

CÓMO INTERPRETAR UN ARTÍCULO SOBRE PRUEBAS DIAGNÓSTICAS

HOW TO INTERPRET A DIAGNOSIS ARTICLE

DR. CARLOS MANTEROLA D. (1)

1. DEPARTAMENTO DE CIRUGÍA Y TRAUMATOLOGÍA, CAPACITACIÓN, INVESTIGACIÓN Y GESTIÓN PARA LA SALUD BASADA EN EVIDENCIA (CIGES) UNIVERSIDAD DE LA FRONTERA. CENTRO COLABORADOR CHILENO UFRO DE LA RED COCHRANE IBEROAMERICANA.
cmantero@ufro.cl

FINANCIADO PARCIALMENTE POR PROYECTO DI09-0060 DE LA DIRECCIÓN DE INVESTIGACIÓN DE LA UFRO.

RESUMEN

Uno de los hechos más frecuentes en la práctica clínica cotidiana es decidir cuándo una prueba diagnóstica es normal o anormal; y qué significado representa este resultado para el paciente en cuestión; pues de eso depende muchas veces la indicación, corrección o suspensión de un tratamiento; la indicación de un procedimiento quirúrgico; e incluso el pronóstico de un paciente.

Todo lo anteriormente expuesto, pasa por la conducción y adecuada interpretación de estudios de pruebas diagnósticas, para los cuales, es indispensable conocer o definir el estándar de referencia o "gold standard", que se utilizará para comparar la prueba en estudio; escoger dos grupos de sujetos a estudio (uno con y otro sin el evento de interés a estudiar); categorizar a los sujetos en estudio como positivos o negativos para el evento de interés en estudio; y construir tablas de contingencia para el cálculo ulterior de los valores de validez (sensibilidad y especificidad) y de seguridad (valores predictivos) de la prueba; con los que posteriormente se podrá determinar las razones de probabilidad; construir una curva ROC; estimar el área bajo la curva y los puntos de corte en relación a la mejor capacidad de discriminar de la prueba diagnóstica en estudio.

Palabras clave: Estudios observacionales, estudios analíticos, estudios de corte transversal, pruebas diagnósticas, estudios

de pruebas diagnósticas, errores diagnósticos, sensibilidad y especificidad, valores predictivos, curvas ROC.

SUMMARY

One of the most frequent clinical problems is to decide if a diagnostic test is normal or abnormal, because this often depends on the indication, correction or suspension of a treatment, the indication of a surgical procedure, or even a patient's prognosis.

But this requires the design and conduction of diagnostic tests studies, for which, it is indispensable to know or define the "gold standard" to compare the test under study; to choose two groups of subjects to study (one with and one without the event of interest or in study disease), to categorize the study subjects as positive or negative for the event of interest in study and construct contingency tables for the calculation of subsequent values of validity (sensitivity and specificity) and security (predictive values) of the test. Later, they may lay the likelihood ratios estimation, to construct a ROC curve, estimate the under the curve area and the cut-off points in relation to the best ability to discriminate the test under study.

Key words: "Epidemiologic Research Design"[Mesh], "Diagnostic Techniques and Procedures"[MeSH], "Cross-Sectional Studies"[Mesh], "Diagnosis" [MeSH], "Diagnostic Errors"[Mesh], "Sensitivity and Specificity"[MeSH], "Predictive Value of Tests"[Mesh], "ROC Curve"[Mesh].

INTRODUCCIÓN

En la publicación anterior se hizo mención a algunos de los diseños que se encuentran agrupados bajo la denominación “estudios observacionales”, entre los cuales, en el subgrupo de estudios analíticos se mencionaron los estudios de pruebas diagnósticas (PD) (1). En esta ocasión se tratarán los aspectos más relevantes para interpretar de forma apropiada un estudio de PD.

Una variable de gran relevancia al momento de solicitar una PD es la valoración de la probabilidad que el sujeto en estudio presente el evento de interés o enfermedad que se investiga (denominado también “probabilidad pre-test”). Al respecto, los clínicos en general empleamos un proceso intuitivo de clasificación de los individuos que consultan en subgrupos de alta, intermedia o baja probabilidad de presentar el evento de interés en estudio; estimaciones que se asocian a la experiencia personal de cada cual y a lo reportado en la literatura. Por ende, se ha de entender que a mayor “probabilidad pre-test”, mayor será el rendimiento de la PD; y por ende, que la prevalencia de la enfermedad o evento de interés en estudio, influye directamente en la “probabilidad pre-test”.

Es así como hace algunos años atrás, un grupo de investigadores de la Universidad de Yale, seleccionaron una muestra probabilística estratificada de clínicos norteamericanos de diferentes especialidades, a los que aplicaron una encuesta referente a su actitud en relación a la interpretación de artículos de PD. La encuesta fue respondida por 300 médicos, con una media de edad de 46 años, los que pasaban una mediana de 90% de su tiempo profesional en actividades asistenciales. El principal resultado fue que muy pocos utilizaban métodos formales de valoración de la precisión de una PD (Tabla 1). Y, a pesar que el 84% reconoció utilizar la sensibilidad y especificidad al interpretar los resultados de una PD, la mayoría de las veces lo hacían informalmente. A partir de lo cual se concluyó que la información sobre la exactitud de las PD debe estar disponible de forma instantánea cuando se solicita una; que la enseñanza formal en esta materia debe mejorarse; y que la información publicada al respecto en su mayor parte es inútil, porque fracasa en reflejar la población de sujetos en los que se aplicaron las PD (2).

No es ningún secreto que para una buena parte de los médicos, términos como sensibilidad, especificidad, valores predictivos, curvas operador receptor (ROC), etc., constituyen conceptos abstractos; tanto así que para algunos representan incluso atentados contra la lengua castellana. Sin embargo, como clínicos necesitamos entender estos conceptos, pues de este modo podremos relacionarnos mejor con quienes realizan PD (desde el laboratorio básico hasta la generación de las imágenes más sofisticadas), entender mejor su real utilidad, y explicar mejor a nuestros pacientes por qué se requiere hacer tal o cual prueba y qué significa tal o cual resultado en su condición en particular.

Uno de los hechos más frecuentes en clínica es decidir cuándo una PD es normal o anormal, pues de esto en buenas cuentas depende muchas veces la instauración o suspensión de un tratamiento, la in-

dicación de una cirugía, o incluso el pronóstico de un paciente. Pero, las definiciones de normalidad son distintas según el punto de vista con que se analice el problema. Si lo enfocamos según la distribución normal de Gauss, se considerará como normal a aquella medida que represente al promedio o media \pm dos desviaciones estándar (lo que supone asumir previamente que la distribución es normal y que todas las anomalías tienen la misma frecuencia de aparición). Otra forma de enfocar la normalidad es con el concepto de factor de riesgo, lo que indica que el resultado obtenido en la prueba diagnóstica no comporta riesgo adicional; esto permite identificar a los pacientes que se comportan en forma extrema (denominados también “outliers”), a los que quizás no puede ofrecerse ningún tipo de prevención o tratamiento. Sin embargo, la aceptación más empleada en relación con las PD es la que considera normal al intervalo de resultados fuera del cual el evento de interés o enfermedad objeto de estudio resulta altamente probable.

Parece razonable que para el reporte referente a los resultados obtenidos con la aplicación de PD, debieran considerarse los siguientes: valores de validez y seguridad de la prueba, las razones de probabilidad, el área bajo la curva y los puntos de corte en relación a la mejor capacidad de discriminar de la prueba en estudio. Por otro lado, es relevante considerar la información de los intervalos de confianza para cada una de estas medidas (3).

TABLA 1. FRECUENCIA DE UTILIZACIÓN DE MÉTODOS DE VALORACIÓN DE EXACTITUD DE PD (N = 300)

Tipo de clínicos encuestados	Método Bayesiano	Curvas ROC	Razones de probabilidad
Médicos especialistas (N = 50)	5	1	1
Médicos generales (N = 50)	2	0	1
Pediatras (N = 50)	1	1	0
Cirujanos generales (N = 50)	0	1	0
Médicos de familia (N = 50)	0	0	0
Obstetras/Ginecólogos (N = 50)	0	0	0
Porcentaje global (N = 300)	3,0	1,0	1,0

Se entiende como “Médicos especialistas” aquellos con subespecialidades médicas.

Al respecto existe la iniciativa STARD, que fue generada para mejorar la precisión en el reporte de estudios de exactitud diagnóstica, de modo tal de permitir a los lectores evaluar la posibilidad de sesgo en un estudio de esta naturaleza y valorar por ende la capacidad de generalización de éste (4).

EFICACIA DE UNA PRUEBA DIAGNÓSTICA

Una de las características fundamentales de una PD es su poder discriminatorio, y éste, tiene relación con la variabilidad de la prueba, la reproducibilidad de los hallazgos, y la variabilidad de la población sana, o la determinación de los rangos de valores normales de la prueba en cuestión en esa población. Los criterios de evaluación de una PD que con mayor frecuencia se utilizan son sensibilidad, especificidad y valores predictivos. Estos, proporcionan información acerca de la capacidad de discriminación de la prueba, son de utilidad para comparar el estado de una prueba diagnóstica, debieran permitir obtener los mismos resultados cuando se aplica en diferentes grados de enfermedad, y constituyen marcadores de la proporción de enfermos y no enfermos que son clasificados correctamente.

Mención aparte requiere el concepto de "estándar de referencia", que representa la prueba diagnóstica más cercana a la veracidad del fenómeno en estudio, de la que se puede disponer en un momento determinado (5); y que será utilizada para comparar con las características de la PD en estudio.

Para calcular sensibilidad, especificidad y valores predictivos de una PD se deben seguir los siguientes pasos (6):

1. Definir el estándar de referencia o "gold standard", es decir cuál es hasta el momento del estudio, la mejor alternativa diagnóstica existente para estudiar una determinada enfermedad o evento de interés en términos de sensibilidad, especificidad y valores predictivos; por ende la mejor opción para identificar sujetos con y sin dicha enfermedad o evento de interés.

2. Escoger dos grupos de sujetos a estudio. Uno que presente la enfermedad o evento de interés (idealmente en distintos estados evolutivos de la enfermedad en estudio) y otro que no lo tenga, definido esto por el estándar de referencia. Es decir, a ambos grupos se les ha de aplicar el estándar de referencia y la PD en evaluación.

3. Categorizar a los individuos en estudio como positivos o negativos para la enfermedad o evento de interés en estudio, según la prueba diagnóstica en evaluación; es decir, clasificar cada paciente como enfermo o libre de la enfermedad en estudio.

4. Aplicar la definición de sensibilidad y especificidad, y finalmente calcular los valores correspondientes, para lo cual es necesario construir tablas de 2 x 2 ó de contingencia (Figuras 1 y 2).

VALIDEZ DE UNA PRUEBA DIAGNÓSTICA

SENSIBILIDAD Y ESPECIFICIDAD

Es el grado en que una prueba diagnóstica mide lo que se supone que debe medir (7). Se puede aclarar a través de la siguiente pregunta: ¿con qué frecuencia el resultado de una prueba diagnóstica es confirmado por procedimientos diagnósticos más complejos y rigurosos?

Sensibilidad

La sensibilidad corresponde a la proporción de aquellos sujetos que, teniendo la enfermedad o evento de interés en estudio definida por el estándar de referencia, ésta es identificada por la prueba diagnóstica en evaluación; es decir se relaciona con el concepto de "positividad para enfermedad o evento de interés". La sensibilidad, responde a la pregunta: ¿si el paciente tiene realmente la enfermedad, cuál es la probabilidad de que la prueba empleada sea positiva? Dicho de otra forma, la sensibilidad es la probabilidad de clasificar correctamente a un individuo cuyo estado real sea el definido como positivo respecto a la condición que estudia la prueba, razón por la que también es denominada fracción de verdaderos positivos (FVP) (7, 8).

FIGURA 1.

Resultado de la prueba en estudio	Estado respecto a la enfermedad según el estándar de referencia	
	Presente	Ausente
Positivo	a (enfermos con prueba +)	b (no enfermos con prueba +)
Negativo	c (enfermos con prueba -)	d (no enfermos con prueba -)

Figura 1. Tabla de contingencia o de 2 x 2 primaria en la que se explica la generación posterior de las celdas con las que se generan los cálculos de sensibilidad, especificidad y valores predictivos.

FIGURA 2.

Resultado de la prueba en estudio	Estado respecto a la enfermedad según el estándar de referencia	
	Enfermo	Sano
Positivo	Verdadero positivo (VP)	Falso positivo (FP)
Negativo	Falso negativo (FN)	Verdadero negativo (VN)

Figura 2. Tabla de contingencia o de 2 x 2 en la que se explica la generación de los conceptos de VP, VN, FP y FN.

De ello se desprende que para calcular la sensibilidad de una prueba diagnóstica se ha de dividir el número de enfermos con prueba positiva por la sumatoria de los enfermos con prueba positiva y los enfermos con prueba negativa; es decir $a / (a + c)$; ó $VP / VP + FN$. Ver figuras 1 y 2.

De lo anteriormente expuesto, se puede resumir que una prueba diagnóstica de alta sensibilidad es útil en contextos clínicos donde el hecho de no diagnosticar genera más problemas que el exceso de diagnóstico. Es el caso de tamizaje o "screening", que se realiza aplicando una PD que otorgue resultados validos y confiables, que sea de bajo costo, de fácil realización y mínima incomodidad para el usuario.

Especificidad

Por su parte, la especificidad corresponde a la proporción de sujetos libres de la enfermedad o evento de interés en estudio, definida por el estándar de referencia, a los que la prueba diagnóstica en evaluación identifica como no enfermos o sin el evento de interés en estudio; es decir se relaciona con el concepto de "negatividad para enfermedad". La especificidad responde a la pregunta: ¿si el paciente no tiene la enfermedad, cuál es la probabilidad de que la prueba sea negativa? Dicho de otra forma, la especificidad es la probabilidad de clasificar correctamente a un individuo cuyo estado real sea el definido como negativo. Es igual al resultado de restar a uno la fracción de falsos positivos (FFP) (7, 8).

De ello se desprende que para calcular la especificidad de una prueba diagnóstica se ha de dividir el número de sujetos "no enfermos" con prueba positiva por la sumatoria de los sujetos "no enfermos" con prueba positiva y los sujetos "no enfermos" con prueba negativa; es decir $d / (b + d)$; ó $VN / FP + VN$. Ver figuras 1 y 2.

De lo anteriormente expuesto, se puede deducir que una PD de alta especificidad es útil para confirmar o descartar una enfermedad o evento de interés. Un ejemplo es la utilización de pruebas más sofisticadas o "pruebas de confirmación"; cuya principal aplicación es confirmar o descartar una enfermedad, debido a las implicancias diagnósticas de esta.

Como se entenderá entonces, existe una estrecha relación entre sensibilidad y especificidad. Ésta, generalmente es de tipo inversa, es decir que para cada resultado específico de una PD expresado en una escala de tipo continuo, la sensibilidad puede incrementarse solamente a expensas de la especificidad. Otra forma de describirlo es a través de las denominadas curvas ROC ("receiver operator characteristic"); concepto que será tratado en párrafos posteriores.

LA SEGURIDAD DE UNA PRUEBA DIAGNÓSTICA LOS VALORES PREDITIVOS

Los conceptos de sensibilidad y especificidad permiten, por lo tanto, valorar la validez de una prueba diagnóstica; sin embargo, carecen de utilidad en la práctica clínica. Tanto la sensibilidad como la especificidad

proporcionan información acerca de la probabilidad de obtener un resultado concreto (positivo o negativo) en función de la verdadera condición del enfermo con respecto a la enfermedad (7). Sin embargo, cuando a un paciente se le realiza alguna prueba, el médico carece de información "a priori" acerca de su verdadero diagnóstico, y más bien la pregunta se plantea en sentido contrario: ante un resultado positivo o negativo en la prueba, ¿cuál es la probabilidad de que el paciente esté realmente enfermo o "no enfermo"? Así pues, resulta obvio que hasta el momento sólo se ha abordado el problema en una dirección. Por medio de la estimación de los valores predictivos completaremos esta información.

Valor predictivo positivo

El valor predictivo positivo (VPP) de una prueba diagnóstica corresponde a la proporción de individuos con una prueba positiva para una enfermedad o evento de interés determinado, que están realmente enfermos de ella. El VPP de una prueba se puede explicar con el siguiente escenario: si el resultado de una PD es positivo ¿qué probabilidad tiene el paciente de presentar la enfermedad en estudio?

Dicho de otra manera, el VPP de una PD se define por la proporción de resultados positivos de la prueba diagnóstica en quienes la enfermedad está presente (confirmada por el estándar de referencia); o, por la probabilidad de que un paciente sea un VP teniendo el resultado de la PD positiva; o, por la probabilidad de tener la enfermedad dado que el resultado de la PD fue positiva (9).

De esto se desprende que para calcular el VPP de una PD se ha de dividir el número enfermos con prueba positiva por la sumatoria de los enfermos con prueba positiva y los sujetos "no enfermos" con prueba positiva; es decir $a / (a + b)$; ó $VP / VP + FP$. Ver figuras 2 y 3.

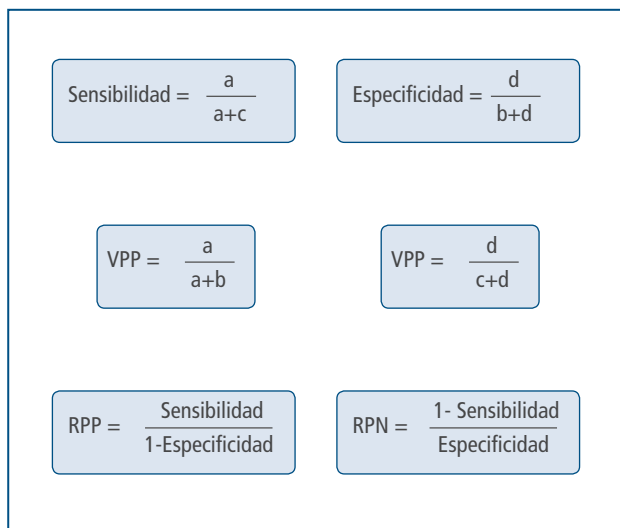
Valor predictivo negativo

Por su parte, el valor predictivo negativo (VPN) de una prueba diagnóstica corresponde a la proporción de individuos con una prueba negativa para una determinada enfermedad o evento de interés en estudio, que no tienen la enfermedad o evento de interés. Así, este concepto se podría explicar con el siguiente escenario: si el resultado de una prueba es negativo, ¿cuál es la probabilidad que tiene el paciente de no presentar la enfermedad en estudio?

Dicho de otra manera, el VPN de una PD se define por la proporción de personas con resultados negativos de la PD en quienes la enfermedad está ausente; por la probabilidad de que el paciente sea un VN teniendo una PD negativa; y por la probabilidad de que la enfermedad esté ausente dado un resultado negativo de la PD (9).

De lo anteriormente expuesto, se puede colegir que para calcular el VPN de una PD se ha de dividir el número enfermos con prueba negativa por la sumatoria de los enfermos con prueba negativa y los sujetos "no enfermos" con prueba negativa; es decir $d / (c + d)$; ó $VN / FN + VN$. Ver figuras 2 y 3.

FIGURA 3. FÓRMULAS PARA LA REALIZACIÓN DE LOS CÁLCULOS DE SENSIBILIDAD, ESPECIFICIDAD, VALORES PREDITIVOS Y RAZONES DE PROBABILIDAD



Es importante entonces recalcar que al igual que para el VPP, el VPN también depende de la sensibilidad y especificidad de la PD y de la prevalencia del fenómeno en evaluación en la población en estudio.

LA INFLUENCIA DE LA PREVALENCIA DE LA ENFERMEDAD O EVENTO DE INTERÉS EN ESTUDIO EN EL COMPORTAMIENTO DE UNA PD

Los valores de sensibilidad y especificidad, a pesar de definir completamente la validez de la PD, presentan la desventaja de que no proporcionan información relevante a la hora de tomar una decisión clínica ante un determinado resultado de la prueba. Sin embargo, tienen la ventaja adicional de que son propiedades intrínsecas a la PD y definen su validez independientemente de cuál sea la prevalencia de la enfermedad en la población a la cual se aplica (probabilidad pre-test o probabilidad estimada antes de la aplicación de la PD).

Por el contrario, el concepto de valor predictivo, a pesar de ser de enorme utilidad a la hora de tomar decisiones clínicas y transmitir a los pacientes información sobre su diagnóstico, presenta la limitación que depende en gran medida de lo frecuente que sea la enfermedad o el evento de interés a diagnosticar en la población objeto de estudio.

Es decir, cuando la prevalencia de una enfermedad o evento de interés es baja, un resultado negativo permitirá descartar la enfermedad con mayor seguridad, siendo así el VPN mayor. Por el contrario, un resultado positivo no permitirá confirmar el diagnóstico, resultando en un bajo VPP (5, 7).

LAS RAZONES DE PROBABILIDAD

Queda claro entonces cómo la prevalencia del evento de interés en estudio puede influir en los valores predictivos de una PD. Por lo tanto, éstas, no pueden ser utilizadas como índices a la hora de comparar dos métodos diagnósticos diferentes, ni tampoco a la hora de extrapolar los resultados de otros estudios en datos propios de cada clínico o de un centro en particular. Por ello, resulta necesario determinar otros índices de valoración que sean a la vez clínicamente útiles y no dependan de la prevalencia de la enfermedad en la población a estudiar. Así, además de los conceptos de sensibilidad, especificidad y valores predictivos, se suele hablar del concepto de razón de probabilidad, de verosimilitud o cociente de probabilidades; o "likelihood ratios". Estos, miden cuánto más probable es un resultado concreto (positivo o negativo) según la presencia o ausencia de enfermedad. Calcular las razones de probabilidad permite conocer mayor precisión en la información de una PD (10).

Mediante la combinación de la sensibilidad y especificidad de una PD se pueden obtener resultados más confiables sobre la validez de esta. La razón de probabilidad puede ser positiva o negativa. La razón de probabilidad nos indica cuánto más probable es un resultado determinado de una PD en un paciente con una enfermedad dada comparado con un paciente sin la enfermedad o evento de interés.

La razón de probabilidad ofrece la ventaja de relacionar la sensibilidad y la especificidad de una prueba en un solo índice. Además, pueden obtenerse razones de probabilidad según varios niveles de una nueva medida y no es necesario expresar la información de forma dicotómica, como resultado de normal o anormal o bien positivo y negativo (Tabla 2). Por último, al igual que sucede con la sensibilidad y la especificidad, no varía con la prevalencia. Esto permite utilizarlas como índices de comparación entre diferentes pruebas para un mismo diagnóstico y se mantienen constantes aunque la prevalencia de la enfermedad varíe en los sujetos en quienes se aplica la prueba (10).

Por otra parte, la razón de probabilidad es particularmente útil para el clínico debido a que le permite un mejor entendimiento de los resultados de una prueba ya que puede entender con qué fuerza el resultado positivo de una PD indica la presencia real de la enfermedad y la fuerza de un resultado negativo para descartar la enfermedad. En otras palabras, la razón de probabilidades nos indicará, como el resultado de una prueba hará cambiar la probabilidad pre-test a la probabilidad post-test de la enfermedad.

Razón de probabilidad positiva

La razón de probabilidad positiva (RPP) o "likelihood ratio positivo" de una PD describe la probabilidad de tener la enfermedad en oposición a no tenerla, teniendo un resultado positivo de la prueba en estudio. Corresponde a la relación entre el porcentaje de enfermos que presentan una prueba diagnóstica positiva y el porcentaje de "no enfermos" que presentan una prueba diagnóstica positiva.

Se calcula dividiendo la probabilidad de un resultado positivo en los

pacientes enfermos entre la probabilidad de un resultado positivo entre los sanos. Es, en definitiva, el cociente entre la fracción de VP (sensibilidad) y la fracción de FP (1-especificidad); o la relación entre la sensibilidad y el complemento de la especificidad (Figura 3).

Razón de probabilidad negativa

La razón de probabilidad negativa (RPN) o "likelihood ratio negativo" de una PD describe la probabilidad de no tener la enfermedad en oposición a tenerla, teniendo un resultado negativo de la prueba en evaluación.

Corresponde a la relación entre el porcentaje de enfermos que presentan una PD negativa y el porcentaje de no enfermos que presentan una PD negativa.

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. Es decir, corresponde al cociente entre la fracción de falsos negativos (1-sensibilidad) y la fracción de verdaderos negativos (especificidad); o en otras palabras, constituye la relación entre el complemento de la sensibilidad y la especificidad (Figura 3).

CURVAS ROC (RECEIVER OPERATING CHARACTERISTIC)

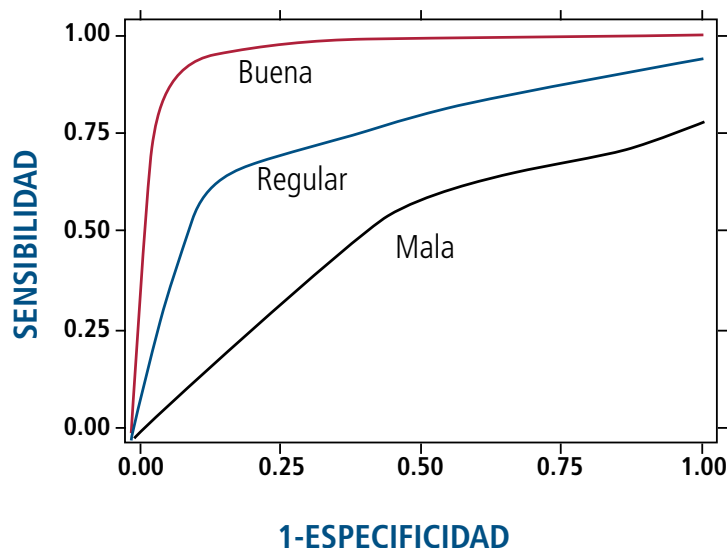
La limitación principal del enfoque hasta ahora expuesto consiste en que habitualmente se piensa en que las PD nos dan una respuesta de tipo dicotómico, es decir positiva o negativa. Esta situación obviamente dejaría al margen a una gran cantidad de PD cuyos resultados se miden en una escala continua o, al menos, discreta ordinal. Por

ejemplo, las evaluaciones de función renal y hepática a través de la exactitud de la creatinina, la bilirrubina y las transaminasas respectivamente; o de inmunohistoquímica, en que los resultados se expresan en porcentajes de presencia de la reacción.

Entonces, la generalización a estas situaciones se consigue mediante la elección de "puntos de corte" que permitan una clasificación dicotómica de los valores de la prueba según sean superiores o inferiores al valor elegido. La diferencia esencial con el caso más simple es que ahora contaremos con algo más que un único par de valores de sensibilidad y especificidad que definan la exactitud de la prueba; sino que con un conjunto de pares correspondientes cada uno a un nivel de decisión distinto. Este procedimiento constituye la esencia del análisis ROC, una metodología desarrollada en el seno de la teoría de la decisión en la década de los 50 (11).

De este modo, mediante la representación de los pares (1-especificidad y sensibilidad) obtenidos al considerar todos los posibles "puntos de corte" de la PD, la curva ROC nos proporciona una representación global de la exactitud diagnóstica. La curva ROC es necesariamente creciente, propiedad que refleja el compromiso existente entre sensibilidad y especificidad: si se modifica el valor de corte para obtener mayor sensibilidad, sólo puede hacerse a expensas de disminuir al mismo tiempo la especificidad. Si la prueba no permitiera discriminar entre grupos, la curva ROC sería la diagonal que une los vértices inferior izquierdo y superior derecho. La exactitud de la prueba aumenta a medida que la curva se desplaza desde la diagonal hacia el vértice superior izquierdo (Figura 4). Si la discriminación fuera perfecta

FIGURA 4. ESQUEMA EXPLICATIVO DE DISTINTAS POSIBILIDADES DE CURVAS ROC.



(100% de sensibilidad y 100% de especificidad) pasaría por dicho punto (12, 13).

Una curva ROC se construye a partir de los métodos no paramétricos. Estos se caracterizan por no hacer ninguna suposición sobre la distribución de los resultados de la PD. El más simple de estos métodos es el que suele conocerse como empírico, que consiste simplemente en representar todos los pares; es decir todos los pares (1-especificidad y sensibilidad) para todos los posibles "puntos de corte" que se puedan considerar con la muestra particular de la que se disponga. Desde un punto de vista técnico, este método sustituye las funciones de distribución teórica por una estimación no paramétrica de ellas.

No obstante ello, se pueden construir también aplicando métodos paramétricos, que se basan en postular un determinado tipo de distribución para la variable de decisión en las dos poblaciones en estudio. El modelo más frecuentemente utilizado es el binormal, que supone la normalidad de las variables tanto en la población enferma como en la "no enferma"; pero existen otros modelos posibles que surgen al considerar distintas distribuciones, similares a la normal como la logística (modelo bilogístico) o la exponencial negativa. El problema ahora se reduce a estimar los parámetros de cada distribución por un método estadísticamente adecuado; en general el método de máxima verosimilitud. Se obtiene así una curva ROC suave, pero puede ocurrir una sustancial falta de ajuste si los supuestos distribucionales resultan ser erróneos. Por ello, si se va a emplear este método debe previamente someterse la hipótesis sobre la naturaleza de las distribuciones a un contraste de significación.

En base a los párrafos anteriores se desprenden algunos conceptos que es necesario aclarar.

1. El área bajo la curva ROC que corresponde al área bajo la curva (ABC) y que se puede emplear como un índice conveniente de la exactitud global de la PD: la exactitud máxima correspondería a un valor de ABC de 1 y la mínima a uno de 0,5 (si fuera menor de 0,5 debería invertirse el criterio de positividad de la prueba).

2. Elección del "punto de corte". Para ello es imprescindible un conocimiento detallado de los riesgos y beneficios de las decisiones médicas derivadas del resultado de la PD. Un enfoque sencillo que utiliza la razón de costes de un resultado FP frente a un FN, lo que requiere calcular el siguiente coeficiente: donde "P" representa a la prevalencia de la enfermedad o evento de interés; o sea el costo de tener un falso positivo, versus el costo de un falso negativo. El valor de corte óptimo se determina hallando el punto de la curva ROC con la siguiente propiedad: la tangente a la curva en ese punto tiene una pendiente "m".

$$m = \frac{\text{Coste de los FP}}{\text{Coste de los FN}} * \frac{1 - P}{P}$$

En consecuencia, las curvas ROC son útiles para: conocer el rendimiento global de una PD (área bajo la curva), comparar dos PD o dos puntos de corte de una misma; comparar dos curvas o dos puntos sobre una curva; y elegir el "punto de corte" apropiado para un determinado paciente.

Sin embargo, tienen limitaciones, las que dicen relación con que sólo contemplan dos estados clínicos posibles: enfermo y "no enfermo"; y no sirven para situaciones en que se trata de discernir entre más de dos enfermedades o eventos de interés (Figura 4).

CRITERIOS A CONSIDERAR EN LA VALORACIÓN DE UNA PD

Características de la población. La sensibilidad o especificidad de una prueba dependen de las características de la población estudiada.

Si se altera o cambia la población en estudio, cambiarán también estos índices. Los datos informados de sensibilidad y especificidad, que son evaluados en poblaciones con una tasa significativa de enfermedad, pueden no ser aplicables en otras poblaciones diferentes en las que se utilice la prueba. Para que este criterio se cumpla, el artículo debe contener información sobre los siguientes aspectos: género y edad de los sujetos en evaluación, resumen de los síntomas clínicos iniciales o estadio de la enfermedad, y criterios de elección para los sujetos que son enrolados en el estudio.

Subgrupos adecuados. La sensibilidad y la especificidad pueden representar valores promedios para una población determinada. A menos que el problema para el cual se utiliza la prueba haya sido definido con mucha precisión, aquellas pueden variar en diferentes subgrupos poblacionales. Para que la prueba pueda ser utilizada con éxito deberían tenerse en cuenta distintos niveles de precisión según los distintos subgrupos existentes en la población estudiada. Este criterio se cumple cuando se informa sobre la precisión de la prueba en relación con cualquier subgrupo demográfico o clínico (por ejemplo en sujetos sintomáticos y sujetos asintomáticos).

Sesgo de selección. Puede producirse cuando los sujetos con los resultados positivos o negativos de una prueba son derivados de forma preferente para verificar el diagnóstico mediante otra prueba considerada el estándar de referencia. Para que este criterio se cumpla, todos los sujetos deberían de haber sido asignados para recibir tanto la prueba diagnóstica en estudio como el estándar de referencia a través de un procedimiento directo o mediante el seguimiento clínico.

Sesgo de medición. Podría introducirse si la PD o el estándar de referencia se realizan sin tomar precauciones para garantizar la objetividad de su interpretación (similar al enmascaramiento utilizado en los ensayos clínicos para tratamiento). Se puede obviar si la PD en evaluación y el estándar de referencia son interpretadas de forma separada y enmascarada por personas independientes que desconocen los resultados de una y otra.

Precisión de los resultados. La precisión de la sensibilidad y la especificidad depende del número de pacientes evaluados. Igual que otras medidas, el resultado estimado debe tener los intervalos de confianza o el error estándar reportados independientemente de la magnitud encontrada.

Presentación de resultados indeterminados. No todos las PD dan lugar a un sí o un no como respuesta, a veces dan lugar a resultados equívocos o indeterminados. La frecuencia de resultados indeterminados limitará la aplicabilidad de la prueba o la hará más cara si da lugar a otros procedimientos diagnósticos posteriores. La frecuencia de resultados indefinidos y el modo en el que se usan en el cálculo de la precisión de la prueba constituyen una información de importancia crítica para conocer la eficacia de la misma. Para que este criterio se cumpla el trabajo debe reflejar de forma apropiada todos los resultados positivos, negativos o indeterminados generados durante el estudio, así como si los resultados indeterminados se incluyeron o excluyeron al calcular los indicadores de precisión de la prueba.

Reproducibilidad de la prueba. Las pruebas no siempre dan el mismo resultado, por motivos relacionados con la variabilidad de éstas o de la interpretación del observador. Los motivos y el impacto de este asunto deben ser tenidos en cuenta. Para que se cumpla este crite-

rio en pruebas que requieren interpretación del observador, al menos alguna de las pruebas debería ser evaluada con alguna medida que resuma la variabilidad interobservador. Para pruebas sin interpretación del observador, el criterio se cumple cuando se refleja una media que resuma la variabilidad del instrumento (14).

EJEMPLO

Se diseñó y validó una escala de síntomas de fácil aplicación para la detección de enfermedad por reflujo gastroesofágico (ERGE). Los investigadores, de acuerdo a los resultados de la escala, exploraron 5 posibles puntos de corte para medir el constructo ERGE (15). La sensibilidad, especificidad, valores predictivos, razones de probabilidad y área bajo la curva para cada punto de corte se aprecia en la Tabla 2.

Para este ejemplo, el punto de corte recomendable para un estudio de prevalencia en población general sería el punto 3. Esto debido a un adecuado balance entre sensibilidad, especificidad; VPP y VPN; y RPP y RPN. Por otro lado, con este punto de corte, se verificó una fuerza de asociación entre los sujetos en estudio y la escala generada, con un OR superior a 200. Finalmente, en la curva ROC se logra apreciar que para este punto de corte, el área bajo la curva es apropiado (Figura 5).

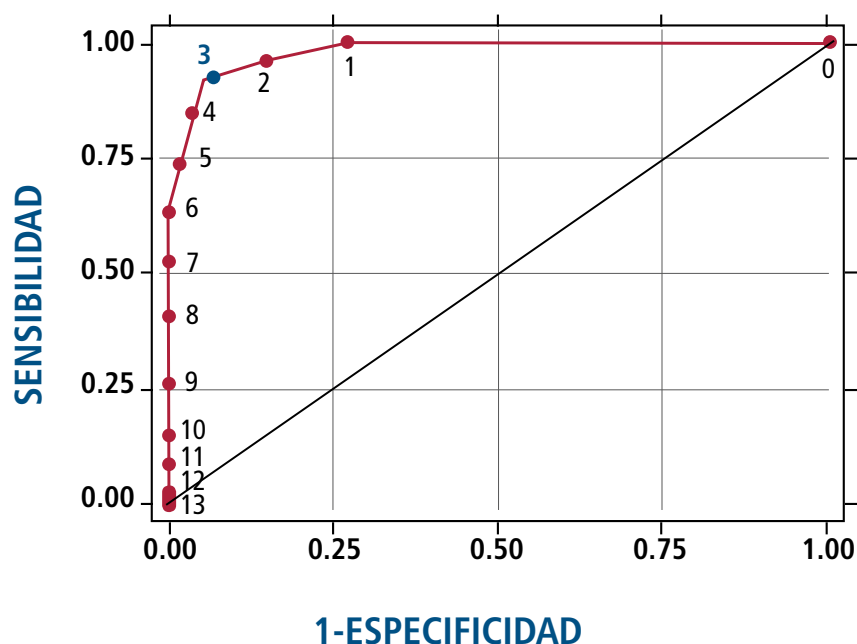
TABLA 2. PROPIEDADES DE UNA ESCALA DE SÍNTOMAS PARA DIAGNOSTICAR ERGE

Propiedades de la escala de ERGE	Puntos de corte				
	1	2	3	4	5
Sensibilidad (%)	100,0	95,6	91,6	84,4	73,7
Especificidad (%)	72,9	86,4	94,9	96,6	98,3
VPP (%)	91,8	95,5	98,2	98,7	99,3
VPP (%)	100,0	86,4	78,9	67,1	55,2
RPP	3,69	7,03	17,96	24,8	43,4
RPN	0,0	0,05	0,89	0,16	0,27
Clasificación correcta (%)	91,8	93,3	92,4	87,4	79,8
Área bajo la curva	50,0	90,9	93,3	90,5	86,0
Asociación entre sujetos en estudio y la escala generada (OR)	*	136.3	204.1	153.7	162.9

OR = Odds ratio

* = Predice error de forma perfecta

FIGURA 5. CURVA ROC DEL INSTRUMENTO DE DIAGNÓSTICO DE ERGE MENCIONADO EN EL EJEMPLO



COMENTARIOS

En términos generales, se observa que los estudios de precisión diagnóstica necesitan incluir en su reporte varios tipos de tablas y gráficos con el fin de proporcionar un panorama detallado de los resultados, de forma tal de poder comunicar mejor la información que posteriormente será utilizada en la práctica clínica (el valor predictivo); situación que suele confundir más que allanar el camino al clínico (11).

Es posible entonces, que el reporte de estudios de PD necesite un nuevo punto de partida. Si los métodos que tenemos para expresar la precisión de estas pruebas, no la expresan de forma apropiada o amigable; entonces debemos encontrar nuevos métodos que lo hagan. Estos, deben ser entendibles por el clínico, de tal forma que su interpretación pueda ser relevante a una amplia variedad de situaciones clínicas y poblaciones de pacientes; es decir, fáciles de utilizar en la práctica diaria y de disponibilidad instantánea (16).

Mientras ello ocurre, habrá que seguir comunicando, leyendo e interpretando los conceptos de sensibilidad, especificidad, valores predictivos, razones de probabilidad, curva ROC, puntos de corte y área bajo la curva de la PD de la que necesitemos información para el cuidado de nuestros pacientes.

REFERENCIAS BIBLIOGRÁFICAS

1. Manterola C. Estudios observacionales. Los diseños utilizados con mayor frecuencia en investigación clínica. *Rev Med Clin Condes* 2009;20:539-548.
2. Reid MC, Lane DA, Feinstein AR. Academic calculations versus clinical judgments: practicing physicians' use of quantitative measures of test accuracy. *Am J Med* 1998;104:374-380.
3. Fischer JE, Bachmann LM, Jaeschke R. A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. *Intensive Care Med* 2003;29:1043-1051.
4. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ* 2003;326:41-44.
5. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Epidemiología Clínica: Ciencia básica para la medicina clínica*. Editorial Médica Panamericana, Buenos Aires, Argentina, 1994.
6. Young JM, Glasziou P, Ward JE. General practitioners' self ratings

of skills in evidence based medicine: validation study. *BMJ* 2002;324:950-951.

7. Hulley SB and Cummings SR. Designing clinical research. Williams and Wilkins, Second Edition, Philadelphia, 2001.

8. Altman DG, Bland JM. Statistics Notes: Diagnostic tests 1: sensitivity and specificity. *BMJ* 1994;308:1552.

9. Altman DG, Bland JM. Statistics Notes: Diagnostic tests 2: predictive values. *BMJ* 1994;309:102.

10. Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. *BMJ* 2004;329:168-169.

11. Whiting PF, Sterne JA, Westwood ME, Bachmann LM, Harbord R, Egger M, Deeks JJ. Graphical presentation of diagnostic information. *BMC Med Res Methodol* 2008;8:20.

12. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* 1993;39:561-577.

13. Altman DG, Bland JM. Statistics Notes: Diagnostic tests 3: receiver operating characteristic plots. *BMJ* 1994;309:188.

14. Steurer J, Fischer JE, Bachmann LM, Koller M, ter Riet G. Communicating accuracy of tests to general practitioners: a controlled study. *BMJ* 2002;324:824-826.

15. Manterola C, Muñoz S, Grande L, Bustos L. Initial validation of a questionnaire for detecting gastroesophageal reflux disease in epidemiological settings. *J Clin Epidemiol* 2002;55:1041-1045.

16. Reid MC, Lane DA, Feinstein AR. Academic calculations versus clinical judgments: practicing physicians' use of quantitative measures of test accuracy. *Am J Med* 1998;104:374-380.