# Interpretación de las pruebas diagnósticas en medicina clínica y hepatología

## *Interpretation of diagnostic tests in clinical medicine and hepatology*

*Gilberto J. Vizcaíno-Salazar*[1], *Mervin Villalobos-Orozco*[2]

**Resumen**

La elección de una prueba a realizar para un paciente, así como su interpretación, constituye una situación cotidiana a la cual el médico debe enfrentarse y para la cual debe aplicar su juicio crítico basado en las evidencias disponibles. En hepatología, la creciente disponibilidad de biomarcadores séricos, pruebas moleculares, métodos no invasivos para la evaluación de la fibrosis hepática y estudios de imagen ha incrementado la necesidad de interpretar adecuadamente el rendimiento diagnóstico de estas herramientas. Es común que, al describir una prueba de diagnóstico clínico o de laboratorio, se mencionen parámetros como la sensibilidad, la especificidad, los valores predictivos positivo y negativo y las razones de verosimilitud. Estos parámetros reflejan diferentes características del rendimiento de una prueba diagnóstica y permiten comprender el significado de sus resultados en un paciente en particular. Cuando se trata de comparar estos parámetros en diferentes pruebas y optar por la que puede ofrecer mayor utilidad para el diagnóstico de una enfermedad determinada, es indispensable que el médico conozca cómo se obtienen estas medidas y cuál es su interpretación para decidir la conducta más apropiada para el paciente. El objetivo de esta revisión es proporcionar al hepatólogo y al especialista del laboratorio clínico herramientas para interpretar correctamente el rendimiento diagnóstico de las pruebas utilizadas en hepatología, utilizando ejemplos de aplicación clínica.

**Palabras clave:** sensibilidad y especificidad, valores predictivos, razones de verosimilitud, curvas ROC, técnicas y procedimientos diagnósticos, enfermedades hepáticas.

**Abstract**

Selecting a diagnostic test for a patient, as well as interpreting its results, is a common situation that physicians face in daily clinical practice and requires critical judgment based on the available evidence. In hepatology, the increasing availability of serum biomarkers, molecular assays, noninvasive

[1] Médico, Especialista en Hematología, PhD en Ciencias Médicas. Profesor Titular Emérito e Investigador adscrito al Instituto de Investigaciones Clínicas "Dr. Américo Negrette", Facultad de Medicina, Universidad del Zulia. Maracaibo, Venezuela. Email: gilvizcaino@gmail.com.

[2] Médico, Especialista en Medicina Interna, Especialista en Gastroenterología, MSc en Gerencia de Proyectos de Investigación y Desarrollo. Gastroestudio Unidad Videoendoscópica, Intergastro. Medellín, Colombia.

Conflicto de intereses: los autores declaran que no tienen conflicto de intereses.
Hepatología 2026;7:240-265. https://doi.org/10.59093/27112330.180.
Asociación Colombiana de Hepatología, 2026.
Recibido el 16 de junio de 2026; aceptado el 3 de agosto de 2026.

methods for liver fibrosis assessment, and imaging techniques has increased the need to appropriately interpret the diagnostic performance of these tools. When describing a clinical or laboratory diagnostic test, parameters such as sensitivity, specificity, positive and negative predictive values, and likelihood ratios are commonly reported. These parameters reflect different characteristics of the performance of a diagnostic test and allow its usefulness to be assessed in different clinical scenarios and the meaning of its results in an individual patient to be understood. When comparing these parameters among different tests and choosing the one that may offer greater usefulness for the diagnosis of a particular disease, physicians need to understand how these measures are obtained and how they should be interpreted in order to make the most appropriate clinical decision. The aim of this review is to provide hepatologists and clinical laboratory specialists with tools to correctly interpret the diagnostic performance of tests used in hepatology, using examples of clinical application.

**Keywords:** sensitivity and specificity, predictive values, likelihood ratios, receiver operating characteristic curves, diagnostic techniques and procedures, liver diseases.

## Introducción

Como sucede con los elementos de una historia clínica y del examen físico, cada procedimiento de diagnóstico clínico o prueba de laboratorio reúne una serie de características que reflejan la información esperada acerca de un paciente del cual el médico desea saber si posee o no determinada enfermedad. En el presente manuscrito analizaremos varios puntos necesarios para comprender la importancia y aplicabilidad del cálculo de la sensibilidad, la especificidad y otros parámetros estadísticos en la práctica clínica diaria de la hepatología para determinar: a) qué tan exactas pueden ser las pruebas de diagnóstico clínico y de laboratorio para clasificar a un sujeto como sano o enfermo, de acuerdo con su estado de salud real; es decir, ¿cuál es el desempeño diagnóstico de la prueba (sensibilidad y especificidad)?, b) cuál es la confiabilidad o reproducibilidad de los resultados de una prueba, por ejemplo, al ser nuevamente aplicada por otro sujeto o por el mismo sujeto y c) cómo verificar qué tan de acuerdo están dos observadores o dos métodos frente a un mismo fenómeno.

En esta revisión se describen los principales parámetros utilizados para interpretar el rendimiento de las pruebas diagnósticas en la práctica clínica, entre ellos la sensibilidad, la especificidad, los valores predictivos, las razones de verosimilitud y las probabilidades pre y posprueba. También se revisan los métodos para su cálculo, su interpretación clínica y sus principales limitaciones. Estos conceptos se ilustran con ejemplos relacionados con enfermedades hepáticas y pruebas diagnósticas de uso frecuente en hepatología, con el propósito de facilitar su aplicación en la práctica clínica y en la toma de decisiones basada en la evidencia.

En la última década el desarrollo de biomarcadores séricos, pruebas moleculares, elastografía hepática y modelos no invasivos para fibrosis ha ampliado las posibilidades para el diagnóstico y la estratificación de las enfermedades hepáticas, haciendo indispensable que el clínico comprenda las características operativas de las pruebas diagnósticas antes de incorporarlas a la práctica clínica [1,2]. Esta evolución ha permitido reducir la necesidad de procedimientos invasivos como la biopsia hepática en múltiples escenarios clínicos y ha incrementado la importancia de interpretar correctamente medidas como la sensibilidad, la especificidad, las razones de

verosimilitud y el área bajo la curva ROC. El objetivo de esta revisión es proporcionar al hepatólogo y al especialista del laboratorio clínico las herramientas para interpretar correctamente el rendimiento diagnóstico de las pruebas utilizadas en hepatología, utilizando ejemplos basados en la evidencia más reciente.

## Definición de términos

A partir de la necesidad de conocer si un paciente realmente posee o no determinada enfermedad se establecen dos preguntas fundamentales: a) si la enfermedad está presente ¿cuál es la probabilidad de que el resultado de la prueba diagnóstica sea positivo? y b) si la enfermedad no está presente ¿cuál es la probabilidad de que el resultado sea negativo? La primera pregunta define lo que se conoce como sensibilidad de una prueba diagnóstica y la segunda incluye el concepto de especificidad de una prueba diagnóstica. La definición precisa de estos conceptos sería [3]:

- Sensibilidad: proporción de individuos enfermos que poseen una prueba positiva.

- Especificidad: proporción de individuos sin la enfermedad que poseen una prueba negativa.

De estos dos conceptos se desprenden, a su vez, dos reglas nemotécnicas con los siguientes acrónimos:

1. SENDES: cuando un signo, síntoma o prueba diagnóstica posee una alta **SE**nsibilidad, un resultado **N**egativo **DES**carta el diagnóstico, especialmente cuando la probabilidad preprueba es baja o intermedia. Por ejemplo, la detección del ARN del virus de la hepatitis C mediante técnicas de amplificación de ácidos nucleicos posee una alta sensibilidad para identificar infección activa. En consecuencia, si un paciente con sospecha clínica de hepatitis C presenta un resultado negativo para ARN del VHC, es altamente improbable que tenga una infección activa, siempre que la prueba se haya realizado en el momento clínico apropiado y no existan factores que comprometan su validez [3,4]. Otro ejemplo corresponde a la elastografía hepática transitoria para la evaluación de la fibrosis hepática. Cuando se emplean los puntos de corte recomendados para el contexto clínico y la etiología de la enfermedad, esta prueba puede presentar una elevada sensibilidad para descartar enfermedad hepática crónica avanzada compensada; por ello, un resultado de rigidez hepática bajo puede hacer poco probable la presencia de fibrosis avanzada o cirrosis y, en muchos pacientes, evita la necesidad de procedimientos diagnósticos invasivos como la biopsia hepática [1].

2. ESPIN: cuando un signo o prueba posee una alta **ES**pecificidad, un resultado **P**ositivo **IN**dica o confirma el diagnóstico, especialmente cuando la probabilidad preprueba no es baja. Por ejemplo, los anticuerpos antimitocondriales (AMA) presentan una alta especificidad para el diagnóstico de colangitis biliar primaria. En consecuencia, un resultado positivo para AMA en un paciente con elevación persistente de la fosfatasa alcalina aumenta considerablemente la probabilidad diagnóstica de esta enfermedad [5].

En razón a lo anteriormente expuesto, el conocimiento de las características de una prueba determinada no permite, por sí solo, tener una interpretación exacta del resultado. La sensibilidad y la especificidad describen el comportamiento de una prueba en individuos con y sin la enfermedad, pero no permiten establecer directamente cuál es la probabilidad de que un paciente tenga la enfermedad después de obtener un resultado positivo o negativo [6]. En este sentido, el interés del médico es determinar la presencia o ausencia de la

enfermedad. En consecuencia, surgen las siguientes preguntas: a) si el resultado es positivo ¿cuál es la probabilidad de que la enfermedad esté presente? y b) si el resultado es negativo ¿cuál es la probabilidad de que la enfermedad no esté presente? La primera pregunta refleja lo que se conoce como valor predictivo positivo (VPP) de una prueba diagnóstica y la segunda está relacionada con lo que se denomina valor predictivo negativo (VPN) del resultado de una prueba diagnóstica.

En hepatología, estos conceptos son especialmente relevantes durante la evaluación de pacientes con sospecha de fibrosis hepática avanzada, hepatitis virales o carcinoma hepatocelular, ya que el significado clínico de un resultado positivo o negativo depende no solo del rendimiento de la prueba diagnóstica, sino también de la probabilidad preprueba y de la frecuencia de la enfermedad en la población evaluada. La definición precisa de estos conceptos sería [7]:

- Valor predictivo positivo (VPP): proporción de individuos con una prueba positiva que presentan la enfermedad.

- Valor predictivo negativo (VPN): proporción de individuos con una prueba negativa que no presentan la enfermedad.

De acuerdo con esto se puede inferir que la sensibilidad y la especificidad son medidas de la exactitud diagnóstica de una prueba, mientras que el valor predictivo positivo y el valor predictivo negativo expresan la probabilidad de que un resultado positivo o negativo refleje correctamente el estado real del paciente. Sin embargo, los valores predictivos dependen de la frecuencia de la enfermedad o de la probabilidad preprueba en la población o en el paciente evaluado, por lo que pueden variar cuando una misma prueba se aplica en diferentes escenarios clínicos [7,8,9].

## Cálculo de sensibilidad, especificidad, valor predictivo positivo y valor predictivo negativo

La determinación de los parámetros de sensibilidad, especificidad, valor predictivo positivo y valor predictivo negativo se realiza construyendo, en primer lugar, una tabla binaria o de 2x2, a partir de la cual se obtienen los diversos resultados probables de una prueba diagnóstica, como se esquematiza en la **Tabla 1** [3,6,7,9].

Los posibles resultados de una prueba diagnóstica se interpretan de la manera siguiente [3,6,7,9]:

- Verdaderos positivos: la enfermedad está presente y la prueba identifica correctamente al paciente como positivo.

- Falsos positivos: la enfermedad no está presente y la prueba identifica al paciente como positivo.

- Verdaderos negativos: la enfermedad no está presente y la prueba identifica correctamente al paciente como negativo.

- Falsos negativos: la enfermedad está presente y la prueba identifica al paciente como negativo.

Aunque la estructura de la tabla de 2×2 es una herramienta general para la evaluación de cualquier prueba diagnóstica con resultados binarios, su aplicación resulta particularmente útil en hepatología, donde permite interpretar el rendimiento de pruebas no invasivas como el índice FIB-4, el APRI, la elastografía hepática transitoria, la resonancia magnética y los biomarcadores utilizados para el diagnóstico de fibrosis hepática, la enfermedad hepática crónica avanzada y el carcinoma hepatocelular. Posteriormente, para el cálculo de los parámetros anteriormente mencionados, se toman los datos registrados en la tabla de

**Tabla 1.** Tabla de 2x2 de una prueba diagnóstica y resultados probables de la misma.

| | Enfermedad presente | Enfermedad ausente |
|---|---|---|
| Prueba positiva | Verdaderos positivos (VP)<br>a | Falsos positivos (FP)<br>b |
| Prueba negativa | Falsos negativos (FN)<br>c | Verdaderos negativos (VN)<br>d |

2x2 y se aplican las fórmulas presentadas en la **tabla 2** [3,6,7,9].

Como regla nemotécnica, las fórmulas expresadas anteriormente se pueden resumir en la representación gráfica de la **figura 1**.

Una de las ventajas de los parámetros de sensibilidad y especificidad es que al ser calculados en forma «vertical», dentro de la matriz de decisiones (tabla 2x2), no dependen de la prevalencia de la enfermedad, como sí ocurre con los valores predictivos positivos y negativos, que son calculados en forma horizontal (**figura 1**) [3,6,7,9].

Los valores predictivos son de gran utilidad para interpretar los resultados de una prueba en la práctica clínica, pero dependen de la probabilidad preprueba o de la frecuencia de la enfermedad en la población evaluada. De esta manera, cuando la probabilidad preprueba es baja, un resultado negativo suele tener un alto valor predictivo negativo, mientras que un resultado positivo puede presentar un valor predictivo positivo menor. Por el contrario, cuando la probabilidad preprueba es alta, un resultado positivo puede tener un mayor valor predictivo positivo, mientras que un resultado negativo puede no ser suficiente para excluir la enfermedad.

Este concepto tiene especial importancia en hepatología. Por ejemplo, el valor predictivo positivo de la ecografía para la detección de carcinoma hepatocelular puede ser mayor en pacientes con cirrosis sometidos a programas de vigilancia que en la población general, debido a la

**Tabla 2.** Fórmulas para el cálculo de los parámetros de una prueba diagnóstica.

| Parámetro | Fórmula |
|---|---|
| Sensibilidad | a / (a + c) = VP / (VP + FN) |
| Especificidad | d / (b + d) = VN / (FP + VN) |
| Valor predictivo positivo | a / (a + b) = VP / (VP + FP) |
| Valor predictivo negativo | d / (c + d) = FN / (FN + VN) |
| Porcentaje de falsos negativos | 100 – sensibilidad |
| Porcentaje de falsos positivos | 100 – especificidad |
| Exactitud | (a + d) / n = (VP + VN) / n |

VP: verdaderos positivos; VN: verdaderos negativos; FP: falsos positivos; FN: falsos negativos; n: número total de individuos evaluados.

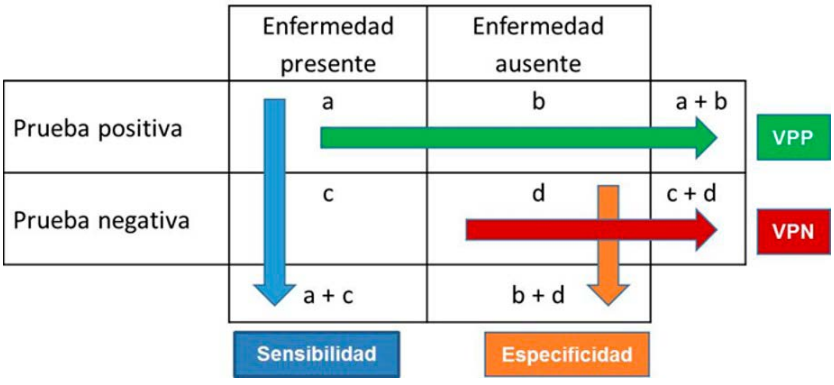


**Figura 1.** Representación esquemática direccional de las fórmulas para calcular la sensibilidad, la especificidad, el valor predictivo positivo (VPP) y el valor predictivo negativo (VPN). Nótese que cada flecha comienza con la letra dividendo y termina con el divisor (p. ej. a / (a + c) = sensibilidad).

mayor prevalencia de la enfermedad en el primer grupo [10]. Del mismo modo, el valor predictivo negativo de pruebas no invasivas como el índice FIB-4 puede variar según la población evaluada, incluso cuando la sensibilidad y la especificidad permanezcan relativamente constantes [1]. Esta diferencia resalta la importancia de interpretar los resultados diagnósticos siempre dentro del contexto clínico y epidemiológico del paciente.

## Ejemplos para el cálculo de los parámetros de una prueba diagnóstica

En un estudio incluido en un metaanálisis reciente [11], el índice FIB-4 fue comparado con la biopsia hepática para el diagnóstico de fibrosis hepática avanzada en pacientes con enfermedad hepática esteatósica asociada a disfunción metabólica (MASLD). La biopsia hepática fue considerada el estándar de referencia para establecer la presencia de fibrosis avanzada. En este estudio se evaluaron 551 pacientes, de los cuales 129 presentaron fibrosis avanzada según el estudio histopatológico. El índice FIB-4 identificó correctamente 107 de estos pacientes, mientras que 22 tuvieron un resultado falsamente negativo. Entre los 422 pacientes sin fibrosis avanzada, 337 fueron clasificados correctamente y 85 presentaron un resultado falsamente positivo. A partir de estos datos puede construirse la tabla de 2×2 (**tabla 3**), a partir de la cual se calculan la sensibilidad, la especificidad, el valor predictivo positivo y el valor predictivo negativo.

Ahora, analicemos esquemáticamente este ejemplo consignando los datos en una tabla de 2x2 (**tabla 3**). A partir de estos datos calculamos los parámetros de la prueba diagnóstica aplicando las fórmulas consignadas en la **tabla 2**.

De acuerdo con estos resultados, el índice FIB-4 presentó una sensibilidad del 82,9 %, por lo que un resultado negativo disminuye la probabilidad de fibrosis hepática avanzada; sin embargo, no permite descartarla por completo, ya que cerca del 17,1 % de los pacientes presentaron resultados falsamente negativos. Por tanto, en este ejemplo no se cumple el acrónimo **SENDES**. Asimismo, la especificidad fue del 79,9 %,

**Tabla 3.** Tabla de 2×2 para el cálculo de los parámetros diagnósticos del índice FIB-4.

| Prueba diagnóstica | Fibrosis avanzada presente | Fibrosis avanzada ausente | Total |
|---|---|---|---|
| FIB-4 positivo | 107 (a) | 85 (b) | 192 (a+b) |
| FIB-4 negativo | 22 (c) | 337 (d) | 359 (c+d) |
| Total | 129 (a+c) | 422 (b+d) | 551 (a+b+c+d) |

Sensibilidad: 107 / (107 + 22) = 82,9 %
Especificidad: 337 / (337 + 85) = 79,9 %
VPP: 107 / (107 + 85) = 55,7 %
VPN: 337 / (337 + 22) = 93,9 %

indicando que un resultado positivo aumenta la probabilidad de fibrosis avanzada, pero no confirma el diagnóstico, debido a la presencia de falsos positivos; en consecuencia, tampoco se cumple el acrónimo **ESPIN**. El valor predictivo positivo fue del 55,7 %, mientras que el valor predictivo negativo alcanzó el 93,9 %, lo que respalda el uso del índice FIB-4 como prueba inicial para descartar fibrosis hepática avanzada y seleccionar los pacientes que requieren estudios complementarios.

## Examinemos otro ejemplo:

En un estudio de precisión diagnóstica realizado por Bandin y colaboradores [12], se evaluó el rendimiento diagnóstico de los anticuerpos anti-gp210 mediante ELISA para el diagnóstico de la colangitis biliar primaria, utilizando como referencia el diagnóstico de esta enfermedad establecido según los criterios utilizados en el estudio. Se estudiaron 782 individuos, de los cuales 285 presentaban colangitis biliar primaria y 497 correspondían a controles con otras enfermedades hepáticas, enfermedades autoinmunes o individuos sanos. La prueba fue positiva en 73 pacientes con CBP y en 2 sujetos sin la enfermedad. La representación esquemática de estos resultados se muestra en la **tabla 4**, a partir de la cual pueden calcularse la sensibilidad, la especificidad y los valores predictivos. Nuevamente, a partir de los datos de la tabla de 2x2 (**tabla 4**) calculamos los parámetros de la prueba diagnóstica aplicando las fórmulas consignadas en la **tabla 2**.

De acuerdo con estos resultados, la determinación de los anticuerpos anti-gp210 presentó una baja sensibilidad (25,6 %), por lo que un resultado negativo no permite descartar la colangitis biliar primaria; en consecuencia, no se cumple el acrónimo **SENDES**. En contraste, la especificidad fue del 99,6 %, indicando que un resultado positivo aumenta de manera importante la probabilidad de presentar la enfermedad, por lo que sí se cumple el acrónimo **ESPIN**. Asimismo, el valor predictivo positivo fue del 97,3 %, lo que demuestra que un resultado positivo de la prueba es altamente sugestivo de colangitis biliar primaria. Por el contrario, el valor predictivo negativo fue del 70,0 %, indicando que un resultado negativo no excluye la enfermedad. Estos hallazgos sugieren que los anticuerpos anti-gp210 pueden ser útiles como prueba complementaria para apoyar el

**Tabla 4.** Tabla 2x2 para la evaluación del rendimiento diagnóstico de los anticuerpos anti-gp210 en la colangitis biliar primaria.

| Prueba diagnóstica | Colangitis biliar primaria presente | Colangitis biliar primaria ausente | Total |
|---|---|---|---|
| Anti-gp210 positivo | 73 (a) | 2 (b) | 75 |
| Anti-gp210 negativo | 212 (c) | 495 (d) | 707 |
| Total | 285 | 497 | 782 |

Sensibilidad: 73 / (73 + 212) = 25,6 %
Especificidad: 495 / (495 + 2) = 99,6 %
Valor predictivo positivo: 73 / (73 + 2) = 97,3 %
Valor predictivo negativo: 495 / (495 + 212) = 70,0 %

diagnóstico de colangitis biliar primaria, especialmente en determinados pacientes con sospecha clínica de la enfermedad y resultados negativos para anticuerpos antimitocondriales.

## Curva ROC en términos de la sensibilidad y la especificidad

La curva ROC (del inglés, *Receiver Operating Characteristic*), o curva característica de funcionamiento del receptor, fue desarrollada por ingenieros eléctricos para medir la eficacia en la detección de objetos enemigos en campos de batalla mediante pantallas de radar, a partir de lo cual se elaboró la teoría de detección de señales (TDS). El análisis ROC se aplicó posteriormente en medicina, radiología, psicología y otras áreas durante varias décadas. En la curva ROC se presenta la sensibilidad de una prueba diagnóstica, que produce resultados continuos en función de la tasa de falsos positivos (1 – especificidad) para distintos puntos de corte [13].

Por su parte, la separación de los grupos, con y sin enfermedad, representa la capacidad discriminatoria de una prueba para diferenciar a los individuos enfermos de aquellos sin la enfermedad. Un parámetro para evaluar la capacidad discriminativa de una prueba diagnóstica que produce resultados continuos es el área bajo la curva (AUC, del inglés, *Area Under the Curve*), la cual puede interpretarse como la probabilidad de que, al seleccionar aleatoriamente un individuo enfermo y uno sin la enfermedad, la prueba asigne un resultado de mayor valor al individuo enfermo que al individuo sin la enfermedad [13].

La realidad nos indica que una amplia gama de pruebas diagnósticas reporta sus resultados cuantitativamente utilizando escalas continuas. En hepatología son ejemplos el índice FIB-4, el APRI, la rigidez hepática medida mediante elastografía transitoria, la concentración sérica de alfafetoproteína (AFP) y diversos biomarcadores séricos de fibrosis hepática. En estas pruebas, para establecer o estratificar el riesgo de una determinada enfermedad, se establece un punto de corte a partir del cual el resultado puede apoyar o hacer menos probable un diagnóstico, o indicar la necesidad de realizar estudios complementarios [13–15]. Por ejemplo, en la evaluación no invasiva de la fibrosis hepática, distintos puntos de corte del índice FIB-4 permiten clasificar a

los pacientes en categorías de bajo, intermedio y alto riesgo de fibrosis avanzada. De igual forma, diferentes valores de rigidez hepática obtenidos mediante elastografía transitoria modifican la sensibilidad y la especificidad para el diagnóstico de fibrosis significativa o cirrosis.

El análisis con base en las curvas ROC constituye un método estadístico para evaluar la capacidad discriminativa de las pruebas que utilizan escalas continuas con tres propósitos específicos: a) evaluar diferentes puntos de corte y seleccionar aquel que resulte más apropiado según el objetivo clínico, b) evaluar la capacidad discriminativa de la prueba diagnóstica, es decir, su capacidad para diferenciar sujetos con y sin la enfermedad y c) comparar la capacidad discriminativa de dos o más pruebas diagnósticas que expresan sus resultados como escalas continuas [14–16].

Los ejes del gráfico de curva ROC adoptan valores entre 0 y 1 (0 % y 100 %), lo que delimita un cuadrado de área igual a 1,00. Una prueba diagnóstica se considera no discriminativa cuando su curva ROC coincide con la línea de no discriminación, la cual posee un área bajo la curva de 0,50 (**figura 2**). A medida que el AUC de una prueba diagnóstica se acerca a 1,00, mayor será su capacidad discriminativa, siendo máxima cuando el AUC es igual a 1,00 [14]. En hepatología, las curvas ROC se utilizan con frecuencia para evaluar el rendimiento diagnóstico de pruebas no invasivas destinadas a identificar fibrosis hepática avanzada, hipertensión portal clínicamente significativa y carcinoma hepatocelular. El área bajo la curva permite comparar la capacidad discriminativa de diferentes biomarcadores o modelos diagnósticos.

El punto de corte de una escala continua que maximiza conjuntamente la sensibilidad y la especificidad puede seleccionarse mediante el índice de Youden, calculado según la fórmula J = sensibilidad - (1 – especificidad), o de manera equivalente, J = sensibilidad + especificidad – 1. Gráficamente, el punto de corte óptimo suele localizarse en la región de la curva ROC más cercana al ángulo superior izquierdo del gráfico, aunque ambos criterios no son necesariamente equivalentes en todas las situaciones (**figura 2**). El índice de Youden identifica el punto de corte que maximiza conjuntamente la sensibilidad y la especificidad para un mismo umbral; sin embargo, dicho punto de corte no necesariamente determina la sensibilidad ni la especificidad más altas que podría alcanzar la prueba por separado, ya que, generalmente, la sensibilidad más alta se obtiene con un punto de corte, mientras que la especificidad más alta se obtiene con otro [14]. Por ejemplo, durante la validación de nuevos biomarcadores séricos de fibrosis hepática o de algoritmos para la detección de carcinoma hepatocelular, el índice de Youden suele emplearse para seleccionar el punto de corte que proporciona el mejor equilibrio entre sensibilidad y especificidad.

Para comprender mejor el concepto de discriminación es más simple pensar que el eje Y del gráfico de la curva ROC corresponde a la proporción de verdaderos positivos sobre el total de pacientes enfermos, es decir, la sensibilidad, y que el eje X corresponde a la proporción de falsos positivos sobre el total de sujetos sin la enfermedad, es decir, 1 - especificidad. Visto de esta manera, un gráfico de curva ROC ilustra la «proporción de verdaderos positivos» (eje Y) versus la «proporción de falsos positivos» (eje X) para cada punto de corte de una prueba diagnóstica cuya escala de medición es continua [13].

Para una construcción e interpretación correcta de los gráficos de curvas ROC es necesario tener en cuenta tres conceptos finales. En primer lugar, la sensibilidad, la

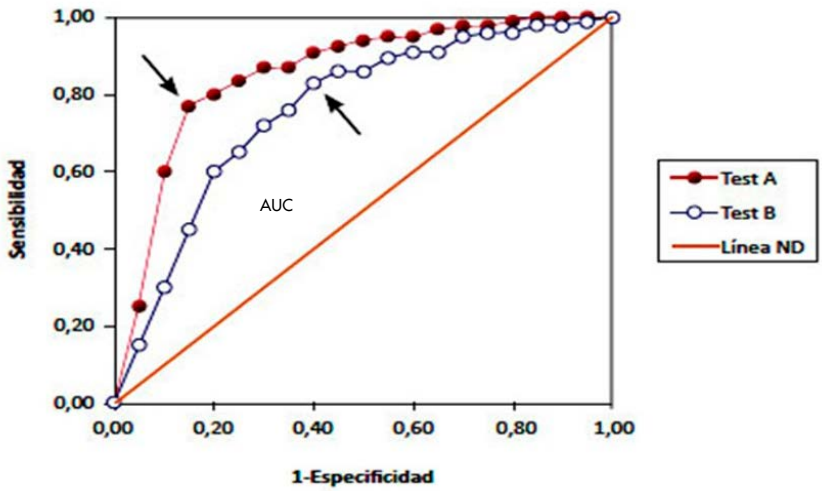


Figura 2. Representación gráfica de curvas ROC de dos pruebas diagnósticas hipotéticas (A y B). Para cada curva ROC, las flechas indican el punto de corte que maximiza conjuntamente la sensibilidad y la especificidad, de acuerdo con el índice de Youden. AUC denota el área bajo la curva. La línea recta, en naranja, representa la línea de no discriminación (línea ND), la cual divide en dos mitades iguales el cuadrado de área de 1,00, por lo que su AUC es igual a 0,50.

especificidad y el área bajo la curva son estimadores muestrales de parámetros poblacionales; por consiguiente, cada uno tiene asociado un error de estimación, lo que hace necesario reportar sus respectivos intervalos de confianza (IC). En segunda instancia, los estudios de exactitud diagnóstica, a partir de los cuales se construyen las curvas ROC pueden utilizar diferentes diseños, aunque los diseños transversales con inclusión consecutiva de pacientes con sospecha de la enfermedad suelen ser apropiados para evaluar el rendimiento de una prueba diagnóstica. Los diseños de casos y controles deben interpretarse con cautela, debido al riesgo de sesgo de espectro y a la posible sobreestimación del rendimiento diagnóstico. Finalmente, el riesgo de sesgo y la aplicabilidad de los estudios de exactitud diagnóstica deben evaluarse críticamente, de acuerdo con las recomendaciones de la medicina basada en la evidencia [17]. Actualmente, el riesgo de sesgo y la aplicabilidad de los estudios de exactitud diagnóstica se recomienda evaluarlos mediante la herramienta QUADAS-2 [18], mientras que la certeza global de la evidencia puede valorarse utilizando el enfoque GRADE adaptado para pruebas diagnósticas [19].

De esta manera, las curvas ROC son útiles para:

1. Conocer el rendimiento discriminativo global de una prueba (área bajo la curva o AUC).

2. Comparar dos o más pruebas o diferentes puntos de corte de una misma prueba: comparación de dos curvas o de dos puntos sobre una curva.

3. Ayudar a seleccionar un punto de corte apropiado para una determinada población o contexto clínico.

Entre las limitaciones de su uso se encuentra que las curvas ROC convencionales están diseñadas para evaluar la capacidad de una prueba para discriminar entre dos condiciones, generalmente la presencia o ausencia de una enfermedad.

Para ilustrar este concepto, consideremos un ejemplo hipotético en el que se comparan tres pruebas mediante curvas ROC. Considérense 100 resultados positivos y 100 resultados negativos, presentados en la figura 3. En la figura 4 se representan, en el espacio ROC, los puntos correspondientes a los tres métodos. El método A muestra un mejor rendimiento discriminativo que los métodos B y C, mientras que el método B se encuentra sobre la línea de no discriminación (diagonal). Si aplicamos la fórmula para calcular la exactitud, esta es de 73,5 %, 50,0 % y 18,0 % para los métodos A, B y C, respectivamente. El método A se sitúa como el de mejor desempeño, mientras que el método C muestra un rendimiento muy pobre.

Dado que el interés de esta revisión es presentar al profesional de la salud una metodología simple, sencilla y de fácil comprensión al lector, no se profundizará en los cálculos matemáticos para las curvas ROC, los cuales ameritan una mayor explicación. En este caso, se sugiere consultar otras referencias que abordan el tema con mayor profundidad [20–22].

## Razón de verosimilitud

La razón de verosimilitud expresa cuánto más probable es obtener un determinado resultado de una prueba diagnóstica en un paciente con una enfermedad que en un paciente sin dicha enfermedad. La razón de verosimilitud positiva (LR+) indica cuántas veces es más probable obtener un resultado positivo en un paciente con la enfermedad que en un paciente sin ella, mientras que la razón de verosimilitud negativa (LR–) expresa cuántas veces es más probable obtener un resultado negativo en un paciente con la enfermedad que en un paciente sin ella [7]. Por ejemplo, en un paciente con sospecha de colangitis biliar primaria (CBP) y un resultado positivo para AMA, si un estudio reporta una sensibilidad del 94,7 % y una especificidad del 99,0 %, la probabilidad de obtener un resultado positivo en un paciente con CBP es mucho mayor que en un individuo sin la enfermedad. En este caso, la razón de verosimilitud positiva es 94,7/(100–99,0)=94,7, lo que indica que un resultado positivo de AMA es aproximadamente 95 veces más probable en un paciente con colangitis biliar primaria que en un individuo sin esta enfermedad.

### Cálculo de la razón de verosimilitud positiva y negativa

*Razón de verosimilitud positiva*

LR+ = sensibilidad/(1-especificidad)

La razón de verosimilitud positiva se calcula dividiendo la probabilidad de un resultado positivo en los pacientes con la enfermedad entre la probabilidad de obtener un resultado positivo en los pacientes sin la enfermedad. Es, en definitiva, el cociente entre la fracción de verdaderos positivos (sensibilidad) y la fracción de falsos positivos (1-especificidad) [3].

*Razón de verosimilitud negativa*

LR- = (1-sensibilidad)/especificidad

La razón de verosimilitud negativa se calcula dividiendo la probabilidad de obtener un resultado negativo en presencia de la enfermedad entre la probabilidad de un resultado negativo en ausencia de esta. Se calcula, por lo tanto, como el cociente entre la fracción de falsos negativos (1-sen-

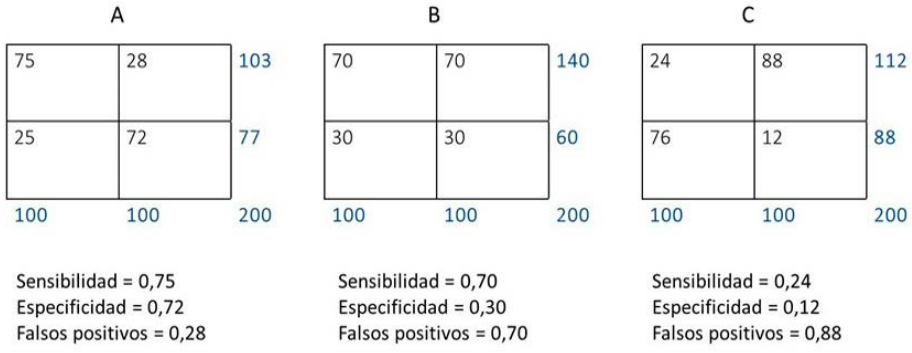


Figura 3. Tablas de contingencia de los métodos A, B y C.

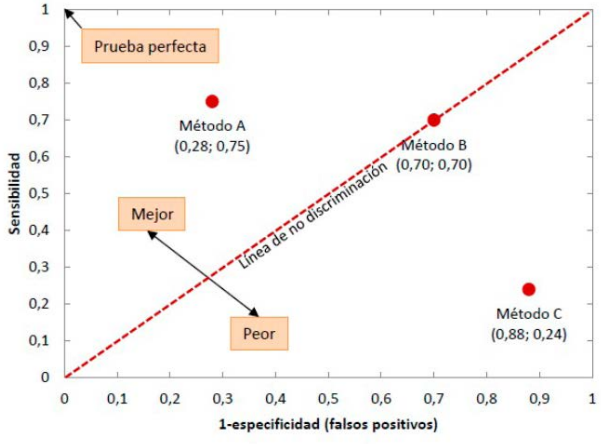


Figura 4. Representación gráfica mediante elaboración de curvas ROC de los datos de la figura 3.

sibilidad) y la fracción de verdaderos negativos (especificidad) [3].

En términos generales, una razón de verosimilitud positiva superior a 10 proporciona evidencia fuerte para aumentar la probabilidad de una enfermedad, mientras que una razón de verosimilitud negativa inferior a 0,1 aporta evidencia sólida para disminuirla. Los valores intermedios producen cambios graduales en la probabilidad diagnóstica y siempre deben interpretarse en conjunto con la probabilidad preprueba y el contexto clínico del paciente [23].

## Interpretación de los parámetros diagnósticos en diferentes escenarios clínicos

Existen situaciones donde se desea comparar las bondades de una prueba diagnóstica antes y después de un tratamiento específico para determinada enfermedad. En un estudio realizado por Gisbert y colaboradores (2003) [24], en el que utilizaron la prueba de aliento con $^{13}$C-urea para el diagnóstico de la infección con *Helicobacter pylori*, encontraron en el grupo pretratamiento los siguientes resultados:

sensibilidad = 96 % (IC95% = 81 % a 99%), especificidad = 100 % (IC95% = 69 % a 100 %), valor predictivo positivo = 100 % (IC95% = 87 % a 100 %), valor predictivo negativo = 92 % (IC95% = 59 % a 100 %), razón de verosimilitud positiva = infinito, y razón de verosimilitud negativa = 0,04. Luego del tratamiento los resultados fueron los siguientes: sensibilidad = 100 % (IC95% = 77 % a 100 %), especificidad = 97 % (IC95% = 90 % a 99 %), valor predictivo positivo = 88 % (IC95% = 62 % a 98 %), valor predictivo negativo = 100 % (IC95% = 95 % a 100 %), razón de verosimilitud positiva = 35 y razón de verosimilitud negativa = 0. Estos resultados muestran que el rendimiento de la prueba puede variar según el contexto clínico y el momento en que se realiza la evaluación, en este caso presentó un excelente rendimiento.

## Probabilidad preprueba

La probabilidad preprueba es la probabilidad de presentar una enfermedad determinada antes de conocer el resultado de una prueba diagnóstica. Esta probabilidad se estima a partir de la información clínica disponible, que puede incluir las manifestaciones clínicas, los factores de riesgo, la prevalencia de la enfermedad en una población similar y los resultados de pruebas realizadas previamente. En pacientes con sospecha de un diagnóstico determinado, la probabilidad preprueba ($PD^+$) se puede interpretar como la proporción de pacientes con la enfermedad ($D^+$) entre todos los pacientes evaluados ($D^+ + D^-$), siendo la fórmula la siguiente:

$$PD^+ = D^+/( D^+ + D^-)$$

Donde ($D^+$) es el número de pacientes con la enfermedad, ($D^-$) es el número de pacientes sin la enfermedad y $PD^+$ es la probabilidad de presentar la enfermedad.

Como ejemplo hipotético, supóngase que en una consulta especializada se evalúan 100 pacientes con elevación persistente de aminotransferasas e hipergammaglobulinemia. Tras el estudio completo, se confirma hepatitis autoinmune en 20 de ellos. En este escenario, la probabilidad preprueba de hepatitis autoinmune es del 20 %. Antes de solicitar pruebas de autoanticuerpos como ANA, ASMA o anti-LKM, un paciente con características clínicas similares tendría una probabilidad ($PD^+$) inicial del 20 % de presentar esta enfermedad.

La probabilidad preprueba es especialmente útil en cuatro situaciones:

1. En la interpretación de los resultados de una prueba diagnóstica.

2. En la selección de una o más pruebas diagnósticas.

3. En la decisión de comenzar el tratamiento:

- Sin pruebas adicionales
- Mientras se espera el resultado de una prueba

4. En la decisión sobre la utilidad clínica de una prueba diagnóstica.

Todo paciente en el que se sospecha una enfermedad tendrá una probabilidad preprueba de presentarla, la cual dependerá de la **prevalencia** de dicha condición clínica en la población, las características del paciente (p. ej. edad, género, factores de riesgo), los signos y síntomas presentes, entre otros. Así, antes de realizar cualquier prueba diagnóstica, el clínico, explícita o implícitamente, le asigna a su paciente una **probabilidad preprueba** de presentar la enfermedad que está siendo evaluada. Una vez realizada la prueba diagnóstica, la probabilidad de presentar la enferme-

dad aumentará o disminuirá, dependiendo del resultado de la prueba en cuestión. Esta nueva probabilidad se llama **probabilidad posprueba** [25].

Una forma de aproximarse a la probabilidad preprueba en un paciente determinado es utilizar la prevalencia de la enfermedad en el estudio que se está analizando (total de pacientes con la enfermedad, o (a + c) en nuestra tabla de contingencia, sobre el total de pacientes del estudio, o (a + b + c + d)). Si nuestro paciente es similar a la población del estudio, sería razonable utilizar este valor. Sin embargo, la probabilidad preprueba también puede estimarse sobre la base del juicio clínico, la prevalencia de la enfermedad en la población a la que pertenece el paciente o mediante variables clínicas y modelos de predicción. Por ejemplo, en pacientes con MASLD, las pruebas no invasivas, como el índice FIB-4, pueden utilizarse para la estratificación inicial del riesgo de fibrosis hepática avanzada y, en pacientes con resultados que indiquen un riesgo aumentado, orientar la realización de una evaluación secundaria mediante técnicas como la elastografía transitoria [26,50].

La probabilidad preprueba de la enfermedad afecta el valor predictivo positivo y el valor predictivo negativo de una prueba diagnóstica y, por lo tanto, determina la utilidad de una prueba diagnóstica. La determinación sérica de AFP, utilizando un punto de corte de 20 ng/mL para el diagnóstico del carcinoma hepatocelular, presenta una sensibilidad del 60,0 % y una especificidad del 90,6 % en pacientes con enfermedad hepática crónica [27]. En este contexto, cuando la prevalencia (probabilidad preprueba) de carcinoma hepatocelular es del 5 %, el valor predictivo positivo de la AFP es del 25,1 % [28]. A medida que aumenta la probabilidad preprueba, también se incrementa el valor predictivo positivo de la prueba, lo que demuestra que la utilidad clínica de un mismo resultado depende en gran medida de la prevalencia de la enfermedad en la población evaluada.

## Diferencias entre razón de verosimilitud y probabilidad preprueba

La razón de verosimilitud y la probabilidad preprueba son conceptos relacionados, pero cumplen funciones diferentes en la interpretación de una prueba diagnóstica. La razón de verosimilitud es una característica del resultado de una prueba y expresa cuánto cambia la probabilidad de enfermedad ante un resultado positivo o negativo. La probabilidad preprueba, en cambio, representa la probabilidad de que el paciente presente la enfermedad antes de conocer el resultado de la prueba y se estima a partir de la información clínica disponible y de la prevalencia de la enfermedad en una población similar. La combinación de la probabilidad preprueba con la razón de verosimilitud permite estimar la probabilidad posprueba. En las **tablas 5** y **6** se presentan ejemplos de estos dos conceptos.

Según los resultados obtenidos, es posible evaluar el rendimiento de las pruebas diagnósticas mediante las razones de verosimilitud. De esta manera, una prueba con una razón de verosimilitud positiva (LR+) mayor de 10 proporciona evidencia fuerte para aumentar la probabilidad de una enfermedad; entre 5 y 10 aporta evidencia moderada; entre 2 y 5 evidencia débil, y menor de 2 tiene escasa utilidad para aumentar la probabilidad diagnóstica. Por su parte, aquella prueba con una razón de verosimilitud negativa (LR-) menor de 0,1 proporciona evidencia fuerte para disminuir la probabilidad de una enfermedad; entre 0,1 y 0,2 aporta evidencia moderada; entre 0,2 y 0,5 aporta evidencia limitada, y entre 0,5 y 1,0 tiene poca utilidad clínica [46].

**Tabla 5.** Ejemplos del cálculo de la razón de verosimilitud en algunas condiciones patológicas.

| Condición patológica | Síntoma, signo o prueba diagnóstica | Tipo de pacientes | Razón de verosimilitud positiva (LR+) |
|---|---|---|---|
| Colangitis biliar primaria [29] | Anticuerpo antimitocondrial (AMA) positivo | Con colestasis bioquímica | 25,2 |
| Hepatitis autoinmune [30] | Anticuerpos anti-SLA/LP positivos | Con elevación persistente de aminotransferasas | 11,1 |
| Hemocromatosis hereditaria [31] | Saturación de transferrina >45 % | Con sobrecarga férrica | 98 |
| Enfermedad de Wilson [32] | Ceruloplasmina sérica <15 mg/dL | Adultos jóvenes con hepatopatía o síntomas neurológicos compatibles | 21,2 |
| Carcinoma hepatocelular [28] | Alfafetoproteína sérica (AFP) | Pacientes con cirrosis o enfermedad hepática crónica:<br>AFP ≥20 ng/mL<br>AFP ≥200 ng/mL<br>AFP ≥400 ng/mL | <br>4,71<br>23,29<br>33,02 |
| Ascitis [33] | Presencia de onda líquida | Sexo masculino | 9,6 |
| Cirrosis hepática [34] | Ascitis<br>Trombocitopenia<br>Telangiectasias<br>Puntaje en la escala de Bonacini para cirrosis hepática mayor que 7 | Con hepatomegalia | 7,2<br>6,3<br>4,3<br>9,4 |
| Tuberculosis intestinal frente a enfermedad de Crohn [35] | Reacción en cadena de la polimerasa (PCR) para tuberculosis | Con enfermedad crónica intestinal | 10,68 |
| Cirrosis hepática [36] | Determinación de interleuquina-6 (IL-6) | Infección bacteriana en pacientes cirróticos | 8,99 |

La principal ventaja de las razones de verosimilitud es que se calculan a partir de la sensibilidad y la especificidad de la prueba y, a diferencia de los valores predictivos, no dependen directamente de la prevalencia de la enfermedad. Por esta razón, constituyen uno de los parámetros más útiles para interpretar y comparar el rendimiento diagnóstico de diferentes pruebas. En la práctica clínica, la probabilidad preprueba puede estimarse a partir de la prevalencia de la enfermedad en una población similar al paciente evaluado, del juicio clínico o mediante modelos de predicción clínica [46].

Fagan, en 1975 [47], describió un nomograma basado en el teorema de Bayes que permite estimar gráficamente la probabilidad posprueba a partir de la probabilidad preprueba y la razón de verosimilitud. El nomograma de Fagan (**figura 5**) tiene tres columnas: la primera corresponde a la probabilidad preprueba de tener la enferme-

**Tabla 6.** Ejemplos del cálculo de la probabilidad preprueba.

| Paciente problema | Enfermedad o prueba diagnóstica | Probabilidad preprueba (%) |
|---|---|---|
| Paciente con colestasis crónica [37] | Colangitis biliar primaria demostrable mediante anticuerpos antimitocondriales (AMA) | 20 a 30 |
| Paciente con elevación persistente de aminotransferasas e hipergammaglobulinemia [38] | Hepatitis autoinmune demostrable mediante ANA, ASMA o anti-SLA/LP | 20 |
| Paciente con ferritina elevada y saturación de transferrina >45 % [39] | Hemocromatosis hereditaria confirmable mediante estudio del gen *HFE* | 15 a 25 |
| Adulto joven (<40 años) con hepatitis inexplicada [40] | Enfermedad de Wilson evaluable mediante ceruloplasmina sérica y excreción urinaria de cobre | 5 a 10 |
| Paciente con cirrosis y nódulo hepático ≥2 cm [41] | Carcinoma hepatocelular diagnosticable mediante TC o RM dinámica con criterios LI-RADS 5 | 68 |
| Paciente con MASLD y FIB-4 elevado [42] | Fibrosis hepática avanzada | 22 |
| Paciente con dispepsia crónica [43] | Infección por *Helicobacter pylori* demostrable por prueba de aliento con $^{13}$C-urea | 25 |
| Paciente con dolor abdominal en hipocondrio derecho [44] | Sospecha de litiasis biliar por ultrasonido abdominal | 40,8 |
| Paciente con sospecha de cáncer pancreático [45] | Diagnóstico por tomografía computarizada de tumor irresecable previo laparoscopia | 41,4 |

dad antes de aplicar la prueba, la segunda a la razón de verosimilitud y la tercera a la probabilidad posprueba. Para utilizarlo, se traza una línea recta entre la probabilidad preprueba y la razón de verosimilitud correspondiente; la prolongación de esta línea permite estimar en la tercera columna la probabilidad posprueba de tener la enfermedad en función del resultado de la prueba.

De acuerdo con este nomograma, la probabilidad posprueba puede estimarse uniendo mediante una línea recta la probabilidad preprueba y la razón de verosimilitud correspondiente. Las pruebas con una LR+ elevada (>10) incrementan considerablemente la probabilidad de enfermedad y pueden ser especialmente útiles para apoyar o confirmar un diagnóstico. En contraste, una LR– inferior a 0,1 reduce de forma importante la probabilidad de enfermedad y puede ser especialmente útil para apoyar su exclusión. Este enfoque constituye una aplicación práctica del teorema de Bayes y facilita la interpretación clínica de los resultados de las pruebas diagnósticas [3,47].

**Ejemplo clínico:** Supóngase un paciente con colestasis crónica cuya probabilidad preprueba de colangitis biliar primaria es

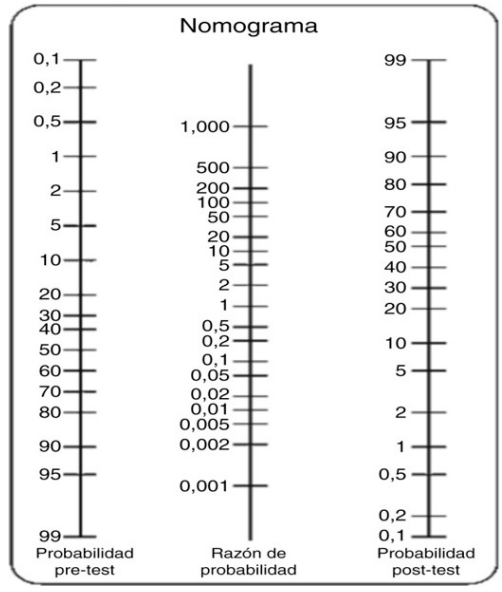


**Figura 5.** Nomograma de Fagan.

del 30 %. Si una prueba para la detección de anticuerpos antimitocondriales (AMA) presenta una LR+ de 69, la aplicación del teorema de Bayes incrementaría la probabilidad posprueba a aproximadamente 96 %. En cambio, si el resultado fuera negativo y la prueba presentara una LR- de 0,05, la probabilidad posprueba disminuiría a aproximadamente 2 %. Este ejemplo hipotético ilustra cómo la combinación de la probabilidad preprueba y la razón de verosimilitud permite estimar la probabilidad posprueba y comprender el impacto de los resultados de una prueba diagnóstica en la toma de decisiones clínicas.

## Influencia de la prevalencia en los resultados de las pruebas diagnósticas

La prevalencia se define como la proporción de individuos de una población que presentan una determinada enfermedad o condición en un momento específico o durante un período determinado. En otras palabras, corresponde al número de individuos que presentan la enfermedad dividida por el total de individuos de la población evaluada [49]. Con base en la tabla 2×2 mostrada anteriormente (**tabla 1**), la prevalencia de la enfermedad en la población estudiada se expresaría mediante la siguiente fórmula:

Prevalencia = (a+c) /(a+b+c+d)

Es importante aclarar que la prevalencia es distinta de la incidencia. La prevalencia representa la proporción de individuos de una población que presentan una enfermedad en un momento determinado o durante un período específico, mientras que la incidencia se refiere a la ocurrencia de nuevos casos de una enfermedad en una población en riesgo durante un período determinado [49]. Si la prevalencia de la enfermedad en la población estudiada se utiliza como una aproximación de la

probabilidad preprueba, es posible calcular las posibilidades (odds) preprueba y posprueba y estimar posteriormente la probabilidad posprueba sin necesidad de utilizar el nomograma de Fagan. Para ello, se aplican las siguientes fórmulas:

- Posibilidades (odds) preprueba = probabilidad preprueba/(1 – probabilidad preprueba)

- Posibilidades (odds) posprueba = posibilidades (odds) preprueba × razón de verosimilitud correspondiente al resultado de la prueba

- Probabilidad posprueba = posibilidades (odds) posprueba/[posibilidades (odds) posprueba + 1]

Cuando el resultado de la prueba es positivo se utiliza la LR+, mientras que cuando el resultado es negativo se utiliza la LR–. A continuación, se aplican los parámetros descritos en esta revisión utilizando los resultados de un estudio de validación del *Enhanced Liver Fibrosis* (ELF) test para el diagnóstico de fibrosis hepática avanzada, empleando la biopsia hepática como estándar de referencia (**tabla 7**) [48].

Nuevamente, a partir de los datos de la tabla de 2x2 (**tabla 7**) calculamos los parámetros de las pruebas diagnósticas aplicando las fórmulas revisadas hasta el momento.

*Sensibilidad = 36 / 54 = 0,667 × 100 = 66,7 %*

*Especificidad = 90 / 116 = 0,776 × 100 = 77,6 %*

*Valor predictivo positivo = 36 / 62 = 0,581 × 100 = 58,1 %*

*Valor predictivo negativo = 90 / 108 = 0,833 × 100 = 83,3 %*

*Porcentaje de falsos negativos = 100 – 66,7 = 33,3 %*

*Porcentaje de falsos positivos = 100 – 77,6 = 22,4 %*

*Razón de verosimilitud positiva = 66,7 / 22,4 = 2,98*

*Razón de verosimilitud negativa = 33,3 / 77,6 = 0,43*

*Prevalencia = 54 / 170 = 0,318 × 100 = 31,8 %*

*Posibilidades (odds) preprueba = 31,8 / 68,2 = 0,47*

*Posibilidades (odds) posprueba = 0,47 × 2,98 = 1,40*

*Probabilidad posprueba = 1,40 / (1,40 + 1) = 1,40 / 2,40 = 0,583 × 100 = 58,3 %*

De acuerdo con estos resultados, el ELF test presentó una sensibilidad del 66,7 %, por lo que un resultado negativo disminuye la probabilidad de fibrosis hepática avanzada, pero no permite descartarla de manera suficiente (no se cumple el acrónimo **SENDES**). Asimismo, la especificidad fue del 77,6 %, por lo que un resultado positivo aumenta la probabilidad de fibrosis avanzada, pero no permite confirmar por sí solo el diagnóstico (no se cumple el acrónimo **ESPIN**). El valor predictivo positivo fue del 58,1 %, lo que indica que poco más de la mitad de los pacientes con un resultado positivo presentaron fibrosis avanzada en esta población, mientras que el valor predictivo negativo fue del 83,3 %, indicando que un resultado negativo disminuye la probabilidad de fibrosis avanzada, aunque no la excluye completamente. Por su parte, la razón de verosimilitud positiva (LR+ = 2,98) incrementa de manera limitada la probabilidad de enfermedad, mientras que la razón de verosimilitud negativa (LR– = 0,43) la reduce de forma limitada.

El nomograma de Fagan (**figura 5**) también permite comprobar gráficamente la relación entre la probabilidad preprueba, la razón de verosimilitud y la probabilidad posprueba. En este ejemplo, al utilizar una probabilidad posprueba de aproximadamente 58,3 % y una LR+ de 2,98, se ob-

**Tabla 7.** Tabla 2x2. Ejemplo de validación diagnóstica del *Enhanced Liver Fibrosis* (ELF) test para fibrosis hepática avanzada.

| Prueba diagnóstica (ELF) | Fibrosis avanzada (+) | Fibrosis avanzada (–) | Total |
|---|---|---|---|
| ELF positivo | 36 (VP) | 26 (FP) | 62 |
| ELF negativo | 18 (FN) | 90 (VN) | 108 |
| Total | 54 | 116 | 170 |

tiene una probabilidad preprueba cercana al 32 %, concordante con la prevalencia observada en la población estudiada (31,8 %). Esta concordancia ilustra la relación matemática entre estos parámetros y muestra cómo la integración de la probabilidad preprueba con la razón de verosimilitud permite estimar la probabilidad posprueba.

La prevalencia de la enfermedad influye en los valores predictivos y en la interpretación clínica de una prueba diagnóstica. Por ello, al evaluar estudios de exactitud diagnóstica es importante considerar las características de la población estudiada, las posibles diferencias en el espectro clínico de los pacientes y el riesgo de sesgos que puedan afectar la generalización de sus resultados a otros contextos [49].

## Consideraciones sobre las pruebas de laboratorio en el diagnóstico clínico

En cada prueba de laboratorio utilizada para el diagnóstico de una enfermedad determinada se deben considerar, además del control de calidad como requisito indispensable para garantizar la confiabilidad de los resultados, su exactitud, precisión y reproducibilidad. También es importante evaluar su validez diagnóstica y su aplicabilidad en la población en la que se pretende utilizar. Así, por ejemplo, en el caso de la fibrosis hepática es importante determinar la presencia y el grado de fibrosis debido a sus implicaciones pronósticas y terapéuticas. Para ello, se encuentran disponibles diferentes pruebas no invasivas, que incluyen índices séricos como el FIB-4, el APRI, el NAFLD Fibrosis Score y el ELF test, además de métodos de imagen como la elastografía transitoria y la elastografía por resonancia magnética [1].

En 2023, Kjaergaard y colaboradores evaluaron el rendimiento diagnóstico del ELF test, el FIB-4 y el NAFLD Fibrosis Score para la detección de fibrosis hepática avanzada en pacientes con enfermedad hepática crónica. Los autores compararon el desempeño de estas pruebas con un estándar de referencia y observaron que su rendimiento diagnóstico variaba según el punto de corte utilizado y las características de la población evaluada. Estos resultados resaltan la importancia de interpretar el rendimiento de las pruebas no invasivas en función del contexto clínico y de la población en la que se aplican [48].

Con base en lo anterior, tomemos otro ejemplo. Supóngase que se decide modificar el punto de corte del índice FIB-4 de 1,3 a 2,67 para identificar fibrosis hepática avanzada. Este cambio modificaría los parámetros de rendimiento diagnóstico de la prueba, especialmente la sensibilidad y la especificidad. En general, al utilizar un punto de corte más bajo aumenta la sensibilidad y disminuye la especificidad, lo que puede favorecer la identificación de pacientes con enfermedad y reducir el número de

falsos negativos. Por el contrario, un punto de corte más alto aumenta la especificidad y disminuye la sensibilidad, lo que reduce el número de falsos positivos. Los valores predictivos también pueden variar con el cambio del punto de corte, aunque su interpretación depende además de la prevalencia o probabilidad preprueba de la enfermedad en la población evaluada [50].

Los valores de sensibilidad y especificidad permiten caracterizar el rendimiento diagnóstico de una prueba, aunque por sí solos no permiten determinar la probabilidad de enfermedad en un paciente individual después de conocer el resultado de la prueba. La sensibilidad y la especificidad no dependen directamente de la prevalencia de la enfermedad como ocurre con los valores predictivos; sin embargo, sus estimaciones pueden variar según las características y el espectro clínico de la población evaluada [50].

Por el contrario, los valores predictivos, a pesar de ser de gran utilidad para la toma de decisiones clínicas y para comunicar a los pacientes la probabilidad de enfermedad asociada con un determinado resultado, presentan la limitación de que dependen de la prevalencia o probabilidad preprueba de la enfermedad en la población evaluada. En general, cuando la prevalencia de la enfermedad es baja, el valor predictivo negativo tiende a ser elevado, mientras que el valor predictivo positivo tiende a ser bajo. Por ello, un resultado negativo puede ser más útil para disminuir la probabilidad de enfermedad, mientras que un resultado positivo puede no ser suficiente para confirmarla, especialmente cuando la prueba no presenta una sensibilidad o especificidad adecuadas [51].

Al igual que en los estudios sobre tratamiento, los resultados de los estudios de exactitud diagnóstica deben informarse acompañados de sus respectivos intervalos de confianza. El intervalo de confianza (IC) expresa la incertidumbre asociada a la estimación de un parámetro y permite valorar la precisión de los resultados. Habitualmente se utiliza un intervalo de confianza del 95 % (IC95 %). En términos generales, cuanto más estrecho es el intervalo de confianza, mayor es la precisión de la estimación; por el contrario, un intervalo amplio indica mayor incertidumbre alrededor del valor estimado [52].

## No inferioridad y concordancia en la comparación de pruebas de laboratorio

El concepto de no inferioridad se aplica principalmente en los ensayos clínicos, especialmente en el campo de la terapéutica, para determinar si una nueva intervención no es inferior a una intervención de referencia por más de un margen preespecificado, que debe establecerse antes de realizar el estudio. Este enfoque puede aplicarse también a la evaluación comparativa de pruebas diagnósticas cuando se define previamente un margen clínicamente aceptable de diferencia en su rendimiento. Sin embargo, la no inferioridad y la concordancia son conceptos diferentes. La concordancia se refiere al grado de acuerdo entre dos o más métodos, técnicas u observadores que evalúan el mismo fenómeno, mientras que la no inferioridad busca demostrar que el rendimiento de una nueva intervención o prueba no es inferior al de una referencia más allá de un límite previamente establecido [13,53].

La concordancia entre los métodos y sus mediciones puede alterarse por los siguientes elementos o fuentes de error: a) la variabilidad entre observadores, b) la variabilidad asociada con el instrumento o método de medición y c) la variabilidad derivada de mediciones realizadas en diferentes momentos. Por ello, cuando

se evalúa la reproducibilidad o concordancia de una prueba, es importante estandarizar las condiciones de medición y valorar la variabilidad intraobservador e interobservador, cuando corresponda. Por otra parte, cuando el objetivo es evaluar la exactitud diagnóstica de una nueva prueba, su rendimiento debe compararse con un estándar de referencia apropiado [13,53].

Durante la evaluación de las características operativas de una nueva prueba diagnóstica pueden introducirse diferentes fuentes de sesgo. Una de ellas es la selección inadecuada del estándar de referencia, especialmente cuando este no representa de manera apropiada la condición clínica que se pretende diagnosticar o cuando su aplicación difiere entre los participantes. Estas situaciones pueden afectar la estimación de la exactitud diagnóstica de la nueva prueba y conducir a resultados sesgados [52,53]. Una medida utilizada para evaluar la concordancia entre dos métodos o entre dos observadores cuando las variables son categóricas es el índice o coeficiente kappa (κ), que se calcula a partir de una tabla de contingencia. Este coeficiente cuantifica el grado de acuerdo observado entre las mediciones después de considerar el acuerdo esperado por azar. Un valor de $\kappa = 1$ indica concordancia perfecta, mientras que un valor de $\kappa = 0$ indica que la concordancia observada no es mayor que la esperada por azar. Los valores negativos indican una concordancia inferior a la esperada por azar [52]. El cálculo se realiza de la siguiente manera:

*La proporción de concordancia observada se calcula como:*

*Concordancia observada = (a + d) / n*

*La concordancia esperada por azar se calcula como:*

*Concordancia esperada = [(a + b)(a + c) + (c + d)(b + d)] / $n^2$*

*Por lo tanto, el coeficiente kappa se calcula mediante la siguiente fórmula:*

$\kappa$ = *(concordancia observada – concordancia esperada) / (1 – concordancia esperada)*

De esta manera, el coeficiente kappa representa la proporción de concordancia observada que excede la concordancia esperada por azar, en relación con la máxima concordancia posible que no se atribuye al azar.

De acuerdo con una clasificación convencional propuesta por Landis y Koch, un valor de κ entre 0,21 y 0,40 indica una concordancia débil; entre 0,41 y 0,60, una concordancia moderada; entre 0,61 y 0,80, una concordancia sustancial, y superior a 0,80, una concordancia casi perfecta [50,53]. Estas categorías son orientativas y deben interpretarse considerando el contexto clínico y las características de los métodos evaluados. Es importante señalar que la concordancia medida mediante el coeficiente kappa no demuestra por sí misma la no inferioridad o equivalencia entre dos pruebas diagnósticas. La demostración de no inferioridad requiere establecer previamente un margen aceptable de diferencia y aplicar un análisis estadístico diseñado específicamente para este propósito.

## Veámos el siguiente ejemplo:

Para el diagnóstico de la hepatitis autoinmune existen diferentes sistemas de clasificación que apoyan el diagnóstico clínico. Con el propósito de evaluar la concordancia entre los criterios simplificados del *International Autoimmune Hepatitis Group* (IAIHG, 2008) y las recomendaciones actualizadas del IAIHG publicadas en 2022, se analizaron biopsias hepáticas

de pacientes con sospecha de hepatitis autoinmune. El índice kappa para la concordancia global entre ambos sistemas fue de 0,638, lo que, de acuerdo con la clasificación convencional de Landis y Koch, corresponde a una concordancia sustancial. Asimismo, la concordancia para la evaluación de la hepatitis portal fue de κ = 0,681, también correspondiente a una concordancia sustancial, mientras que para la hepatitis lobulillar fue de κ = 0,483, lo que indica una concordancia moderada. Estos resultados sugieren que el grado de concordancia puede variar según el hallazgo histológico evaluado [54].

Ilustraremos otro ejemplo:

La reproducibilidad de los hallazgos por resonancia magnética también ha sido evaluada en pacientes con hepatitis autoinmune. En un estudio dirigido a determinar la concordancia interobservador de diferentes hallazgos imagenológicos, la nodularidad de la superficie hepática presentó un índice κ = 0,83, mientras que la ascitis alcanzó un κ = 0,89, lo que representa una excelente concordancia entre los radiólogos. En contraste, otros hallazgos como la fibrosis reticular (κ=0,12) y el realce heterogéneo del parénquima (κ=0,25) mostraron una concordancia pobre, lo que evidencia que no todas las características evaluadas mediante resonancia magnética poseen la misma reproducibilidad diagnóstica [55].

En conclusión, estos ejemplos ilustran que el índice kappa constituye una herramienta útil para cuantificar la concordancia entre diferentes métodos diagnósticos o entre observadores cuando se evalúan variables categóricas. Los valores elevados de κ indican un mayor grado de acuerdo más allá del esperado por azar, mientras que los valores bajos sugieren una mayor variabilidad entre las mediciones. Sin embargo, la concordancia no debe interpretarse como evidencia de exactitud diagnóstica o utilidad clínica por sí sola. Para una mayor comprensión del índice o coeficiente kappa se recomienda consultar revisiones metodológicas específicas, como las de Landis y Koch (1977) y Cerda y Villarroel (2008) [53,54].

## Conclusiones

La importancia de los parámetros utilizados para analizar los resultados de las pruebas diagnósticas y de laboratorio radica en que permiten valorar su capacidad para identificar correctamente a los pacientes con una determinada condición clínica y diferenciarlos de aquellos que no la presentan. En condiciones ideales, una prueba diagnóstica sería aquella capaz de identificar a todos los pacientes con la enfermedad y, al mismo tiempo, clasificar correctamente a todos aquellos que no la presentan, alcanzando una sensibilidad y una especificidad del 100 %.

Los resultados de una prueba diagnóstica permiten estimar la probabilidad de que un paciente presente o no una determinada enfermedad, para lo cual deben interpretarse en función de la probabilidad preprueba y de las características de la prueba utilizada. De allí la relevancia de analizar su sensibilidad, especificidad y valores predictivos positivo y negativo. Además, las razones de verosimilitud permiten cuantificar el cambio en la probabilidad de enfermedad después de conocer el resultado de una prueba y facilitan su interpretación en el proceso de toma de decisiones clínicas. En este sentido, es importante reconocer que el rendimiento de una prueba diagnóstica puede variar según la población evaluada, el punto de corte utilizado y el contexto clínico en el que se aplique.

Dado que el rendimiento de una prueba diagnóstica puede variar según las carac-

terísticas de la población estudiada, los estudios de exactitud diagnóstica deben proporcionar una descripción detallada de los pacientes incluidos y, cuando sea pertinente, de los diferentes subgrupos evaluados. Asimismo, deben considerarse factores como la historia clínica, las comorbilidades, las características del protocolo de la prueba y el espectro clínico o grado de severidad de la enfermedad, ya que estos elementos pueden influir en la estimación del rendimiento diagnóstico y en la aplicabilidad de los resultados a otros contextos [56]. Estos aspectos adquieren especial relevancia en hepatología, donde la interpretación de pruebas no invasivas, biomarcadores y métodos de imagen puede variar según la etiología y el estadio de la enfermedad hepática. Del mismo modo, la evaluación de la reproducibilidad y la concordancia entre métodos u observadores constituye un aspecto complementario en la valoración de nuevas pruebas diagnósticas y de laboratorio, particularmente cuando se pretende determinar la consistencia de sus resultados [51].

## Nota editorial

Este artículo corresponde a una versión revisada, actualizada y adaptada al contexto de la hepatología de un trabajo publicado previamente en 2017 por el mismo autor [57]. En esta versión se actualizan los conceptos relacionados con la medicina basada en la evidencia y la evaluación de las pruebas diagnósticas, y se incorporan nuevos ejemplos y referencias aplicados específicamente al ámbito de la hepatología clínica.

## Referencias


1. **Berzigotti A, Tsochatzis E, Boursier J, Castera L, Cazzagon N, Friedrich-Rust M, et al.** EASL Clinical Practice Guidelines on non-invasive tests for evaluation of liver disease severity and prognosis – 2021 update. J Hepatol 2021;75:659-689. https://doi.org/10.1016/j.jhep.2021.05.025.
2. **Kanwal F, Neuschwander-Tetri BA, Loomba R, Rinella ME.** Metabolic dysfunction-associated steatotic liver disease: Update and impact of new nomenclature on the American Association for the Study of Liver Diseases practice guidance on nonalcoholic fatty liver disease. Hepatology 2024;79:1212-1219. https://doi.org/10.1097/HEP.0000000000000670.
3. **Vizcaíno-Salazar G.** Medicina basada en la evidencia y análisis de diseños de investigación clínica. Maracaibo, Venezuela: EDILUZ; 2002.
4. **Centre for Evidence-Based Medicine.** SpPin and SnNout. Oxford: University of Oxford; 2017. https://www.cebm.ox.ac.uk/resources/ebm-tools/sppin-and-snnout.
5. **Hirschfield GM, Beuers U, Corpechot C, Invernizzi P, Jones D, Marzioni M, et al.** EASL Clinical Practice Guidelines: The diagnosis and management of patients with primary biliary cholangitis. J Hepatol 2017;67:145-172. https://doi.org/10.1016/j.jhep.2017.03.022.
6. **Griner PF, Mayewski RJ, Mushlin AI, Greenland P.** Selection and interpretation of diagnostic tests and procedures. Principles and applications. Ann Intern Med 1981;94:557-592.
7. **Pita-Fernández S, Pértegas-Díaz S.** Pruebas diagnósticas: sensibilidad y especificidad. Cad Aten Primaria 2003;10:120-124.
8. **Bossuyt PM, 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. https://doi.org/10.1136/bmj.h5527.
9. **Sierra-Arango F.** La sensibilidad y especificidad: entendiendo su origen y utilidad real. Rev Colomb Gastroenterol 2003;18:180-182.
10. **Singal AG, Llovet JM, Yarchoan M, Mehta N, Heimbach JK, Dawson LA, et al.** AASLD Practice Guidance on prevention, diagnosis, and treatment of hepatocellular carcinoma. Hepatology 2023;78:1922-1965. https://doi.org/10.1097/HEP.0000000000000466.
11. **Ghosal A, Ghosal S.** Global bivariate meta-analysis of FIB-4 cut-offs to rule out advanced fibrosis in MASLD. Int J Hepatol 2026;2026:9419416. https://doi.org/10.1155/ijh/9419416.

12. **Bandin O, Courvalin JC, Poupon R, Dubel L, Homberg JC, Johanet C.** Specificity and sensitivity of gp210 autoantibodies detected using an enzyme-linked immunosorbent assay and a synthetic polypeptide in the diagnosis of primary biliary cirrhosis. Hepatology 1996;23:1020-1024. https://doi.org/10.1002/hep.510230512.
13. **Cortés-Reyes É, Rubio-Romero JA, Gaitán-Duarte H.** Métodos estadísticos de evaluación de la concordancia y la reproducibilidad de pruebas diagnósticas. Rev Colomb Obstet Ginecol 2010;61:247-255. https://doi.org/10.18597/rcog.271.
14. **Cerda J, Cifuentes L.** Uso de curvas ROC en investigación clínica: aspectos teórico-prácticos. Rev Chil Infectol 2012;29:138-141. https://doi.org/10.4067/S0716-10182012000200003.
15. **Akobeng AK.** Understanding diagnostic tests 3: Receiver operating characteristic curves. Acta Paediatr 2007;96:644-647. https://doi.org/10.1111/j.1651-2227.2006.00178.x.
16. **Altman DG, Bland JM.** Diagnostic tests 3: receiver operating characteristic plots. BMJ 1994;309:188. https://doi.org/10.1136/bmj.309.6948.188.
17. **Valenzuela DL, Cifuentes AL.** Validez de estudios de tests diagnósticos. Rev Med Chile 2008;136:401-404. https://doi.org/10.4067/S0034-98872008000300018.
18. **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;155:529-536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009.
19. **Schünemann HJ, Mustafa RA, Brozek J, Santesso N, Bossuyt PM, Steingart KR, et al.** GRADE guidelines: 22. The GRADE approach for tests and strategies—from test accuracy to patient-important outcomes and recommendations. J Clin Epidemiol 2019;111:69-82. https://doi.org/10.1016/j.jclinepi.2019.02.003.
20. **López de Ullibarri-Galparsoro I, Pita-Fernández S.** Curvas ROC. Cad Aten Primaria 1998;5:229-235.
21. **XLSTAT by Lumivero.** ROC curves. París, Francia: Lumivero; 2026. https://www.xlstat.com/solutions/features/roc-curves.
22. **Ortiz AT.** Curvas ROC para datos de supervivencia. Aplicación a datos biomédicos. Santiago de Compostela: Universidade de Santiago de Compostela; 2010.
23. **McGee S.** Simplifying likelihood ratios. J Gen Intern Med 2002;17:646-649. https://doi.org/10.1046/j.1525-1497.2002.10750.x.
24. **Gisbert J, Ducons J, Gomollón F, Dominguez-Munoz JE, Celaya F, Miño G, et al.** Validation of the C-13-urea breath test for the initial diagnosis of Helicobacter pylori infection and to confirm eradication after treatment. Rev Esp Enferm Dig 2003;95:115-121.
25. **Salech F, Mery V, Larrondo F, Rada G.** Estudios que evalúan un test diagnóstico: interpretando sus resultados. Rev Med Chile 2008;136:1208-1208. https://doi.org/10.4067/S0034-98872008000900018.
26. **Rinella ME, Neuschwander-Tetri BA, Siddiqui MS, Abdelmalek MF, Caldwell S, Barb D, et al.** AASLD Practice Guidance on the clinical assessment and management of nonalcoholic fatty liver disease. Hepatology 2023;77:1797-1835. https://doi.org/10.1097/HEP.0000000000000323.
27. **Trevisani F, D'Intino PE, Morselli-Labate AM, Mazzella G, Accogli E, Caraceni P, et al.** Serum alpha-fetoprotein for diagnosis of hepatocellular carcinoma in patients with chronic liver disease: influence of HBsAg and anti-HCV status. J Hepatol 2001;34:570-575. https://doi.org/10.1016/S0168-8278(00)00053-2.
28. **Zhang J, Chen G, Zhang P, Zhang J, Li X, Gan D, et al.** The threshold of alpha-fetoprotein (AFP) for the diagnosis of hepatocellular carcinoma: A systematic review and meta-analysis. PLoS One 2020;15:e0228857. https://doi.org/10.1371/journal.pone.0228857.
29. **Xu Q, Zhu W, Yin Y.** Diagnostic value of anti-mitochondrial antibody in patients with primary biliary cholangitis: A systematic review and meta-analysis. Medicine (Baltimore) 2023;102:e36039. https://doi.org/10.1097/MD.0000000000036039.

30. **Zhang WC, Zhao FR, Chen J, Chen WX.** Meta-Analysis: Diagnostic Accuracy of Antinuclear Antibodies, Smooth Muscle Antibodies and Antibodies to a Soluble Liver Antigen/Liver Pancreas in Autoimmune Hepatitis. PLoS One 2014;9:e92267. https://doi.org/10.1371/journal.pone.0092267.
31. **Kane SF, Roberts C, Paulus R.** Hereditary Hemochromatosis: Rapid Evidence Review. Am Fam Physician 2021;104:263-270.
32. **Xu R, Jiang YF, Zhang YH, Yang X.** The optimal threshold of serum ceruloplasmin in the diagnosis of Wilson's disease: A large hospital-based study. PLoS One 2018;13:e0190887. https://doi.org/10.1371/journal.pone.0190887.
33. **Williams JW, Simel DL.** The rational clinical examination. Does this patient have ascites? How to divine fluid in the abdomen. JAMA 1992;267:2645-2648. https://doi.org/10.1001/jama.267.19.2645.
34. **Udell JA, Wang CS, Tinmouth J, FitzGerald JM, Ayas NT, Simel DL, et al.** Does this patient with liver disease have cirrhosis? JAMA 2012;307:832-842. https://doi.org/10.1001/jama.2012.186.
35. **Jin T, Fei B, Zhang Y, He X.** The Diagnostic Value of Polymerase Chain Reaction for Mycobacterium tuberculosis to Distinguish Intestinal Tuberculosis from Crohn's Disease: A Meta-analysis. Saudi J Gastroenterol 2017;23:3-10. https://doi.org/10.4103/1319-3767.199135.
36. **Wu Y, Wang M, Zhu Y, Lin S.** Serum interleukin-6 in the diagnosis of bacterial infection in cirrhotic patients: A meta-analysis. Medicine (Baltimore) 2016;95:e5127. https://doi.org/10.1097/MD.0000000000005127.
37. **Lindor KD, Bowlus CL, Boyer J, Levy C, Mayo M.** Primary Biliary Cholangitis: 2018 Practice Guidance from the American Association for the Study of Liver Diseases. Hepatology 2019;69:394-419. https://doi.org/10.1002/hep.30145.
38. **Mack CL, Adams D, Assis DN, Kerkar N, Manns MP, Mayo MJ, et al.** Diagnosis and Management of Autoimmune Hepatitis in Adults and Children: 2019 Practice Guidance and Guidelines From the American Association for the Study of Liver Diseases. Hepatology 2020;72:671-722. https://doi.org/10.1002/hep.31065.
39. **Bacon BR, Adams PC, Kowdley KV, Powell LW, Tavill AS, American Association for the Study of Liver Diseases.** Diagnosis and management of hemochromatosis: 2011 practice guideline by the American Association for the Study of Liver Diseases. Hepatology 2011;54:328-343. https://doi.org/10.1002/hep.24330.
40. **European Association for the Study of the Liver.** EASL Clinical Practice Guidelines: Wilson's disease. J Hepatol 2012;56:671-685. https://doi.org/10.1016/j.jhep.2011.11.007.
41. **Bremner KE, Bayoumi AM, Sherman M, Krahn MD.** Management of solitary 1 cm to 2 cm liver nodules in patients with compensated cirrhosis: A decision analysis. Can J Gastroenterol 2007;21:491-500. https://doi.org/10.1155/2007/182383.
42. **Chen Y, Dong B, Jin X, Liu C, Zheng MH, Liang X, et al.** Two-step clinical care pathway to predict MASLD-related advanced fibrosis and long-term outcomes in type 2 diabetes. Gut 2026;75:576-587. https://doi.org/10.1136/gutjnl-2025-337506.
43. **Chey WD, Wong BCY, Practice Parameters Committee of the American College of Gastroenterology.** American College of Gastroenterology guideline on the management of Helicobacter pylori infection. Am J Gastroenterol 2007;102:1808-1825. https://doi.org/10.1111/j.1572-0241.2007.01393.x.
44. **Gurusamy KS, Giljaca V, Takwoingi Y, Higgie D, Poropat G, Štimac D, et al.** Ultrasound versus liver function tests for diagnosis of common bile duct stones. Cochrane Database Syst Rev 2015;2015:CD011548. https://doi.org/10.1002/14651858.CD011548.
45. **Allen VB, Gurusamy KS, Takwoingi Y, Kalia A, Davidson BR.** Diagnostic accuracy of laparoscopy following computed tomography (CT) scanning for assessing the resectability with curative intent in pancreatic and periampullary cancer. Cochrane Database Syst Rev 2016;2016:CD009323. https://doi.org/10.1002/14651858.CD009323.pub3.
46. **Aznar Oroval E, Mancheño Álvaro A, García Lozano T, Sánchez Yepes M.** Razón de verosimilitud y nomograma de Fagan: 2 instrumentos básicos para un uso racional de

las pruebas del laboratorio clínico. Rev Calid Asist 2013;28:390-391. https://doi.org/10.1016/j.cali.2013.06.002.

47. **Fagan TJ.** Letter: Nomogram for Bayes's theorem. N Engl J Med 1975;293:257. https://doi.org/10.1056/NEJM197507312930513.
48. **Kjaergaard M, Lindvig KP, Thorhauge KH, Andersen P, Hansen JK, Kastrup N, et al.** Using the ELF test, FIB-4 and NAFLD fibrosis score to screen the population for liver disease. J Hepatol 2023;79:277-286. https://doi.org/10.1016/j.jhep.2023.04.002.
49. **Ruiz-Canela Cáceres J, García-Vera C.** Las estimaciones de validez de las pruebas diagnósticas varían según la prevalencia en los estudios. Evid Pediatr 2014;10:8.
50. **Diamond GA.** Clinical epistemology of sensitivity and specificity. J Clin Epidemiol 1992;45:9-13. https://doi.org/10.1016/0895-4356(92)90181-l
51. **Altman DG, Bland JM.** How to obtain the confidence interval from a P value. BMJ 2011;343:d2090. https://doi.org/10.1136/bmj.d2090.
52. **Cerda J, Villarroel L.** Evaluación de la concordancia inter-observador en investigación pediátrica: Coeficiente de Kappa. Rev Chil Pediatr 2008;79:54-58. https://doi.org/10.4067/S0370-41062008000100008.
53. **Landis JR, Koch GG.** The measurement of observer agreement for categorical data. Biometrics 1977;33:159-174.
54. **Moral K, Simsek B, Tomar VB, Albayrak C, Ergin M, Kilic G, et al.** Comparison of the recommendation of international autoimmune hepatitis pathology group 2022 and the simplified criteria for autoimmune hepatitis 2008: A preliminary study. Hepatol Forum 2024;5:100-105. https://doi.org/10.14744/hf.2023.2023.0061.
55. **Gomes NBN, Torres US, Silva GSE, Mamone POS, Ferraz MLCG, D'ippolito G.** Magnetic resonance imaging findings in autoimmune hepatitis: how frequent and reproducible are they? Radiol Bras 2023;56:308-316. https://doi.org/10.1590/0100-3984.2023.0044.
56. **Hlatky MA, Pryor DB, Harrell FE, Califf RM, Mark DB, Rosati RA.** Factors affecting sensitivity and specificity of exercise electrocardiography. Multivariable analysis. Am J Med 1984;77:64-71. https://doi.org/10.1016/0002-9343(84)90437-6.
57. **Vizcaíno-Salazar GJ.** Importancia del cálculo de la sensibilidad, la especificidad y otros parámetros estadísticos en el uso de las pruebas de diagnóstico clínico y de laboratorio. Med Lab 2017;23:365-386. https://doi.org/10.36384/01232576.34.