

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 2x2 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.

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

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 resultados que 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)

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 2x2 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.

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 2x2, 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 crítica 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).

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 prevalencia 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”.

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).

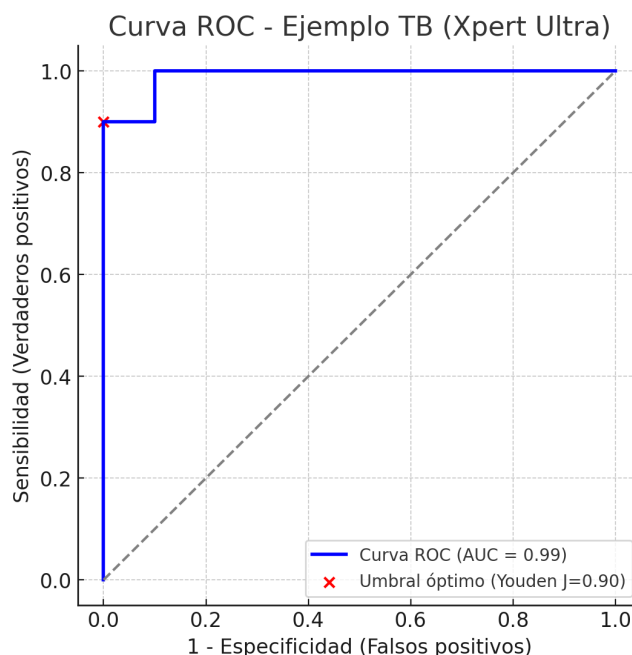
Medida	Definición	Fórmula	Resultado (caso)	Interpretación clínica
Sensibilidad	Proporción de enfermos identificados como positivos por la prueba índice	$VP / (VP + FN)$	90 % (90/100)	9 de cada 10 pacientes con TB serán detectados por Xpert. Importante para evitar falsos negativos.
Especificidad	Proporción 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	$\text{Sensibilidad} / (1 - \text{Especificidad})$	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	$(1 - \text{Sensibilidad}) / \text{Especificidad}$	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 Operating Characteristic) 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.

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 “óptimo” 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 DIAGNOSTIC ODDS RATIO (DOR)

En algunos trabajos aparece el diagnostic odds 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

- 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.

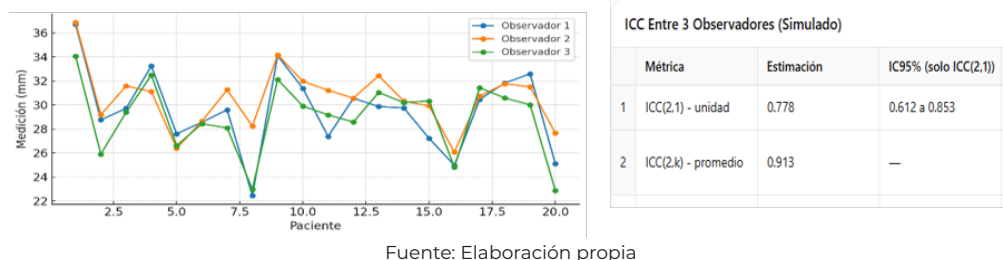
Para su interpretación:

- < 0,5 → acuerdo pobre.
- 0,5 – 0,75 → acuerdo moderado.

- 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



Fuente: Elaboración propia

Para su interpretación:

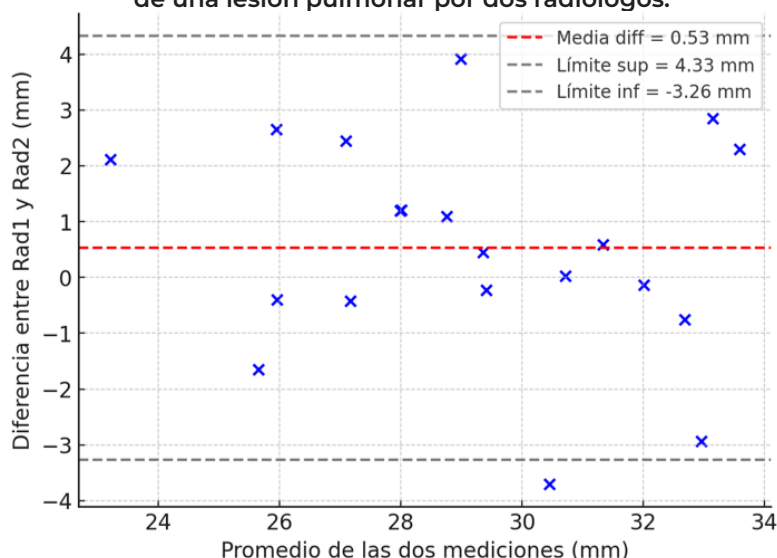
- ICC (2,1) ~0,78 → bueno: un lector individual produce mediciones razonablemente consistentes con otro.
- ICC (2,3) ~0,91 → excelente si se promedian las mediciones de tres observadores; útil para comités, doble lectura o auditoría.
- Si el uso rutinario depende de un único operador, el acuerdo que 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).

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 ($\approx 0,53$ 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 ($\pm 1,96$ 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 baciloscopia) 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.

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 excluirlas, 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”.

- **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 Infectious Diseases*) 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.

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.

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 2x2 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 $\approx 82\%$. Eso justifica decisiones inmediatas: aislamiento respiratorio e inicio de tratamiento sin esperar semanas al cultivo.
- Un Xpert negativo reduce la probabilidad de TB a $\approx 5\%$, 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.

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

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.

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](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](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. *Eur Urol 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, Juárez 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>.