

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 (automation bias)**, decide iniciar una reanimación hídrica agresiva con cristaloideos 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 automation bias 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

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

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 pacientes deben 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

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.

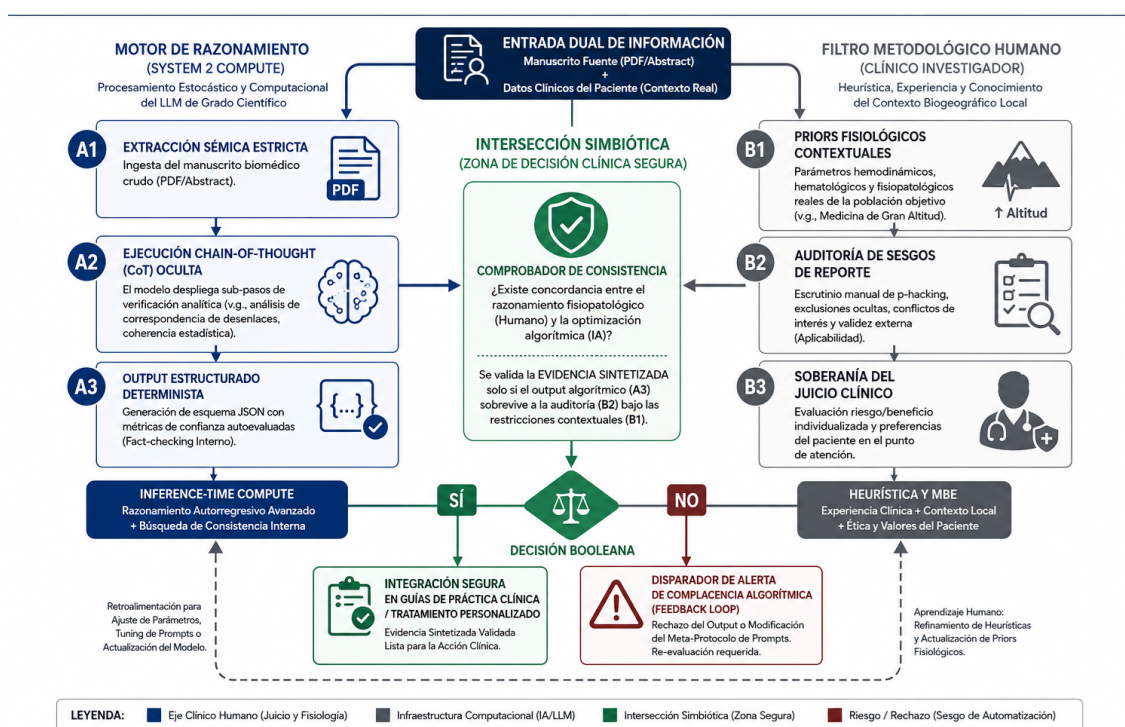
5. INTELIGENCIA ARTIFICIAL Y MEDICINA BASADA EN EVIDENCIA

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

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 LLMs entre 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 LLMs definido 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 confluencia 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.

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: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task’ (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

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 *Journal of the American Medical Informatics Association* (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.

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.

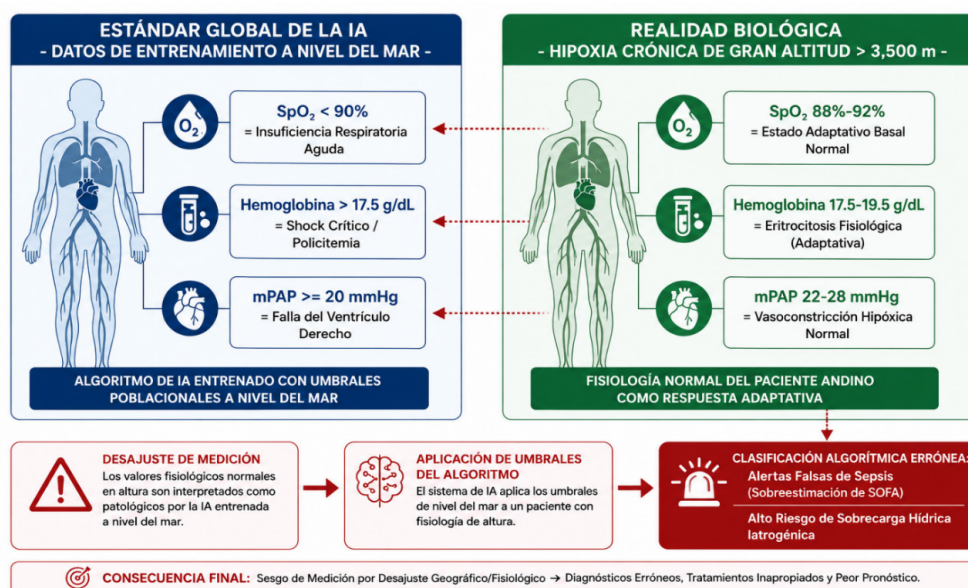
8.1 FISIOLÓGÍA DE ALTURA: BASE BIOLÓGICA DE LA DIFERENCIA

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.

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 SpO_2 , 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 SpO_2 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 SpO_2 y frecuencia respiratoria derivados de poblaciones de nivel del mar. Un paciente andino de 45 años con SpO_2 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 SpO_2 basal de 91% a 85% puede no activar los umbrales de alerta de un sistema diseñado para poblaciones con SpO_2 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 SpO_2 de 84% puede ser considerada 'normal para la altitud' por un algoritmo que incorpore correctivos de altitud mal calibrados, retrasando la intervención

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), eICU Collaborative Research Database (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,

sin alteraciones que hagan difícil su análisis, para realizar Ajuste fino (Fine Tunning) con grandes modelos de IA pre-entrenados y 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 automation bias, 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 automation bias definido 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 automation bias 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.

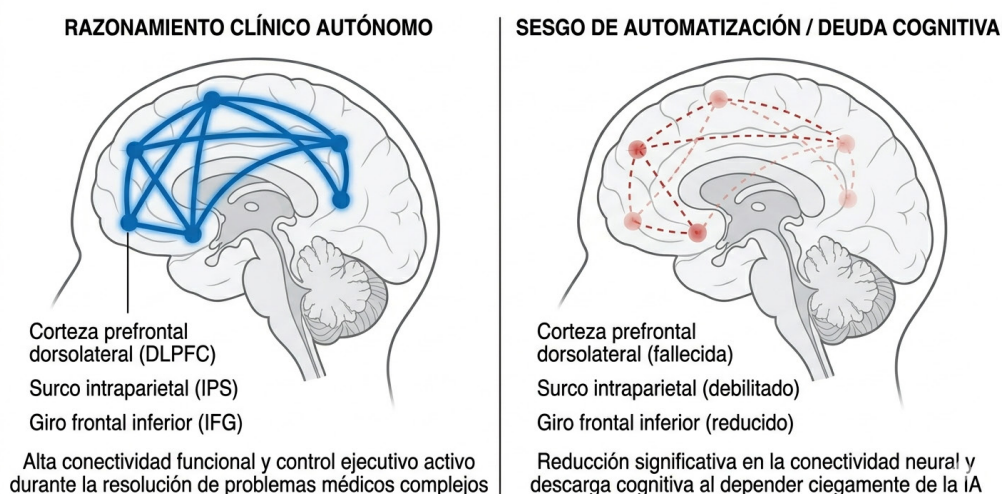
9.2 DESKILLING CLÍNICO

El deskilling es la 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 Intelligence Review 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.

COMPARACIÓN DE REDES CEREBRALES EN LA TOMA DE DECISIONES CLÍNICAS



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.

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 Learning Practice, 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 *Collective Action for Responsible 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.

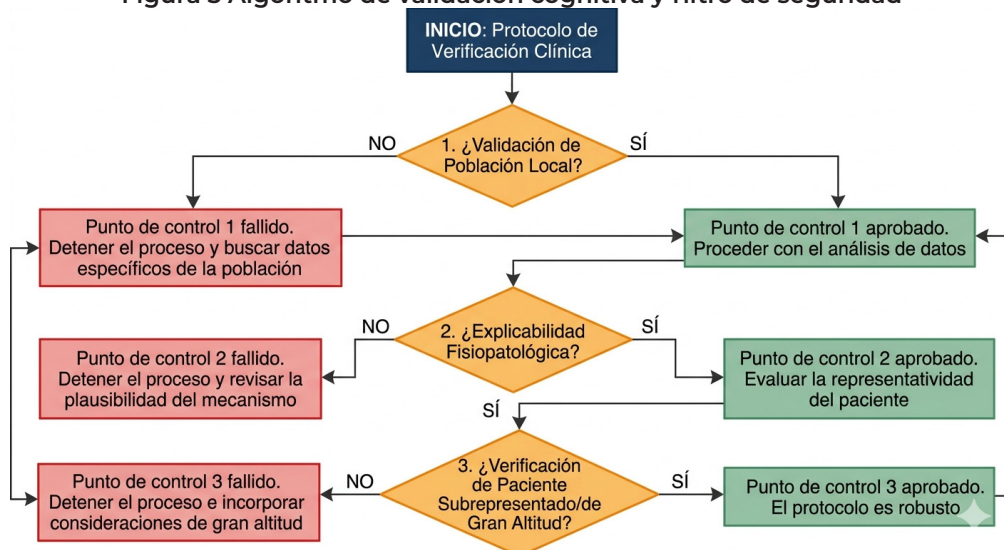
11. PRINCIPIOS PARA UNA INTEGRACIÓN ÉTICA Y SEGURA DE LA IA EN MEDICINA

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

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

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 automation bias. 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 legítimos 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 médico 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 fisiológico o epidemiológicamente diferente.

- 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 automation bias 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.

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 *automation bias* 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.

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](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. Available at: <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. Available at: <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. Available at: <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/NEJMr2204673>.
- 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. Available at: <https://doi.org/10.1007/s10462-025-11352-1>.
- European Parliament and Council. Regulation (EU) 2024/1689 on artificial intelligence (AI Act). *Official Journal of the European 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. Available at: <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. Available at: <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](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. Available at: <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. Available at: <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. Available at: <https://doi.org/10.3389/fphys.2025.1642777>.