

Evaluación de la validez y confiabilidad de una prueba diagnóstica (*Assessment of the validity and reliability of a diagnostic test*)

José H Donis¹ ✉

¹ Instituto de Investigaciones Cardiovasculares, Universidad de los Andes, Mérida - Venezuela.

[ARTICULO DE REVISION]

Recibido: 18 de Marzo de 2012. Aceptado: 18 de Mayo de 2012.

Resumen

En esta revisión se discuten, de manera resumida, los conceptos fundamentales de lo que significa la validez de una prueba a través de sus índices, como son la sensibilidad y especificidad, consideradas como propiedades intrínsecas de una prueba diagnóstica. Así mismo, se exponen los conceptos fundamentales de los valores predictivos positivos y negativos en la práctica clínica y su asociación con la prevalencia de la enfermedad estudiada. Seguidamente se analizan las razones de verosimilitud o de probabilidad, las cuales, como su nombre lo indica, son razones que evidencian con qué frecuencia un individuo padece una enfermedad, cuando son positivas, en relación con otro individuo que no la padece, la cual también es positiva. Adicionalmente, se evalúan las curvas ROC como una metodología para categorizar en forma dicotómica es decir, positivos o negativos, aquellas pruebas con valores en escala continua. Finalmente, se discute la confiabilidad de una prueba, parámetro de gran valor en la veracidad de cualquier prueba que se realice en forma cotidiana en la práctica clínica, ya que asegura la repetitividad de la misma en el sitio de su ejecución.

Palabras clave

Sensibilidad, Especificidad, Valores Predictivos, Razón de Verosimilitud, curvas ROC y Confiabilidad.

Abstract

This review is intended to explain briefly, the significance of the validity of a diagnostic test by using specific indicators such as: the sensitivity, and the specificity, which are considered as intrinsic properties of the test. Additionally, the fundamental concepts of positive and negative predictive values of a diagnostic test in the clinical practice, are described and highlighted, and their connection to the prevalence of a particular disease. The likelihood ratio, meaning how many times a healthy person suffers from a disease, in other words, the percentage of ill people diagnosed with a given test result, divided by the percentage of healthy individuals with the same results. Ideally, abnormal test results should be much more typical in ill individuals, than in those who are healthy (high likelihood ratio), and normal test results should be more frequent in healthy people rather than in sick people (low likelihood ratio). Furthermore, the receiver operating characteristic curve (ROC) is described as a methodology to categorize, in a dichotomy form, as positive or negative those tests with values in continue scale. Finally, the reliability or reproducibility, which is related to the amount of error in any measurement, is analyzed. A more formal definition of reliability is variability between subjects, divided by inter-subject variability plus measurement, which could give as the best reliability of the test in clinical practice.

Keywords

Sensitivity, specificity, predictive value, curves ROC, likelihood ratios, reliability.

Introducción

En un intento de reducir la incertidumbre inherente al manejo de patologías en los seres humanos, se han utilizado las pruebas diagnósticas,

cuya función no solamente es la deducción de un dilema, sino la cuantificación y descripción de la duda que se pueda generar con respecto a una patología (1). El interrogatorio y el examen clínico, son pruebas diagnósticas, que tienen sensibilidad y especificidad con gran dependencia del operador que la realiza. Con

frecuencia, la información que se tiene en un caso particular no es suficiente para la formulación de un diagnóstico, ni para la toma de decisiones, en este caso debe recurrirse a una prueba diagnóstica que tenga la suficiente validez para aproximarse con certeza al diagnóstico de una enfermedad.

La validez de un prueba diagnóstica, se refiere a la extensión a la cual un test o sustituto, mide lo que se desea cuantificar. Existen varios tipos de validez: validez de contenido, esta describe la extensión donde una medida refleja la dimensión de un problema particular; validez de constructo, se refiere a la extensión donde una medida conforma un fenómeno externo establecido; y por último, validez de criterio, la cual se define como la extensión donde una medida se correlaciona con una prueba de referencia (*gold standard*), es decir, que pueda predecir un fenómeno observable (2). Esos tipos de validez son a menudo aplicados a cuestionarios en los cuales la verdad es físicamente no verificable. Por lo tanto, se puede definir la validez de una prueba como su capacidad para distinguir entre los individuos que tienen la enfermedad, de aquellos que no la tienen (3).

En esta revisión, se discutirá acerca de las herramientas utilizadas para validar pruebas

diagnósticas y determinación de su nivel de confiabilidad.

Prueba Diagnóstica

Antes de describir como se evalúa una prueba diagnóstica, es importante definirla. Se define como prueba diagnóstica a cualquier proceso, más o menos complejo, que pretenda determinar en un paciente la presencia de cierta condición, supuestamente patológica, no susceptible de ser observable directamente con algunos de los cinco sentidos elementales (4). Al hablar de pruebas diagnósticas debe tenerse en cuenta los umbrales diagnósticos y terapéuticos, que son los umbrales del pre-examen del diagnóstico o umbral de incertidumbre, donde se justifica la prueba diagnóstica (4) (ver figura 1a).

Las pruebas diagnósticas son utilizadas para diferentes fines: tamizaje de una población, búsqueda de casos, descarte de un diagnóstico, confirmación de un diagnóstico o seguimiento de una patología (4).

Para una correcta evaluación de una prueba diagnóstica se deben conocer los siguientes elementos: sensibilidad, especificidad, valor predictivo

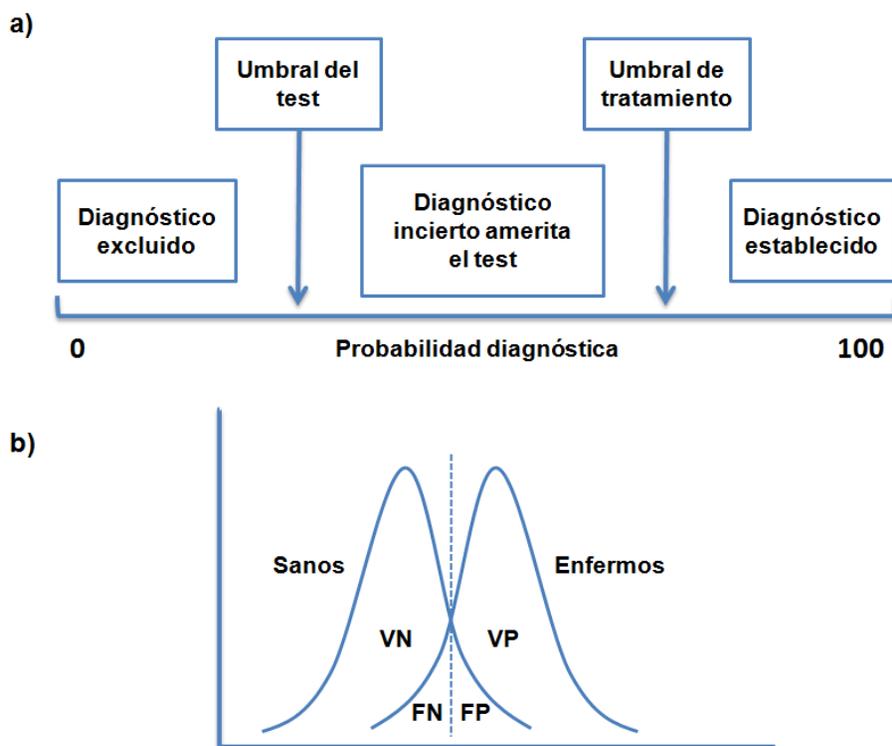


Figura 1. a) Umbrales para diagnóstico y tratamiento, como una función de probabilidades del diagnóstico y tratamiento. b) Umbral de diagnóstico. VN= verdaderos negativos. VP= verdaderos positivos. FN= falsos negativos. FP=falsos positivos.

positivo, valor predictivo negativo, índice de eficiencia pronóstica o razón de verosimilitud tanto positiva como negativa, teorema de Bayes, curvas ROC (del inglés, receiver operating characteristic curve), pruebas múltiples: en paralelo y en series; y confiabilidad o repetitividad de un test o prueba.

Evaluación de las Pruebas Diagnósticas

Sensibilidad: Es definida como la capacidad de una prueba para identificar correctamente aquellos que tienen la enfermedad. Esta es igual al número de sujetos con un test positivo que tienen la enfermedad, dividido entre todos los sujetos que tienen la enfermedad (1-4) (ver figura 2 y tabla 1).

En la figura 2, la flecha señala la dirección de la sensibilidad desde la enfermedad hacia el resultado de la prueba, es decir, es la probabilidad de que en un individuo enfermo, la prueba resulte positiva. Así, la sensibilidad es la probabilidad dada por la proporción de sujetos con resultado positivo entre el total de enfermos.

Especificidad: es definida como la capacidad de una prueba para identificar aquellos que no tienen la enfermedad, y es igual al número de sujetos que resultan negativos a la prueba y que no tienen la enfermedad, dividido entre el número de personas que no tienen la enfermedad o están sanos (ver figura 2 y tabla 1). Así, se evidencia en la figura 2 que la

flecha señala la dirección de la especificidad, y su trayecto indica que va desde la ausencia de enfermedad al resultado de la prueba, es decir, la probabilidad de que en un individuo no enfermo, la prueba resulte negativa.

Como norma general, cuando se desea evaluar la sensibilidad y la especificidad de una prueba, se debe partir del conocimiento de la presencia o no de la enfermedad en los individuos estudiados, y comparar los resultados de la prueba en evaluación, con una prueba de referencia o "gold standard".

La sensibilidad y la especificidad son características intrínsecas de una prueba, son interdependientes ya que un aumento de la sensibilidad está acompañada por una reducción de la especificidad y viceversa, esto es válido en las pruebas con escalas continuas donde el nivel umbral o "cut-off" para un resultado positivo puede variar, por ejemplo, el uso del conteo de leucocitos, como una prueba para diagnosticar infección bacteriana. Si se selecciona un umbral alto para identificar infecciones bacterianas (mayor de 25.000 leucocitos), se corre el riesgo de pasar por alto infecciones bacterianas leves a moderadas. En este caso la prueba tendrá una baja sensibilidad ya que es muy difícil encontrar este valor en infecciones leves, sin embargo tendrá una alta especificidad, ya que todos los pacientes con este valor tienen infección. Sí por el contrario, se disminuye el umbral para el diagnóstico de infección bacteriana a

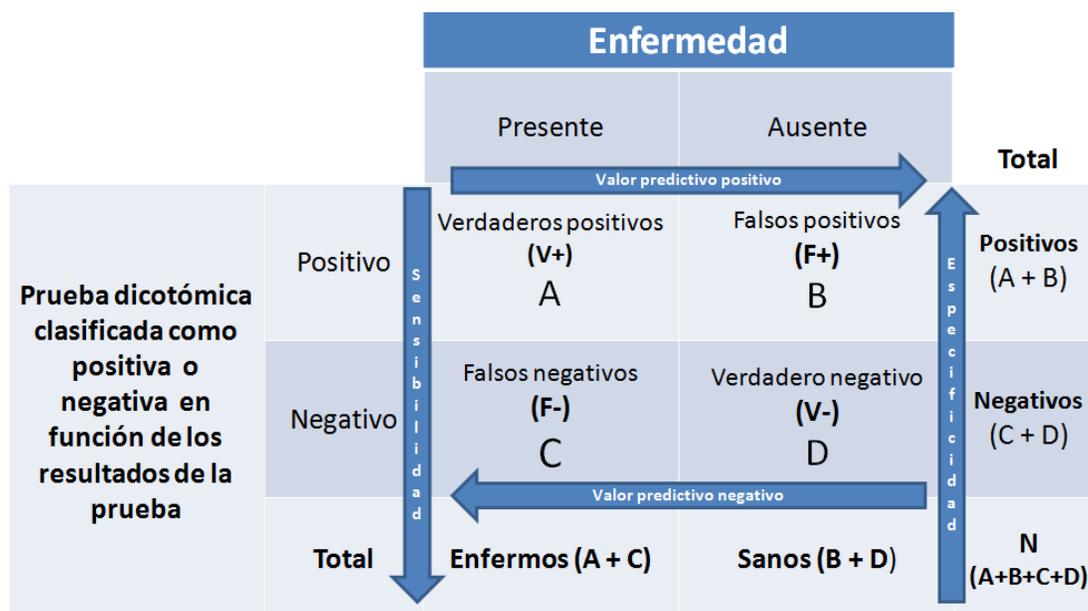


Figura 2. Resultados de una prueba diagnóstica.

Tabla 1. Indicadores importantes en el estudio de las pruebas diagnósticas (ver figura 2).

Indicador	Cálculo / significado
Sensibilidad	$A / (A + C)$
Especificidad	$D / (B + D)$
Valor predictivo positivo	$A / (A + B)$
Valor predictivo negativo	$D / (C + D)$
Probabilidad de falsos positivos	$B / (B + D)$ o 1-especificidad
Probabilidad de falsos negativos	$C / (A + C)$ o 1-sensibilidad
Prevalencia de la enfermedad	$(A + C) / N$
Proporción de sanos	$(B + D) / N$
Razón de verosimilitud positiva	Sensibilidad / (1-especificidad)
Razón de verosimilitud negativa	1-sensibilidad / (especificidad)

10.000 leucocitos/mm³, la especificidad de la prueba disminuirá, a expensas de un incremento de la sensibilidad (2, 5).

Seguridad de una Prueba: La seguridad de una prueba está determinada por los valores predictivos positivos (VPP) y negativos (VPN). Estos índices son importantes para valorar la utilidad de una prueba, en el terreno clínico y de manera individualizada, es decir para cada paciente, contrario a la información suministrada por la sensibilidad y la especificidad (estas últimas carecen de utilidad en la práctica clínica).

Valor Predictivo Positivo (VPP): Es la probabilidad que tiene un individuo de estar enfermo cuando el resultado de la prueba es positiva, en otras palabras, es una probabilidad condicionada a que un paciente que resulte positivo a la prueba, tenga la enfermedad. En la figura 2, la flecha señala la dirección del valor predictivo positivo, desde el resultado de la prueba positiva a la probabilidad de que este individuo esté enfermo (4). En la tabla 1 se muestra la fórmula para el cálculo del VPP.

Valor Predictivo Negativo (VPN): Contrariamente el valor predictivo negativo es la probabilidad de que un individuo que obtenga un resultado negativo a la prueba, no presente la enfermedad o esté sano. En la figura 2, la flecha señala

la dirección del VPN, que proviene de un resultado negativo de la prueba y la probabilidad de que el individuo no esté enfermo. Esta probabilidad está condicionada a que un sujeto con resultado negativo no esté enfermo o esté sano. En otras palabras, es la probabilidad de que el paciente no tenga la enfermedad frente a un resultado negativo de la prueba (2-6). En la tabla 1 se muestra el cálculo para el VPN.

Otra característica de los valores predictivos, es que dependen de la prevalencia de la enfermedad en una población. Una prueba con una sensibilidad y una especificidad dada, puede tener diferentes valores predictivos en diferentes poblaciones de pacientes con diferente prevalencia de la enfermedad. Si una prueba es utilizada en una población con un alta prevalencia de la enfermedad, la prueba en evaluación tendrá un alto VPP, por otro lado, esta misma prueba, con la misma sensibilidad y especificidad, tendrá un VPP bajo, cuando es utilizada en una población con una baja prevalencia de la enfermedad. Un ejemplo sería cuando se realiza una prueba que es considerada útil para detectar sangre oculta en heces. Un resultado positivo en personas ancianas es predictivo de cáncer de colon, hecho contrario sí se compara con una población de personas jóvenes menores de 20 años (2,5-6).

Razón de Verosimilitud (RVP y RVN) Positiva y Negativa: También llamada “Likelihood Ratio” o razón de verosimilitud o índice de eficiencia pronóstica (IEP). Este indicador tiene como característica que es un índice fijo, debido a que se usa cuando la prueba diagnóstica no tiene resultados dicotómicos, sino un umbral o puntos de cortes, como por ejemplo el valor de la glicemia. El cálculo de la razón de verosimilitud, es otra manera o método de valorar la exactitud de una prueba en el terreno clínico. Ofrece la ventaja sobre los otros indicadores, de que es independiente de la prevalencia de la enfermedad en una población. La razón de verosimilitud indica que un resultado de una prueba diagnóstica, elevará o reducirá la probabilidad de tener la enfermedad, es decir es relativo a la probabilidad previa de la enfermedad (prevalencia); en otras palabras, es una razón y no una proporción, y significa: cuál es la probabilidad de tener o no la enfermedad. Cada prueba diagnóstica está caracterizada por dos razones de verosimilitud: la razón de verosimilitud positiva o cociente de probabilidad positiva y la razón de verosimilitud negativa o cociente de probabilidad negativa, estas se describen a continuación.

Razón de Verosimilitud Positiva (RVP) o Cociente de Probabilidad Positiva (CPP): se calcula dividiendo la probabilidad de un resultado positivo en los pacientes enfermos entre la probabilidad de un resultado positivo en los individuos sanos. Es, en definitiva, el cociente entre la fracción de verdaderos positivos (sensibilidad) y la fracción de falsos positivos (1-especificidad), y nos indica la razón de enfermedad o la probabilidad de tener la enfermedad si el resultado es positivo (4, 5), la fórmula se describe en la tabla 1 (7).

Razón de Verosimilitud Negativa (RVN) o Cociente de Probabilidad Negativo (CPN): se calcula dividiendo la probabilidad de un resultado negativo en presencia de enfermedad, entre la probabilidad de un resultado negativo en ausencia de la misma. Por lo tanto, es el cociente entre la fracción de falsos negativos (1-sensibilidad) y la fracción de verdaderos negativos (especificidad). La RVN indica la probabilidad de que un individuo enfermo, obtenga un resultado negativo, en relación con la que un individuo no enfermo o sano, obtenga un resultado negativo. En otras palabras, este indicador muestra, que es poco probable que en un paciente enfermo la prueba resulte negativa, con respecto a un paciente sano, con el mismo resultado negativo (ver tabla 1). Así, una razón de verosimilitud positiva mayor de 1, indica que existe una probabilidad elevada de que el individuo

tenga la enfermedad, y mientras más elevado sea el valor de razón obtenida, mayor será la probabilidad de tener la enfermedad. Inversamente, una razón de verosimilitud negativa menor de 1 disminuirá la probabilidad de que el individuo tenga la enfermedad en estudio (7), la formula se describe en la tabla 1. En la tabla 2 se muestran los valores que pueden ser útiles para orientarnos sobre la capacidad de una prueba a través de las razones de verosimilitud, expresadas como valores de índices de eficiencia pronósticas (8-13).

Tabla 2. Valores que determinan la eficiencia diagnostica de una prueba.

Valores IEP	Capacidad
IEP(+) ≥ 10 ó IEP(-) ≤ 0.1	Suficiente
IEP(+) $\geq 5 < 10$ ó IEP (-) $> 0.1 \leq 0.2$	Moderada
IEP(+) $\geq 2 < 5$ ó IEP (-) $> 0.2 \leq 0.5$	Escasa
IEP(+) $\geq 1 < 2$ ó IEP (-) $> 0.5 < 1$	Insignificante

Teorema de Bayes

Esta herramienta toma en cuenta la probabilidad previa o prevalencia de la enfermedad para el cálculo de la probabilidad posterior a la prueba. Dado que las pruebas clínicas no son perfectas, no pueden ser utilizadas correctamente si no se estima la probabilidad previa, o prevalencia del daño que el paciente estudiado presenta (14).

Probabilidades pretest y postest y el teorema de Bayes: en el contexto de la evaluación de medios diagnósticos, se tienen los elementos siguientes (14):

- La probabilidad pretest denominada también prevalencia de la enfermedad, se denota como P (Enf) o P.
- La probabilidad pretest o prevalencia de no enfermos, se denota como P(No Enf) o Q=1-P.
- La sensibilidad conocida, que puede denotarse como P(T+/Enf), y que representa la Probabilidad de que el test resulte positivo, dado que el paciente tiene la enfermedad.
- La especificidad conocida de un test, que puede denotarse como P(T-/No Enf), o la probabilidad

de que el test resulte negativo dado que el paciente no tiene la enfermedad.

- El Valor Predictivo Positivo o Probabilidad Postest Positiva de una Prueba se escribe como P (Enf/T+), que representa la probabilidad de estar enfermo, debido a que el paciente tiene un resultado positivo en la prueba diagnóstica.

- El Valor Predictivo Negativo o Probabilidad Postest Negativa de una Prueba se escribe como P (No Enf/T-), y representa la probabilidad de no estar enfermo, dado que el paciente tiene un resultado negativo en la prueba diagnóstica

El cálculo de la Probabilidad Postest Positiva o Negativa o Valor Predictivo Positivo o Negativo se muestra en la tabla 3.

Curvas ROC

Hasta ahora se ha discutido la posibilidad de que en una prueba se obtengan dos resultados: positivo o negativo, pero a menudo los resultados de una prueba son en variables continuas, tales como la presión arterial o el nivel de glicemia, no tienen un resultado positivo y otro negativo, sino datos en escala continua. La decisión debe ser, por lo tanto, establecida con un nivel umbral o “cutt-off” encima del cual el resultado de la prueba es considerado como positivo, y por debajo de éste es considerado negativo. Por ejemplo, no hay un nivel de glicemia que claramente separe a los enfermos de los que no lo están, existen valores que están solapados entre los dos grupos, es decir los que están enfermos (diabéticos) de los sanos o no enfermos (no diabéticos), para cada nivel de glicemia (15). Por

estableciendo un nivel umbral de 120 mg/dL, solo el 25% de los individuos diabéticos serán identificados como positivos (sensibilidad), mientras que el 90% de los sujetos sanos serán identificados correctamente como negativos (especificidad) (7).

Por el contrario, sí en el mismo ejemplo hipotético (figura 3b) se establece un umbral de 60 mg/dL el 85% de los individuos diabéticos serán identificados correctamente como positivos (sensibilidad), mientras que solo el 30% de los sujetos sanos serán identificados como negativos (especificidad). De esta manera, se demuestra la relación entre la sensibilidad y la especificidad, en función del punto de corte seleccionado, lo que plantea la necesidad de establecer puntos de corte óptimos en función de los valores continuos medidos, en este sentido, una herramienta que facilita el establecimiento de los puntos de corte, es la construcción de las curvas ROC.

El uso de las curvas ROC minimiza la aparición de falsos positivos, lo cual está asociado con un elevado costo emocional y económico, ya que los pacientes podrían ameritar la repetición de las pruebas, sin que sea del todo necesaria, pues posiblemente pocos de ellos estén realmente enfermos. Adicionalmente, la implementación de las curvas ROC también disminuye el número de falsos negativos. Los falsos negativos constituyen un serio problema en el manejo clínico, sobretodo cuando se trata de enfermedades graves, ya que se pierde la oportunidad de darle al paciente un tratamiento eficaz en etapas iniciales (2).

Construcción de las curvas ROC: éstas se representan en un plano cartesiano, graficando sensibilidad versus 1-especificidad (ver figura 4). En

Tabla 3. Probabilidad Postest Positiva o Negativa (ver tabla 1).

Indicador	Cálculo
Probabilidad Post-Test Positiva =	$\frac{\text{(Sensibilidad * prevalencia de la enf.)}}{\text{[(Sensibilidad * prevalencia de la enf.) + (falsos positivos * proporción de sanos)]}}$
Probabilidad Post-Test Negativa =	$\frac{\text{(Especificidad * proporción de sanos)}}{\text{[(Especificidad * proporción de sanos) + (falsos negativos * prevalencia de la enf.)]}}$

ejemplo, en un muestreo hipotético (figura 3a)

este sentido, las curvas ROC muestran todos los pares

a) Punto de Corte en 120 mg / dL		Enfermos	No enfermos	
		Diabeticos	No diabeticos	Total
Resultado de la prueba	Positivo	5	2	7
	Negativo	15	18	33
Total		20	20	40

b) Punto de Corte en 60 mg / dL		Enfermos	No enfermos	
		Diabeticos	No diabeticos	Total
Resultado de la prueba	Positivo	17	14	31
	Negativo	3	6	9
Total		20	20	40

Figura 3. Distribución hipotética de los resultados de una prueba diagnóstica en un muestreo de controles y de pacientes con diabetes. a) cutt off de 120 mg/dL, b) cutt off de 60 mg/dL.

ordenados (sensibilidad, 1-especificidad), resultantes de la variación continua de los puntos de corte en todo el rango de resultados observados. El gráfico expresará mayor poder discriminatorio de la prueba, cuanto mayor sea el área bajo la curva, que por lo general es considerado como mayor al 0,80 (15-19).

Herramientas para identificar el punto de corte más apropiado: las curvas ROC optimizan el punto de corte, es decir, cuanto más sensible y específica sea la prueba (puntos más hacia arriba y más hacia la izquierda en la gráfica), más se alejará de la diagonal, lo cual es considerado como el mejor punto de corte. Cuando se construyen las curvas ROC existen varios métodos que permiten la identificación del punto de corte ideal. Dos de ellos son el método de distancias mínimas y la determinación de los índices de Youden. En el primero, se trazan líneas que unan los distintos puntos de corte con la coordenada (1,0), que se corresponde a una sensibilidad y especificidad del 100%, la línea que tenga la menor distancia es la que esta conectada con el punto de corte ideal, que representa el mejor par (sensibilidad, 1-especificidad) (18).

Por su parte, el método de Youden requiere la determinación de un índice para cada punto de corte posible. El cual se calcula a través de la fórmula $IY=(S+E) -1$ (20-22). El punto de corte con el mayor índice corresponde al ideal.

En la figura 4 se ilustran el uso de las curvas ROC, aplicado al ejemplo hipotético de los valores de

glicemia, demostrándose que el par ordenado correspondiente a P5 constituye el punto de corte ideal calculado según los métodos descritos previamente.

Las curvas ROC son apropiadas para determinar la utilidad diagnóstica de una prueba, lo cual en la práctica se refleja con valores de área bajo la curva mayores a 0,80; permite además comparar distintas pruebas diagnósticas para una misma enfermedad cotejando sus respectivas curvas (15-19).

Pruebas Múltiples

Ya que las pruebas diagnósticas utilizadas no siempre tienen características operativas muy satisfactorias, usualmente, se recurre al uso de pruebas múltiples que pueden ser indicadas de dos maneras: en paralelo y en serie (1-4).

Pruebas en paralelo: en estas se hacen simultáneamente dos o más pruebas en el mismo individuo a fin de aumentar la sensibilidad diagnóstica. Se considera negativa la prueba en aquellos sujetos que tengan resultados negativos en todos los test y positiva en aquellos que presenten resultados positivos en al menos uno de los test (1-4).

Pruebas en serie: estas se aplican secuencialmente con la finalidad de aumentar la especificidad o dicho de otra manera, para confirmar un primer resultado positivo. Se consideran como positivos, aquellos sujetos que tengan resultados

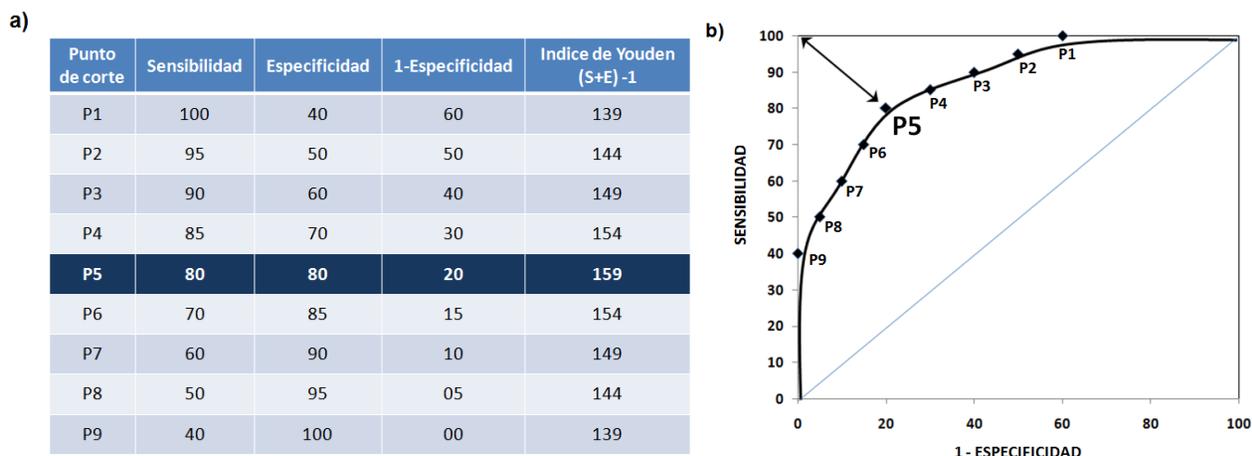


Figura 4. Interpretación de las curvas ROC. a) Donde se muestran todos los pares ordenados resultantes de la variación continua de los puntos de cortes en todo el rango observado (sensibilidad, 1-especificidad) donde el valor P5 es el sitio de mejor punto de corte con 80 % de sensibilidad y 80 % de especificidad como lo muestra la raya en la Gráfica b

positivos en todas las pruebas y negativos los que presenten resultados negativos en las pruebas confirmatorias (1-4).

Confiabilidad o repetitividad de una prueba

El concepto de confiabilidad o reproducibilidad implica la cantidad de error que se comete al realizar cualquier medida (por ejemplo la determinación de la presión arterial).

Existen varios tipos de confiabilidad, incluyendo la confiabilidad inter e intra-sujeto.

En la práctica diagnóstica es común dudar acerca de la confiabilidad o repetibilidad de una prueba independientemente de su sensibilidad o especificidad. Si un resultado no es reproducible, el valor y la utilidad de la prueba son pobres. Existen tres tipos de repetitividad (1-4):

Variación intra-sujeto: se refiere a la variación implícita en las características que se están determinando, las cuales, a menudo, varían en el tiempo, aún en intervalos de tiempo cortos. Por ejemplo, las cifras de presión arterial varían incluso en un período de 24 horas y dependiendo de la actividad realizada por el sujeto en el momento de realizar la medición. Por lo tanto, al evaluar el resultado de una prueba, es importante considerar las condiciones y el momento en que se realizó la misma (1-4).

Variación intra-observador: ocurre entre dos observaciones realizadas por el mismo experimentador. Por ejemplo, la lectura de una

radiografía en dos ocasiones distintas, la cual probablemente puede ser reportada, en ambas oportunidades, con observaciones diferentes. Aquí se debe tomar en cuenta el grado de factor subjetivo del observador que realiza el análisis y sus conclusiones; por lo tanto, a mayor cantidad de elementos subjetivos que intervengan en la lectura, se presentará mayor variación intra-observador (1-4).

Variación inter-observador: dos observadores diferentes pueden reportar el resultado de una prueba de manera distinta. Es relevante, entonces, lograr un consenso entre los observadores, para ello, una herramienta útil para unificar los criterios de observación es la transformación a término cuantitativos de las apreciaciones de los observadores. Existen diversos modelos matemáticos que permiten solventar estas discordancias, como por ejemplo, el índice de kappa, cuyos valores son directamente proporcionales a la confiabilidad de las observaciones del test (1-4).

Síntesis

Al momento de evaluar cualquier prueba diagnóstica es indispensable considerar las propiedades intrínsecas de la misma, tales como, la sensibilidad y la especificidad. Por su parte, los VPN y VPP tienen mayor aplicabilidad en la práctica clínica, ya que determinan la utilidad del test para el diagnóstico de una determinada patología.

Las curvas ROC son instrumentos útiles para la evaluación de las pruebas diagnósticas ya que permite

discriminar la capacidad de un test para diferenciar entre dos estados alternativos de salud, enfermedad o no enfermedad

interpretación o aplicación errada de la misma puede tener consecuencias importantes en el estado de salud de un individuo que es sometido a este test.

Todos estos aspectos deben ser considerados cuando se evalúa la utilidad de una prueba ya que una

Referencias

1. Gordis L. Epidemiology. Third Edition Ed. Philadelphia: Elsevier Saunders, 2004.
2. Israni R. Medpage Tools Guide to Biostatistics. Medpage today 2010; Com: 1-11. [\[Google Scholar\]](#)
3. Fernandez PS, Diaz PS. Pruebas Diagnosticas. Atención Primaria en la Red 2003; 10:120-4. [\[Google Scholar\]](#)
4. Ruiz Morales A, Morriño Zarate L. Epidemiología Clínica Investigación Aplicada. Bogotá DC Colombia: Editorial Médica Panamericana, 2004.
5. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. JAMA 1994; 271:703-7. [\[Google Scholar\]](#) [\[PubMed\]](#)
6. Grimes DA, Schulz KF. Uses and abuses of screening tests. Lancet 2002; 359:881-4. [\[PubMed\]](#) [\[Google Scholar\]](#)
7. Grimes DA, Schulz KF. Refining clinical diagnosis with likelihood ratios. Lancet 2005; 365:1500-5. [\[PubMed\]](#) [\[Google Scholar\]](#)
8. Harper R, Reeves B. Reporting of precision of estimates for diagnostic accuracy: a review. BMJ 1999; 318:1322-3. [\[PubMed\]](#) [\[Google Scholar\]](#)
9. Molinero LM. Valoración de las Pruebas Diagnosticas 1-12: 2002. www.she-lilha.org/stat1.html. [\[Google Scholar\]](#)
10. Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. BMJ 1994; 308:1552. [\[PubMed\]](#) [\[Google Scholar\]](#)
11. Altman DG, Bland JM. Diagnostic tests 2: Predictive values. BMJ 1994; 309:102. [\[PubMed\]](#) [\[Google Scholar\]](#)
12. Centre for Evidence-Bases Medicine (CEBM). Likelihood Ratios, 1-3 2009. www.cebm.net/scrating-post.as [\[Google Scholar\]](#)
13. Loong TW. Understanding sensitivity and specificity with the right side of the brain. BMJ 2003; 327:716-9. [\[PubMed\]](#) [\[Google Scholar\]](#)
14. Díaz FR, López Barón JF. Probabilidad. Universidad de Málaga, Málaga, Spain: Thompson Editores, 2005.
15. Altman DG, Bland JM. Diagnostic tests 3: receiver operating characteristic plots. BMJ 1994; 309:188. [\[PubMed\]](#) [\[Google Scholar\]](#)
16. Obuchowski NA. Sample size tables for receiver operating characteristic studies. AJR Am J Roentgenol 2000; 175:603-8. [\[PubMed\]](#) [\[Google Scholar\]](#)
17. Dwyer AJ. In pursuit of a piece of the ROC. Radiology 1996; 201:621-5. [\[PubMed\]](#) [\[Google Scholar\]](#)
18. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993; 39:561-77. [\[PubMed\]](#) [\[Google Scholar\]](#)
19. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. Radiology 2003; 229:3-8. [\[PubMed\]](#) [\[Google Scholar\]](#)
20. Bohning D, Bohning W, Holling H. Revisiting Youden's index as a useful measure of the misclassification error in meta-analysis of diagnostic studies. Stat Methods Med Res 2008; 17:543-54. [\[PubMed\]](#) [\[Google Scholar\]](#)
21. Perkins NJ, Schisterman EF. The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. Am J Epidemiol 2006; 163:670-5. [\[PubMed\]](#) [\[Google Scholar\]](#)
22. Supo J. Videos de Clase de Análisis de datos Categóricos en la web: www.seminariosdeinvestigacion.com 2011

Como citar este artículo: Donis JH, Evaluación de la validez y confiabilidad de una prueba diagnóstica. *Avan Biomed* 2012; 1: 73-81