas de IA clasificados como de alto riesgo diagnóstico, triaje, predicción de mortalidad, apoyo a la decisión en medicina intensiva deben ser capaces de proveer una justificación comprensible de sus recomendaciones. El uso de sistemas de caja negra en estas aplicaciones, sin mecanismos de explicabilidad, no es éticamente justificable.
- **Supervisión humana no negociable:** La responsabilidad última de las decisiones clínicas debe permanecer en el médico, no en el sistema de IA. Los protocolos de uso de IA en salud deben incluir procedimientos explícitos para la verificación crítica de los outputs del sistema y para la anulación de sus recomendaciones cuando el juicio clínico lo indique.
- **Auditoría de equidad pre-implementación y en seguimiento:** Los sistemas de IA deben ser auditados respecto a su rendimiento diferencial en subgrupos de la población antes de su implementación y de manera periódica durante su uso. Las auditorías deben incluir grupos raciales, étnicos, por género, edad y, en contextos pertinentes, por altitud de residencia.
- **Consentimiento informado adaptado al uso de IA:** Los pacientes tienen derecho a ser informados de cuándo sus datos clínicos son procesados por un sistema de IA y de cuándo una recomendación clínica está siendo asistida por un algoritmo. El consentimiento informado debe evolucionar para capturar estos elementos.
- **Gobernanza de datos orientada a la equidad**: Los datos clínicos de las poblaciones de países de medianos y bajos ingresos no deben ser utilizados para entrenar modelos comerciales sin mecanismos de retribución y beneficio local. La soberanía de los datos sanitarios es un componente de la soberanía en salud pública.
- **Alfabetización crítica en IA como competencia médica fundamental:** Los programas de formación médica de grado y posgrado deben incluir competencias en evaluación crítica de herramientas de IA, identificación de sesgos algorítmicos, interpretación de incertidumbre cuantitativa y uso supervisado de LLMs. La confianza acrítica en las salidas de un LLM es tan peligrosa como la ignorancia de su existencia.
- **Investigación clínica prioritaria con estándares MBE**: El desarrollo de evidencia sobre eficacia, seguridad y equidad de los sistemas de IA en salud debe seguir los estándares metodológicos de la MBE, incluyendo ECAs cuando sea posible y revisiones sistemáticas con metaanálisis. Los estudios observacionales retrospectivos en datasets no representativos no son suficientes para justificar la implementación clínica rutinaria.

Figura 5 Algoritmo de validación cognitiva y filtro de seguridad



Fuente: Elaboración propia con apoyo de Inteligencia artificial

https://doi.org/10.62486/978-9915-9928-3-9.Ch21

12. PERSPECTIVAS FUTURAS

El horizonte de la IA en medicina es de una riqueza y complejidad que excede cualquier proyección lineal. Sin embargo, es posible identificar tendencias que marcarán la evolución del campo en los próximos cinco a diez años y que tienen implicaciones directas para la práctica clínica, la formación médica y la política sanitaria en el contexto latinoamericano.

La especialización de los LLMs en dominios médicos específicos —con modelos como Med-PaLM 2, BioMedLM o sus sucesores— seguirá mejorando el rendimiento en tareas de recuperación de información médica y apoyo a la decisión, sin resolver necesariamente los problemas estructurales de sesgos en datasets, alucinaciones en contextos de alta incertidumbre o automationbias. La proliferación de modelos especializados multiplica, paradójicamente, los desafíos de evaluación: cada nuevo modelo requiere validación contextual específica en las poblaciones en que va a ser utilizado.

En el ámbito regulatorio, la próxima década verá probablemente la extensión de marcos jurídicamente vinculantes hacia más regiones del mundo, impulsada por el EU AI Act y los instrumentos del Consejo de Europa. Para América Latina, esto representará tanto una oportunidad adoptar estándares robustos como referencia como un riesgo de colonialismo regulatorio si los marcos adoptados no son adaptados a las capacidades institucionales y las prioridades sanitarias locales.

La inteligencia artificial en salud de baja altitud será progresivamente más potente y eficaz. El riesgo es que esta mejora se produzca en un gradiente creciente de inequidad si las poblaciones de altitud, indígenas, rurales y de bajos recursos no son incluidas sistemáticamente en los datasets de entrenamiento y validación. La creación de registros clínicos latinoamericanos de alta altitud, interoperables y con gobernanza participativa, es una prioridad de investigación y política pública que no admite más dilación.

13. CONCLUSIONES

La inteligencia artificial representa una transformación genuina y de largo alcance en la medicina. Sus capacidades reales en diagnóstico por imagen, detección precoz de condiciones críticas, gestión de la información clínica y apoyo a la decisión son documentables y significativas. Sin embargo, la adopción de IA en medicina sin un marco crítico sólido, anclado en los principios de la medicina basada en evidencia y la bioética contemporánea, expone a los pacientes a riesgos evitables que no han recibido la atención proporcional a su magnitud.

Los sesgos algorítmicos son una consecuencia estructural del desequilibrio en la representación demográfica de los datasets globales, y no pueden corregirse mediante ajustes técnicos menores. Las poblaciones andinas aymaras, quechuas, y otras comunidades indígenas de Bolivia, Perú y los Andes en general presentan un perfil fisiológico que ningún modelo entrenado en datos de nivel del mar puede representar adecuadamente. La aplicación de tales modelos sin validación local no es una práctica subóptima: es un riesgo clínico concreto que puede traducirse en errores de triaje, tratamiento inadecuado y daño al paciente.

Los grandes modelos de lenguaje, pese a sus capacidades impresionantes en el procesamiento del lenguaje natural médico, exhiben una serie de limitaciones que los clínicos deben comprender cabalmente: generan alucinaciones factológicas, no distinguen de manera confiable entre hipótesis diagnósticas probables e improbables, pueden amplificar el sesgo de confirmación, y su uso no supervisado en la formación médica plantea interrogantes legítimas sobre sus efectos en el desarrollo del razonamiento clínico crítico.

El camino hacia una medicina aumentada por IA, equitativa y segura, es exigente pero no imposible. Requiere validación local de los modelos, gobernanza participativa de los datos, marcos regulatorios vinculantes adaptados a las realidades regionales, y una cultura médica que mantenga el pensamiento crítico, la supervisión humana y la responsabilidad profesional como valores irrenunciables. Las recomendaciones de este capítulo apuntan en esa dirección.

14. PUNTOS CLAVE PARA LA PRÁCTICA CLÍNICA

Lo que todo medico debe saber de la Inteligencia Artificial:

- Los sistemas de IA en salud deben ser validados localmente antes de su implementación. Un modelo con excelente rendimiento en EE.UU. o Europa no garantiza el mismo rendimiento en Bolivia, Perú ni en ningún contexto fisiológica o epidemiológicamente diferente.

https://doi.org/10.62486/978-9915-9928-3-9.Ch21

- La SpO2 basal en poblaciones andinas (>2.500 m.s.n.m.) es normalmente de 87-93%, valor que los algoritmos de nivel del mar clasificarían como hipoxemia. Nunca aplique umbrales de SpO2 de nivel del mar sin considerar la altitud del paciente.
- Los LLMs generan información factualmente incorrecta (alucinaciones) de manera no predecible. Toda recomendación clínica derivada de un LLM debe ser verificada con fuentes primarias antes de su aplicación.
- El automationbias es un riesgo clínico documentado: los médicos pueden seguir recomendaciones incorrectas de sistemas de IA por efecto de la confianza en la automatización. Desarrolle el hábito de verificar activamente las respuestas de IA con su propio razonamiento clínico.
- Los sesgos algorítmicos en IA médica afectan desproporcionadamente a minorías raciales, étnicas y a poblaciones subrepresentadas. Sea especialmente crítico con las recomendaciones de IA en estos grupos.
- La IA es una herramienta de apoyo, no un sustituto del juicio clínico. La responsabilidad de las decisiones clínicas permanece siempre en el médico.
- Ante la implementación de cualquier herramienta de IA en su institución, exija información sobre el dataset de entrenamiento, la población de validación, las métricas de equidad y los mecanismos de supervisión.
- Fomente en sus estudiantes y residentes el hábito del razonamiento clínico activo antes de la consulta a herramientas de IA. La IA debe amplificar el pensamiento crítico, no reemplazarlo.

15. PREGUNTAS DE AUTOEVALUACIÓN

PREGUNTAS DE OPCIÓN MÚLTIPLE

1. Un algoritmo predictivo de sepsis basado en IA que utiliza los criterios clásicos del score SOFA se implementa en una Unidad de Cuidados Intensivos a 4,000 m s. n. m. Si el modelo no ha sido reajustado ni sometido a fine-tuning local, ¿cuál es el error sistemático esperado en el triaje de los pacientes basales locales?

a) Un sesgo de infratriaje, debido a que los valores de hemoglobina elevados enmascaran la anemia del paciente crítico.

b) Un sesgo de supertriaje (falsos positivos), debido a que la SpO2 basal de la población andina (87-93%) será interpretada de forma persistente como disfunción respiratoria aguda.

c) Una tasa de error del 0% debido a que los modelos autorregresivos compensan la altitud mediante aprendizaje no supervisado.

d) Un sesgo de confusión temporal secundario al aclaramiento acelerado de la creatinina inducido por la hipoxia crónica.

2. ¿Cuál de los siguientes fenómenos describe mejor la situación en la que un médico especialista revierte su propio diagnóstico correcto (v.g., en una mamografía o endoscopia) para alinearse con una sugerencia errónea emitida por un software de IA?

a) Deskilling clínico estructural.

b) Sesgo de confirmación inducido por RLHF.

c) Sesgo de automatización (*Automation Bias*).

d) Deuda cognitiva persistente.

3. Al evaluar el preprint del MIT Media Lab (Kosmyna et al., 2025) sobre el uso de ChatGPT y la “deuda cognitiva”, ¿cuál es una limitación metodológica interna crítica que impide extrapolar directamente sus resultados a la toma de decisiones en el pase de visita médico?

a) El uso de electroencefalografía (EEG) no está validado para medir actividad cerebral en humanos vivos.

b) El paradigma ecológico del estudio fue una tarea de escritura de ensayos tipo SAT, el cual difiere profundamente de las dinámicas complejas y multiestratificadas del razonamiento clínico en tiempo real.

c) Los datos fueron recolectados en su totalidad en hospitales andinos, limitando su validez en poblaciones urbanas de baja altitud.

d) El modelo utilizado fue exclusivamente de código abierto, omitiendo el comportamiento de los modelos comerciales de razonamiento.

https://doi.org/10.62486/978-9915-9928-3-9.Ch21

PREGUNTAS ABIERTAS DE RAZONAMIENTO CRÍTICO

4. Explique el concepto de "Sycophancy" (complacencia o adulación algorítmica) en los grandes modelos de lenguaje y describa cómo este fenómeno puede amplificar el sesgo de confirmación en el diagnóstico diferencial de un médico interno de pregrado.

5. Desde la perspectiva metodológica de la Medicina Basada en Evidencia (MBE), ¿por qué la técnica de "rebalanceo de clases" en un dataset de entrenamiento de nivel del mar es insuficiente para solucionar el sesgo que sufren las poblaciones andinas ante los algoritmos globales? ¿Qué alternativa técnica propone el texto?

RESPUESTAS:

1. b). Los modelos globales se entrenan predominantemente en datasets de nivel del mar. Al aplicar de forma cruda un umbral de corte estático de SpO2, el modelo comete un sesgo de medición ; confunde el fenotipo adaptativo normal del habitante de gran altitud con un estado patológico de hipoxemia refractaria , lo que dispara alertas innecesarias y sobrecarga el sistema con falsos positivos.

2: c). El *automationbias* es la propensión cognitiva a dar por válidas las respuestas de un sistema automatizado, disminuyendo la vigilancia heurística humana. La evidencia empírica (como el estudio de Dratsch et al. citado en el texto) demuestra que afecta incluso a especialistas experimentados, incrementando las tasas de error cuando el sistema falla.

3: b). Aunque el estudio ofrece señales de alerta sobre la delegación cognitiva pasiva, la validez externa del paradigma de escritura de ensayos universitarios es sumamente débil si se pretende aplicar como un modelo analógico de la heurística y la semiología médica. Además, la menor conectividad neural detectada por el EEG puede interpretarse de manera alternativa como optimización o eficiencia en el procesamiento cognitivo, una hipótesis que el diseño original no puede descartar.

4. Respuesta Esperada: El estudiante o clínico debe detallar que la *sycophancy* ocurre porque los LLMs son entrenados mediante Aprendizaje por Refuerzo con Retroalimentación Humana (RLHF), un proceso que prioriza la generación de respuestas que satisfagan o validen las premisas introducidas por el usuario para evitar la fricción discursiva. Si un interno interroga al modelo asumiendo erróneamente que un paciente con dolor torácico tiene una "pericarditis", el modelo tenderá a estructurar argumentos, e incluso a priorizar datos, que confirmen esa sospecha diagnóstica, omitiendo de forma complaciente las señales que apuntan a un infarto agudo de miocardio o a un síndrome aórtico agudo. Esto anula el pensamiento crítico y consolida el sesgo de confirmación del clínico.

5. Respuesta Esperada: Debe responderse que el mero rebalanceo de clases matemático (v.g., duplicar sintéticamente registros de personas catalogadas como "hispanas" o "latinoamericanas") no soluciona el problema porque la covariable crítica —la altitud geográfica y la hipoxia hipobárica crónica— no está registrada en los metadatos de los datasets de entrenamiento globales (como MIMIC o UK Biobank). La alternativa técnica sólida que propone el texto es la recolección de *Small Data* (datasets pequeños locales de alta fidelidad clínica) para ejecutar un **Fine-Tuning** (ajuste fino de capas profundas) sobre modelos médicos de código abierto, permitiendo recalibrar las fronteras de decisión estadística basándose en los "priors" de la fisiología de altura real de la región.

https://doi.org/10.62486/978-9915-9928-3-9.Ch21

REFERENCIAS

Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56. Available at: https://doi.org/10.1038/s41591-018-0300-7.
Chen F, Wang L, Hong J, Jiang J, Zhou L. Unmasking bias in artificial intelligence: A systematic review of bias detection and mitigation strategies in electronic health record-based models. J Am Med Inform Assoc. 2024;31(7):1511–20. Available at: https://doi.org/10.1093/jamia/ocae109.
Yang Y, Lin M, Zhao H, Peng Y, Huang F, Lu Z. A survey of recent methods for addressing AI fairness and bias in biomedicine. NPJ Digit Med. 2024;7:56. Available at: https://doi.org/10.1038/s41746-024-01051-8.
Kosmyna N, Hauptmann E, Yuan YT, et al. Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. arXiv. 2025:arXiv:2506.08872. Preprint [not peer-reviewed].
Cross S, Bhatta N, Okonkwo EC, et al. Effects of high altitude on respiratory rate and oxygen saturation reference values in healthy infants and children younger than 2 years in four countries: A cross-sectional study. Lancet Glob Health. 2021;9(5):e665–74. Available at: https://doi.org/10.1016/S2214-109X(21)00078-2.
Rojas-Camayo J, Mejia CR, Callacondo D, et al. Reference values for oxygen saturation by pulse oximetry in subjects at high altitude (Peruvian Andes). Thorax. 2018;73(8):776–8. Availableat: https://doi.org/10.1136/thoraxjnl-2017-211162.
Saavedra JM, Rivera-Chira M, Fuentes Ponce AI, et al. Individual variations and sex differences in hemodynamics with percutaneous arterial oxygen saturation (SpO2) in young Andean highlanders in Bolivia. BMC Res Notes. 2020;13(1):504. Availableat: https://doi.org/10.1186/s13104-020-05350-3.
Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447–53. Availableat: https://doi.org/10.1126/science.aax2342.
Sahni NR, Carrus B. Artificial intelligence in US health care delivery. N Engl J Med. 2023;389(4):348–58. Available at: https://doi.org/10.1056/NEJMra2204673.
Natali C, Marconi L, Dias Duran LD, Miglioretti M, Cabitza F. AI-induced deskilling in medicine: A mixed method literature review for setting a new research agenda. ArtifIntell Rev. 2025. Availableat: https://doi.org/10.1007/s10462-025-11352-1.
European Parliament and Council. Regulation (EU) 2024/1689 on artificial intelligence (AI Act). OfficialJournaloftheEuropean Union. 2024 Jun.
World Health Organization. WHO guidance on ethics and governance of artificial intelligence for health. Geneva: World Health Organization; 2021. ISBN 978-92-4-002-4229.
UNESCO. Recommendation on the ethics of artificial intelligence. Paris: UNESCO; 2021. Availableat: https://unesdoc.unesco.org/ark:/48223/pf0000381137.
OECD. Collective action for responsible AI in health. Paris: OECD; 2024 Jan. Available at: https://doi.org/10.1787/a6af1b86-en.
World Health Organization; International Telecommunication Union; World Intellectual Property Organization. Global Initiative on AI for Health (GI-AI4H): Strategic priorities. Geneva: WHO; 2023.
Dratsch T, Chen X, Rezazade Mehrizi M, et al. Automation bias in mammography: The impact of artificial intelligence BI-RADS suggestions on reader performance. Radiology. 2023;307(4):e222176. Availableat: https://doi.org/10.1148/radiol.222176.
United Nations General Assembly. Resolution A/RES/78/265: Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development. New York: UN; 2024 Mar.
Council of Europe. Framework Convention on artificial intelligence and human rights, democracy and the rule of law (CETS No. 225). Vilnius, 2024 Sep.
US Food and Drug Administration. Artificial intelligence and machine learning (AI/ML) software as a medical device (SaMD) action plan. Silver Spring (MD): FDA; 2021 Jan.
Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. Lancet Digit Health. 2019;1(6):e271–97. Available at: https://doi.org/10.1016/S2589-7500(19)30123-2.
Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023;620:172–80. Availableat: https://doi.org/10.1038/s41586-023-06291-2.
Si J, He Y, Yao J, et al. High-altitude cerebral oxygen saturation detection using wireless wearable cerebral oximeter. Front Neurol. 2024;15:1445563. Availableat: https://doi.org/10.3389/fneur.2024.1445563.
Roustan D, Bastardot F. The clinicians' guide to large language models: A general perspective with a focus on hallucinations. Interact J Med Res. 2025;14:e59823. Available at: https://doi.org/10.2196/59823.
Evaluating and addressing demographic disparities in medical large language models: A systematic review. medRxiv. 2024:v1 [preprint, under peer review].
Liao Y, Lu D, Yang J. Changes and monitoring technology of human heart rate and blood oxygen saturation under high-altitude hypoxia. Front Physiol. 2025;16:1642777. Availableat: https://doi.org/10.3389/fphys.2025.1642777.