

Cómo interpretar un artículo sobre pruebas diagnósticas*

Drs. MARÍA EUGENIA BURGOS D.^{1,2}, CARLOS MANTEROLA D.¹

¹ Departamento de Cirugía y Traumatología.

² Programa de Doctorado en Ciencias Médicas.

Facultad de Medicina, Universidad de La Frontera.
Temuco, Chile.

Assessment of diagnostic test studies

Introducción

Las pruebas de diagnóstico (PD), constituyen un elemento fundamental en el ejercicio médico cotidiano. Tal como lo señala su nombre, son útiles para el diagnóstico, pero también tienen mucha utilidad en el tratamiento, la toma de decisiones relacionadas con el pronóstico y en el ámbito de la salud pública que tiene relación con las indicaciones y utilidad de las pruebas de tamizaje o *screening*¹.

En nuestra práctica diaria, los médicos nos enfrentamos frecuentemente a la decisión de solicitar o no exámenes que nos ayuden en el proceso diagnóstico de nuestros pacientes, y una vez que los solicitamos, nos enfrentamos a su vez al, a veces difícil, proceso de cómo interpretarlos. Por otro lado, frente al constante desarrollo tecnológico en el que nos vemos inmersos, dentro del cual un rol importante lo desempeña la constante implementación de nuevas PD, o modificaciones de algunas ya existentes, los médicos deben adquirir y desarrollar la habilidad de interpretar y evaluar estudios relacionados con PD. Sin embargo, esta interpretación no resulta fácil, principalmente porque involucra conceptos abstractos, como son sensibilidad (S), especificidad

(E), valores predictivos (VP), etc. Por este motivo, a continuación intentaremos resumir, en forma sencilla, los principales aspectos que se deben conocer para la interpretación adecuada de un estudio de PD.

Conceptos generales

Una vez que creemos que un determinado artículo de PD es potencialmente útil y relevante en nuestro quehacer clínico, debemos determinar si la metodología que se utilizó en el estudio nos da la suficiente confianza de que se evitaron los sesgos propios de este tipo de diseño. Para determinar lo anterior, existen ciertos criterios que se deben evaluar en su cumplimiento; los que se señalan a continuación:

1. Espectro de los pacientes: En términos generales, una PD es útil si permite diferenciar entre distintas enfermedades o condiciones clínicas, así como entre la condición de “sano” (sin el evento de interés en estudio) y de “enfermo” (con el evento de interés en estudio)¹. Por otro lado, prácticamente cualquier PD podría diferenciar entre un sujeto “gravemente enfermo” y uno “sano” o “no-enfermo”. Por lo tanto,

*Recibido el 21 de Enero de 2010 y aceptado para publicación el 10 de Marzo de 2010.

Este estudio ha sido financiado parcialmente por el proyecto DI09-0060 de la Dirección de Investigación de la Universidad de La Frontera.

Correspondencia: Dr. Carlos Manterola D.
Casilla 54-D, Temuco, Chile. Fax: 56-45-325761
E-mail: cmantero@ufro.cl

el valor práctico de una PD se basa en su utilidad en aquellos pacientes o sujetos que comúnmente vemos en la práctica clínica, los cuales habitualmente se presentan con diversos estadios clínicos de una enfermedad o evento bajo estudio². Por lo tanto, es importante determinar en el estudio de PD qué estamos analizando, qué tan amplio fue el espectro de sujetos que se incluyó en la valoración de la PD. Un espectro apropiado de sujetos a estudiar debe incluir a individuos que comparten los mismos síntomas, pero que presentan distintas patologías y excluir a los extremos, es decir, a los “sanos” o “no enfermos” y a los “muy enfermos”. De esta forma, la población del estudio se asemejará a la población en que en nuestra práctica clínica se aplicará finalmente la PD y es donde mayor utilidad tendrá, ya que podrá discriminar entre patologías distintas, pero con presentaciones clínicas similares.

2. Estándar de referencia: El rendimiento de toda PD se basa finalmente en la comparación con un estándar de referencia (*gold standard*, patrón de oro o estándar de oro), que no es más que la o las técnicas diagnósticas que definen la presencia o ausencia de la condición en estudio, con la máxima certeza conocida. Hay varios aspectos que deben evaluarse en relación al estándar de referencia y que se detallan a continuación.

2.1. Utilización del estándar de referencia más adecuado: Como se señaló anteriormente, para determinar la precisión de una PD, ésta debe ser comparada con “la verdad” o en su defecto, con lo más cercano a la verdad posible³. Esta verdad generalmente está dada por el estándar de referencia que la comunidad científica determine como el más adecuado (biopsia, cirugía, necropsia, seguimiento a largo plazo, etc.), siendo a veces necesaria la combinación de exámenes como patrón de referencia. Ahora bien, si en el estudio que estamos analizando, el estándar de referencia no impresiona ser el más adecuado, es poco probable que los resultados sean válidos para nuestro propósito.

2.2. Evaluación independiente del estándar de referencia y la PD en estudio: Una vez que aceptamos que el estándar de referencia es el más adecuado, debemos asegurarnos que la PD en estudio y el estándar de referencia fueron evaluados en forma independiente. Esto quiere decir que la interpretación de una nueva PD no debe estar influenciada por el conocimiento del resultado del estándar de referencia o viceversa. En la medida en que no se realiza esta evaluación enmascarada, la investigación está propensa a sobreestimar el poder de la PD en estudio.

2.3. Influencia del resultado de la PD en la realización del estándar de referencia: El estándar de referencia en ocasiones puede ser costoso, invasivo

o incluso riesgoso y usualmente se requiere de una nueva PD que reemplace al estándar de referencia⁴. Las propiedades de una PD estarán distorsionadas si su resultado influye en la decisión de llevar a cabo el estándar de referencia. Cuando se produce esta situación, conocida como sesgo de verificación, el investigador sólo lleva a cabo la investigación con el estándar de referencia en aquellos sujetos en que la PD en estudio resulta positiva, asumiendo que en aquellos pacientes en que la prueba resultó negativa no tienen la condición en estudio.

Por lo tanto, si después de analizar el estudio sobre PD concluimos que se han cumplido los criterios antes mencionados, podemos señalar que los resultados del estudio probablemente representan una estimación no sesgada de las características de la PD en estudio. Lo que queda entonces a continuación es realizar una correcta interpretación de los resultados y analizar su aplicabilidad en nuestra práctica clínica habitual.

Interpretación de los resultados

Validez de una prueba diagnóstica

La validez es el grado en que una PD mide lo que se supone que debe medir y es aquí donde entran en juego los conceptos de sensibilidad y especificidad.

Es así que, para poder realizar una correcta interpretación de los resultados que se nos entregan en un estudio sobre PD, debemos primeramente conocer y familiarizarnos con ciertos conceptos que nos resultan muchas veces abstractos y poco amigables, como son la S, E, los VP, las razones de probabilidad (RP), etc.

Antes de ir de lleno a la definición de estos conceptos debemos tener en mente siempre la tabla de 2 x 2, conocida también como tabla de contingencia. (Figuras 1 y 2). En esta tabla, si comparamos una PD con el estándar de referencia, se pueden obtener cuatro combinaciones si los resultados de la PD se expresan en forma dicotómica. En el fondo, enfrentamos el resultado de la PD (en filas) con el estado real de los pacientes (en columnas), determinado por el estándar de referencia que vamos a utilizar. Finalmente, el resultado de la PD puede ser correcto (verdadero positivo y verdadero negativo) o incorrecto (falso positivo y falso negativo).

Es así como tendremos un “Verdadero Positivo” cuando el estándar de referencia es positivo y la PD en estudio es positiva. Un “Verdadero Negativo” cuando el estándar de referencia es negativo y la PD en estudio es negativa. Un “Falso Positivo” cuando el estándar de referencia es negativo y la PD en estudio es positiva. Y, un “Falso Negativo” cuando el estándar de referencia es positivo y la PD en estudio es negativa.

Resultado de la prueba de 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 2 x 2 en la que se explica la generación de las celdas con las que se realizan los cálculos tendientes a obtener las medidas de S, E y VP.

A partir de los valores que se obtienen en esta tabla de contingencia se pueden calcular las distintas formas en que se pueden expresar el rendimiento de una PD (Figuras 1 y 2).

Sensibilidad

La sensibilidad, corresponde a la proporción de sujetos que presentan la enfermedad o evento en estudio determinada por el estándar de referencia, en los cuales la PD que se está probando resulta positiva. En otras palabras, corresponde a la proporción de verdaderos positivos, clasificados como positivos según el estándar de referencia, los cuales son correctamente identificados como positivos por la PD en estudio⁵. La S es, por lo tanto, la capacidad que posee la PD para detectar la enfermedad o evento de interés en estudio. Para calcular entonces la S de una PD se debe 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 $a / (a + c)$; o $VP / VP + FN$ (Figuras 1 y 2).

En el contexto clínico, una prueba altamente sensible es especialmente útil en aquellos casos en los que al no diagnosticar una enfermedad o evento específico, esto puede resultar incluso fatal para los enfermos, como es el caso de enfermedades peligrosas pero tratables, como la tuberculosis; o en enfermedades en las que un FP no conduzca a serios trastornos psicológicos o económicos para el paciente. Es así como, en general, las pruebas de tamizaje deben ser de alta S para poder detectar, en lo posible, a todos los enfermos.

Especificidad

La especificidad, corresponde a la proporción de sujetos que no presentan la enfermedad o evento en estudio determinada por el estándar de referencia, en los cuales la PD que se está probando resulta negativa. En otras palabras, corresponde a la proporción de verdaderos negativos, clasificados como

Resultado de la prueba de estudio	Estado respecto a la enfermedad según el estándar de referencia	
	Enfermo	No enfermo
Positivo	Verdadero positivo (VP)	Falso positivo (FP)
Negativo	Falso negativo (FN)	Verdadero negativo (VN)

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

tales según el estándar de referencia, los cuales son correctamente identificados como negativos por la nueva PD en estudio⁵. Para calcular entonces la E de un PD se debe 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 $b / (b + d)$; o $FP / FP + VN$ (Figuras 1 y 2).

En la práctica clínica, en general, las pruebas que se utilizan para confirmar un diagnóstico, deben ser de alta especificidad, para evitar los falsos positivos. Por lo tanto, una PD de alta especificidad es necesaria ante enfermedades graves, pero sin tratamiento disponible que las haga curables, cuando exista gran interés por conocer la ausencia de enfermedad, o cuando, por el hecho de diagnosticar a un sujeto con una enfermedad o evento de interés que realmente no posee, esto pudiera conllevarles graves consecuencias, ya sean físicas, económicas, psicológicas, etc.; por ejemplo el SIDA.

Como podemos ver entonces, existe una estrecha relación entre S y E, la cual generalmente es de tipo inversa; para el resultado específico de una PD la S puede incrementarse solamente a expensas de la E y viceversa. Sin embargo, S y E son una propiedad intrínseca de una PD, cuyos resultados son independientes de la prevalencia de la enfermedad y en la práctica clínica, como se explica a continuación, carecen de utilidad. Tanto la S como la E entregan información acerca de la probabilidad de obtener un resultado positivo o negativo en función de la verdadera condición del sujeto determinada por el estándar de referencia. No obstante, cuando a un sujeto se le aplica una PD, en general el médico carece de una información *a priori* acerca del diagnóstico y lo que realmente le interesa es, ante una PD con resultado positivo o negativo, ¿cuál es la probabilidad de que el sujeto tenga o no el evento de interés en estudio? Es aquí donde entran en juego los VP, los cuales complementan esta información.

Valor predictivo positivo

El valor predictivo positivo (VPP), es la probabilidad de padecer la enfermedad o el evento de interés si se obtiene un resultado positivo en la PD que está siendo aplicada. Dicho de otra forma, corresponde a la proporción de sujetos con resultado positivo para la PD que verdaderamente presentan la enfermedad o evento de interés⁶. Para calcular entonces el VPP de un PD se debe dividir el número de enfermos con prueba positiva por la sumatoria de los enfermos con prueba positiva y los sujetos “no enfermos” con prueba positiva; es decir $a / (a + b)$; o $VP / VP + FP$ (Figuras 2 y 3).

Valor predictivo negativo

El valor predictivo negativo (VPN) es la probabilidad de no padecer la enfermedad o el evento de interés si se obtiene un resultado negativo en la PD que está siendo aplicada. Dicho de otra forma, corresponde a la proporción de sujetos con resultado negativo para la PD que verdaderamente no presentan la enfermedad o evento de interés⁶. Para calcular entonces el VPN de una PD debemos dividir el número de enfermos con prueba negativa por la sumatoria de los enfermos con prueba negativa y los sujetos “no enfermos” con prueba negativa; es decir $d / (c + d)$; o $VN / FN + VN$ (Figuras 2 y 3).

Sin embargo, a pesar de que los VP nos proporcionan información que es relevante clínicamente, ésta sólo es utilizable si la queremos aplicar a sujetos que son similares a los que se incluyeron en el estudio. Lo anterior, debido a que los VP varían enormemente dependiendo de la prevalencia de la enfermedad, por lo que si nuestro paciente o sujeto en estudio tiene un riesgo mayor o menor de presentar la enfermedad o evento de interés, los resultados no serán aplicables para nuestro paciente.

Las razones de probabilidad

Como podemos darnos cuenta hasta ahora, las características operativas tradicionales de una PD, como son S, E y VP, tienen algunas limitaciones en la práctica clínica habitual. En base a la S o E de una PD, consideradas en forma aislada, es imposible concluir si un paciente tiene o no una determinada enfermedad o evento de interés. La interpretación del resultado de una PD, en el contexto clínico, requiere de la utilización de los VP, pero estos a su vez, pueden tener valores diferentes en cada escenario clínico, dependiendo de la prevalencia de la enfermedad o evento de interés. Adicionalmente, estas medidas se aplican solamente a PD con resultados dicotómicos: positivo o negativo. Como una alternativa para vencer estas limitaciones, se ha propuesto la utilización de otra herramienta en la evaluación de las PD: las RP.

Las RP (coeficiente de probabilidad o likelihood ratio), comparan la probabilidad de encontrar el resultado de la PD (positivo o negativo) en personas enfermas o con el evento de interés, con la probabilidad de encontrar ese mismo resultado en personas sin la enfermedad o evento de interés⁷.

La razón de probabilidad positiva (RPP), se calcula dividiendo la probabilidad de un resultado positivo en los pacientes enfermos o con el evento de interés, por la probabilidad de un resultado positivo en los sujetos sin la enfermedad o evento de interés⁷. En definitiva, corresponde al cociente entre la fracción de VP (S) y la fracción de FP (1-E) (Figura 3).

La razón de probabilidad negativa (RPN), se calcula dividiendo la probabilidad de un resultado negativo en los pacientes enfermos o con el evento de interés, por la probabilidad de un resultado negativo en los sujetos sin la enfermedad o evento de interés. En definitiva, corresponde al cociente entre la fracción de falsos negativos (1-S) y la fracción de VN (E) (Figura 3).

La RP tiene la ventaja de que relaciona en un solo índice la S y E, además de que pueden obtenerse RP según varios niveles de una nueva medida y no es necesario expresar la información en forma dicotómica (normal/anormal, positivo/negativo). Finalmente, al igual que lo que sucede con la S y E, no varía con la prevalencia de la enfermedad o evento de interés. Por lo tanto, se desprende de lo anterior, que la principal ventaja de las RP es que permiten su utilización como índice de comparación entre diferentes PD para un mismo diagnóstico.

Las razones de probabilidad son particularmente útiles en la práctica clínica debido a que nos permiten entender con qué fuerza el resultado positivo de una PD indica la presencia real de la enfermedad o evento de interés y la fuerza de un resultado negati-

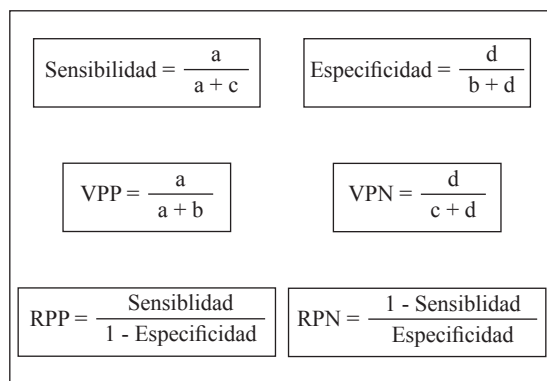


Figura 3. Fórmulas para la realización de los cálculos de S, E, VP y RP.

vo para descartar la enfermedad o evento de interés. Es aquí donde entran en juego otros conceptos importantes: probabilidad pre-test y probabilidad post-test.

En términos sencillos, la probabilidad pre-test es la estimación de la prevalencia de la enfermedad o evento de interés que asumimos antes de realizar una PD. Como todos sabemos, un sujeto en que se sospecha una determinada enfermedad o evento de interés, tendrá una cierta probabilidad de presentarla, la cual dependerá de ciertas variables como son la prevalencia de la enfermedad o evento de interés en la población, de las características del sujeto y de los signos o síntomas que éste presente. Para realizar una estimación de esta probabilidad pre-test, en general la podemos obtener realizando una estimación aproximada basada en nuestra experiencia personal o de nuestros pares o en base a datos estadísticos o epidemiológicos del medio en que trabajamos.

Por otro lado, una vez que se realiza la PD en el estudio, la probabilidad de presentar la enfermedad o evento de interés aumenta o disminuye, dependiendo del resultado de la PD denominándose ahora probabilidad post-test.

En este punto, es importante enfatizar que las propiedades de la PD que nos permite cuantificar la magnitud y el sentido del cambio de la probabilidad pre-test a post-test son las RP. En la práctica, si conocemos o podemos determinar la probabilidad pre-test de una determinada enfermedad o evento de interés y conocemos las RP de la PD, podemos utilizar el Nomograma de Fagan para determinar la probabilidad post-test (Figura 4). En éste, la columna izquierda representa la probabilidad pre-test, la columna del centro a la RP y la columna de la derecha a la probabilidad post-test. Trazando una línea recta se unen los dos primeros valores y podemos determinar así la probabilidad post-test. En este nomograma, los cambios más significativos en la probabilidad de enfermedad o evento de interés ocurren con PD que tienen valores de RP mayores de 10 o menores de 0,1; los cuales usualmente son muy útiles para confirmar o descartar la determinada enfermedad o evento de interés.

La curva ROC (Receiver Operating Characteristic)

Hasta el momento, sólo nos hemos referido a aquellas PD que presentan un resultado dicotómico: positivo/negativo, presente/ausente, etc. Sin embargo, existe una amplia gama de PD cuyos resultados se miden en una escala continua, como por ejemplo el resultado de la glucosa sérica. En estos casos, lo que se pretende es determinar valores de corte o niveles de decisión que permitan lograr una clasificación dicotómica de los valores de la PD según sean superiores o inferiores al valor de corte estimado. Esencialmente además, en estos casos, ya no se cuenta con un único valor de sensibilidad y especificidad que definan la exactitud de la PD, sino que más bien con un conjunto de pares correspondientes cada uno a un distinto nivel de decisión. Para obtener la curva ROC, se representa para cada posible valor de corte la S en la ordenada y 1-E en la abscisa (Figura 5). Mediante esta representación, que considera todos los posibles valores de corte de la PD, la curva ROC nos proporciona una representación global de la exactitud diagnóstica. Como vemos en la figura, la curva ROC es necesariamente creciente, lo que refleja la relación existente entre S y E; si se modifica un determinado valor de corte pretendiendo obtener mayor S, necesariamente lo hace a expensas de una disminución en E^{8,9}. Ahora bien, la exactitud de la PD aumenta a medida que la curva se desplaza desde la diagonal hacia el vértice superior izquierdo (Figura 5). Para construir la curva ROC se pueden utilizar los métodos no paramétricos y los paramétricos. En el primero de los casos, se caracterizan por no hacer ninguna suposición sobre la distribución de los resultados de la PD; dentro de ellos el más simple es el empírico que básicamente

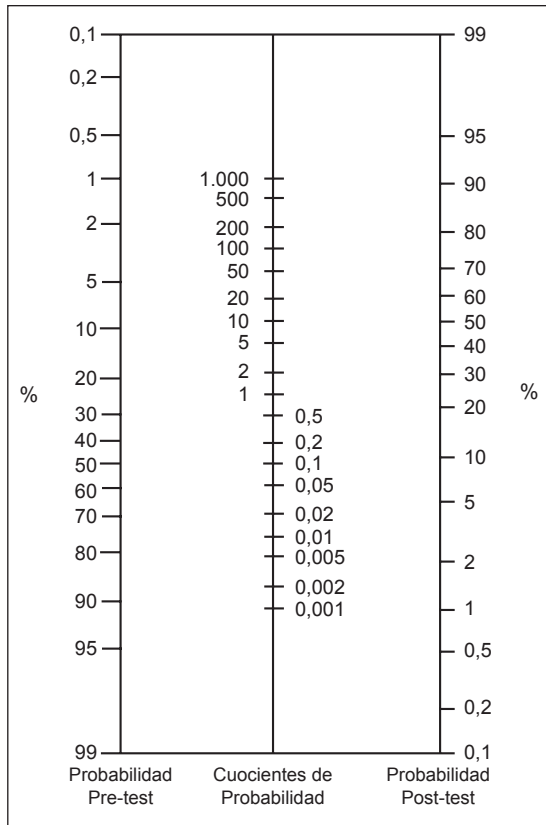


Figura 4. Esquema del nomograma de Fagan.

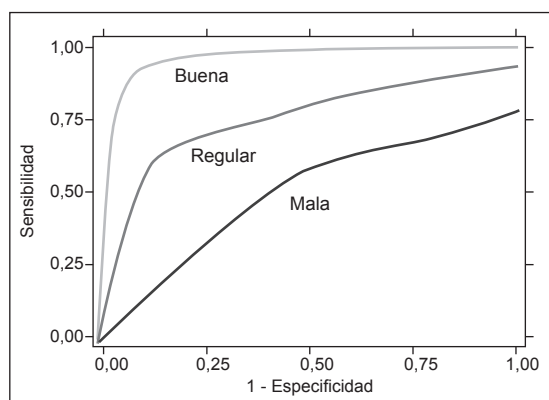


Figura 5. Esquema explicativo de distintas posibilidades de curvas ROC.

consiste en representar todos los pares (1-E, S) para todos los posibles valores de corte que se pueden considerar con la muestra de la cual se dispone. Por otro lado, los métodos paramétricos se basan en postular un determinado tipo de distribución para la variable de decisión en las dos poblaciones que se trata de distinguir. Dentro de estos últimos, el modelo más frecuentemente utilizado es el binormal, el cual supone la normalidad de las variables tanto en la población sana o sin el evento de interés como en la enferma o con el evento de interés.

En relación al análisis estadístico de la curva ROC, como se señala anteriormente, la mayor exactitud de una PD se traduce en un desplazamiento hacia arriba y hacia la izquierda de la curva ROC. De lo anterior, se desprende que el área bajo la curva ROC (ABC) se puede utilizar como un índice conveniente de la exactitud global de la PD donde la exactitud máxima corresponde a un valor de ABC de 1 y la mínima a uno de 0,5. Finalmente, el ABC puede interpretarse como un promedio de la S (E) sobre todos los valores posibles de E (S), aunque en la práctica clínica puede que nos interese solamente los puntos de corte de la curva ROC que nos aseguren altos niveles de S y E. Donde tiene importancia además el cálculo del ABC es en aquellas situaciones en que se requiere discriminar entre dos o más PD para utilizarlas en el diagnóstico de una determinada patología o evento de interés. En términos generales al comparar dos o más PD se debería elegir aquella que tenga mayor ABC, por ser la de mayor exactitud diagnóstica.

Otro punto importante a señalar en relación a la curva ROC es que al realizar la evaluación de una determinada PD, en la práctica clínica es preciso determinar un punto de corte que discrimine entre un sujeto enfermo o con el evento de interés y uno sano o sin el evento de interés. Para esto, es im-

prescindible poseer un conocimiento detallado de los riesgos y beneficios de las decisiones médicas que se adoptan como resultado de la aplicación de la PD. Una manera sencilla y que a la vez utiliza la razón de costes de un resultado falso positivo frente a un falso negativo requiere calcular el “coeficiente *m*”, donde *P* corresponde a la prevalencia de la enfermedad. 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 pendiente *m*.

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

Ejemplo

Un buen ejemplo en el cual podemos observar la aplicación de las propiedades de una PD, lo podemos encontrar en la publicación de Escalona y cols¹⁰. En esta publicación, los autores se plantean como objetivo determinar el rendimiento de la proteína C reactiva y el recuento de leucocitos en el diagnóstico de la apendicitis aguda. En el artículo se reporta finalmente la sensibilidad, especificidad, valores predictivos positivos, negativos y cociente de probabilidad de la medición de proteína C reactiva y el recuento de leucocitos para el diagnóstico de apendicitis aguda, como se muestra en la tabla. Además, el poder discriminatorio de las variables se extrajo del área de curvas ROC, las que se construyeron en base a cruces entre la sensibilidad y especificidad de una PD para diversos puntos de corte.

Aplicabilidad de la prueba diagnóstica

Finalmente, una vez que hemos realizado una correcta interpretación de los resultados obtenidos en la PD, lo que queda es determinar la aplicabilidad de la PD en los sujetos evaluados en nuestra práctica clínica habitual. Para esto debemos considerar si la prueba es lo suficientemente reproducible, independientemente del ámbito, las condiciones de aplicación y de la persona que la interprete.

Un aspecto importante a considerar en este punto también es la similitud o diferencia existente entre nuestro paciente o sujeto en estudio y aquellos incluidos en el estudio. Al respecto, hay que tener en cuenta que el espectro de pacientes no se refiere solamente a la severidad de la condición bajo estudio, sino que también a las características demográficas de los pacientes o sujetos bajo estudio y a la presencia de comorbilidad. En términos generales, si uno va a realizar la PD en un contexto similar al que se presenta en el estudio y si además el paciente o sujeto en estudio cumple con los criterios de selección que se señalan en el estudio, uno podría estar confiado en aplicar los resultados.

Por otro lado y simultáneamente, una vez que hemos evaluado la validez del artículo sobre PD y después de haber realizado incluso algunos cálculos simples que nos hayan permitido entender sus resultados, debemos necesariamente preguntarnos si estos resultados ayudarán en el cuidado de nuestros pacientes¹¹. Es así como indudablemente valoramos aún más una PD si ésta posee riesgos inherentes que resultan aceptables, si al no tratar la enfermedad o evento de interés que está siendo estudiada esto ocasiona consecuencias mayores y si la enfermedad o evento de interés una vez que es diagnosticada puede ser tratada con rapidez.

Reporte de los resultados de un estudio de PD

Para reportar resultados referente a PD, se ha de observar alguna metodología que permita ordenar conceptos y unificar criterios. Algunas iniciativas de este tipo se han generado: STARD, REMARK y QUADAS.

La iniciativa STARD (Standards for Reporting of Diagnostic Accuracy), tiene por objetivo mejorar la calidad en el reporte de los estudios de exactitud diagnóstica, permitir a los lectores evaluar los potenciales sesgos en este tipo de estudios y determinar la generalización de los resultados¹². Para lograr esto, se realizó una extensa búsqueda para identificar publicaciones acerca de conducción y reporte de estudios de PD. Posteriormente, se realizó una reunión de consenso que involucró expertos relacionados con investigación, metodología, editores y profesionales, con el fin de identificar una lista de potenciales ítems que dieran origen a un "checklist". Este, quedó compuesto por 25 ítems y un diagrama de flujo que puede ser utilizado por autores para la descripción de elementos esenciales del diseño y conducción del estudio, la ejecución de la PD y los resultados¹³.

Por su parte, la iniciativa REMARK (REporting recommendations for tumour MARKer prognostic studies) ha enfocado sus recomendaciones en los estudios de marcadores tumorales en el ámbito de la oncología. La iniciativa se gestó ante la observación de la frecuente contradicción en los resultados reportados en los estudios sobre marcadores tumorales¹⁴. Para esto se confeccionó una guía compuesta por 20 ítems cuyo propósito es proporcionar un reporte transparente y completo de modo que la información relevante esté disponible y permita enjuiciar la utilidad de los datos y a entender el contexto en el cual las conclusiones aplican.

Finalmente, la iniciativa QUADAS (Quality Assessment of Diagnostic Accuracy Studies) se creó

como un instrumento para la evaluación de la calidad de los estudios de exactitud diagnóstica incluidos en revisiones sistemáticas. Para este proyecto, se realizó un consenso donde se realizó una revisión de la evidencia en conjunto con la opinión de expertos, con el objeto de desarrollar un instrumento que pueda ser utilizado en revisiones sistemáticas que evalúan la calidad de los estudios primarios de exactitud diagnóstica¹⁵. Este instrumento consta de 14 ítems en los cuales se evalúa el espectro de pacientes, el estándar de referencia, sesgos propios de este tipo de estudios como por ejemplo el sesgo de verificación, la ejecución de la PD, pérdidas en el seguimiento y presencia de resultados indeterminados.

En conclusión, estas herramientas fueron desarrolladas para mejorar la construcción, el reporte y la evaluación de los estudios de exactitud diagnóstica. El mejoramiento en el diseño y reporte de los estudios de exactitud diagnóstica puede influenciar positivamente la calidad de los datos disponibles para los clínicos en relación a la utilidad diagnóstica de pruebas utilizadas para tamizaje, diagnóstico, pronóstico y planificación de tratamiento¹⁶.

Conclusiones

Como cirujanos nos vemos enfrentados a una continua exposición a nuevas PD y a estudios que muchas veces intentan seducirnos con las propiedades de una nueva y prometedora prueba. Por esta razón, determinar la validez de estos estudios, saber interpretar correctamente los resultados y finalmente aplicar estos resultados en nuestros pacientes, son tres pasos fundamentales en la elección e interpretación de una PD. Esto nos ha motivado al desarrollo de esta guía que ha pretendido, de forma sencilla, proporcionar herramientas que permitan evaluar de forma crítica un estudio sobre PD.

Referencias

1. Jaimes F. Pruebas diagnósticas: uso e interpretación. *Acta Med Colomb* 2007; 32: 29-33.
2. Archibald S, Bhandari M, Thoma A. Users' guides to the surgical literature: how to use an article about a diagnostic test. *Can J Surg* 2001; 44: 17-23.
3. Valenzuela L, Cifuentes L. Validez de estudios de tests diagnósticos. *Rev Med Chile* 2008; 136: 401-404.
4. Gilbert R, Logan S, Moyer V, Elliott E. Assessing diagnostic and screening tests: Part 2. How to use the research literature on diagnosis. *West J Med* 2001; 175: 37-41.
5. Altman D, Bland J. Diagnostic tests 1: sensitivity and specificity. *BMJ* 1994; 308 (6943): 1552.

6. Altman D, Bland J. Diagnostic tests 2: predictive values. *BMJ* 1994; 309 (6947): 102.
7. Honest H, Khan KS. Reporting of measures of accuracy in systematic reviews of diagnostic literature. *BMC Health Serv Res* 2002; 2:4.
8. Zweig M, Campbel IG. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* 1993; 39: 561-577.
9. Altman D, Bland J. Statistics Notes: Diagnostic tests 3: receiver operating characteristic plots. *BMJ* 1994; 309 (6948): 188.
10. Escalona A. Utilidad de la proteína C reactiva y recuento de leucocitos en sospecha de apendicitis aguda. *Rev Chil Cir* 2006; 58: 122-126.
11. Bhandari M, Guyatt G. How to Appraise a Diagnostic Test. *World J Surg* 2005; 29: 561-596.
12. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al; Standards for Reporting of Diagnostic Accuracy. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ* 2003; 326: 41-44.
13. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al; Standards for Reporting of Diagnostic Accuracy. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Ann Intern Med* 2003; 138: W1-W12.
14. McShane L, Altman D, Sauerbrei W, Taube S, Gion M, Clark C. REporting recommendations for tumour MARKer prognostic studies (REMARK). *Br J Cancer* 2005; 93: 387-391.
15. Whiting P, Weswood M, Rutjes A, Reitsma J, Bossuyt P, Kleijnen J. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. *BMC Med Res Methodol* 2006; 6: 9.
16. Cook C, Cleland J, Huijbregts P. Creation and Critique of Studies of Diagnostic Accuracy: Use of the STARD and QUADAS Methodological Quality Assessment Tools. *J Man Manip Ther* 2007; 15: 93-102.