ARTÍCULO DE REVISIÓN

¿Son suficientes los indicadores del rendimiento de una prueba o test diagnóstico para evaluar su desempeño?

Are The Performance Indicators of aTest or Diagnostic Test Sufficient to Evaluate Performance?

Dariel Díaz Arce[1]
José Patricio Beltrán Carreño[2]
Johanna Elizabeth Cueva Sarmiento[2]

[1] Unidad Educativa Santana. Ecuador.
[2] Ministerio de Salud Pública. Ecuador.

RESUMEN

Introducción: La sensibilidad, especificidad, valor predictivo, índice de validez y razón de verosimilitud, son comúnmente empleados como indicadores del desempeño de las pruebas diagnósticas en el ámbito médico, pero ¿son suficientes para concluir que un test es aceptable para la práctica clínica?
Objetivo: Evaluar la utilidad de cada uno de estos indicadores y sus limitaciones, basándose en el análisis de un caso concreto.
Métodos: Se desarrolló un estudio bibliográfico para identificar los usos y abusos de los indicadores del desempeño de pruebas diagnósticas que con mayor frecuencia se emplean en la literatura.
Conclusiones: Los resultados indican que las herramientas estadísticas son un apoyo a la toma de decisiones sobre la calidad de un test diagnóstico cuando se interpretan adecuadamente dentro de un marco conceptual y metodológico viable, reproducible y aplicable a la población diana que se desea evaluar.

Palabras clave: Test diagnósticos; sensibilidad; especificidad; razón de verosimilitud; exactitud diagnóstica; valor predictivo.

ABSTRACT

Introduction: Sensitivity, specificity, predictive value and validity index and likelihood ratio are commonly used as indicators of the performance of diagnostic tests, but Are they sufficient to conclude that a test is acceptable for clinical practice?
Objective: To evaluate the usefulness of each one of these indicators and their limitations, based on the analysis of a recent report.
Methods: A bibliographic study was developed to identify the uses and abuses of the performance indicators of diagnostic tests that are most frequently used in the literature.
Conclusion: The results indicate that the statistical tools are a support for making decision on the quality of a diagnostic test but they must be interpreted properly within a viable conceptual and methodological framework, reliable and applicable to the target population to be evaluated.

Keywords: Diagnostic test; sensibility; specificity; *likelihood ratio;* diagnostic accuracy; predictive value.

INTRODUCCIÓN

Las pruebas diagnósticas se han convertido en aliadas ineludibles del accionar médico y de otros muchos especialistas de la salud. Es por ello que este campo de investigación científica es uno de los más activos en los últimos años, registrándose solo desde 2015 hasta enero del 2017 en la base de datos *Pubmed* de la*US National Library of Medicine* del *National Institute of Health* de los Estados Unidos, unos 260 estudios que en su título poseen la frase "*diagnostic test".*

Al mismo tiempo, se desarrollan incontables esfuerzos por mejorar el proceso diagnóstico con estas pruebas "complementarias", llamándose la atención sobre el hecho de que un gran número de los estudios publicados carecen de una calidad suficiente como para adaptarlos a la práctica clínica. Por ello, a nivel internacional surgieron varias iniciativas que pretenden aportar herramientas para evaluar tales trabajos con sus particularidades de diseño, resultados y aplicabilidad práctica. Algunos de estos son el instrumento *Quality Assessment of Diagnostic Accuracy Studies* (*QUADAS* y*QUADAS*-2),[1,2] el *Standard for Reporting of Diagnostic Accuray* (*STARD*), el*REporting recommendations for tumors MARKer prognostic studies* (*REMARK*)[1] y las recomendaciones del *Evidence-Based Medicine Working Group*, detalladas en trabajos como los de *Ochoa y cols.*,[3] *Vera y cols.,*[4] *Valenzuela y Cifuentes,*[5] entre otros.

El objetivo de este trabajo es evaluar la utilidad de cada uno de estos indicadores y sus limitaciones, basándose en el análisis de un caso concreto.

MÉTODOS

Se desarrolló un estudio bibliográfico para identificar los usos y abusos de los indicadores del desempeño de pruebas diagnósticas que con mayor frecuencia se emplean en la literatura.

DESARROLLO

Uno de los aspectos a los que más relevancia se da dentro del análisis de las pruebas diagnósticas es a los indicadores del rendimiento o desempeño, entre los que destacan la sensibilidad, la especificidad, los valores predictivos positivo y negativo, las razones de verosimilitud positiva y negativa, tasa de aciertos o índice de validez, entre otros.[6,7] La comprensión de estos indicadores, su utilidad práctica y sus límites permite al especialista reducir los errores por mal uso de las técnicas diagnósticas o una sobredimensionalización de las mismas.[8-11]

La presente investigación se enfocó en revisar las ventajas y limitaciones de estos indicadores de rendimiento diagnóstico, a través del análisis de un ejemplo concreto de reciente publicación. Para hacer más sencillo y didáctico este tema, en el artículo se emplearon las tablas de contingencia 2 × 2 para resultados dicotómicos, presencia o ausencia de la enfermedad *vs*. resultado positivo o negativo por el test diagnóstico.

Sensibilidad y Especificidad: ¿indicadores de validez?

Un test diagnóstico será válido cuando logre aportar suficiente información para determinar con bastante exactitud la presencia o ausencia de una condición clínica determinada, descartando en el proceso otras condiciones con similar presentación y potencialmente confusoras.[6]

Los términos sensibilidad y especificidad de una prueba se emplean comúnmente como indicadores de seguridad, considerándose como las probabilidades de clasificar correctamente a un individuo como enfermo o sano respectivamente.[12] La forma en la que se calculan estos indicadores se resume en la tabla 1.

En un trabajo reciente, *Tenezaca Sari y cols.,*[13] evaluaron la eficacia de la secuencia de Murphy en el diagnóstico clínico de apendicitis aguda, obteniendo una sensibilidad de 68,03 % y una especificidad de 71,43 %, concluyendo que esta secuencia es "aceptable para valorar a pacientes con sospecha diagnóstica de apendicitis aguda". Se propone así, la introducción de esta cronología en la elaboración de la anamnesis "(…) como referencia para realizar un examen físico dirigido, con el propósito de reducir el número de complicaciones (...)".

La sensibilidad y especificidad aportadas indican respectivamente que, de cada 100 pacientes con apendicitis aguda, la secuencia de Murphy ayuda en detección de 68 de ellos, mientras que de igual número de sanos se detectan 71. Aunque estos pudieran interpretarse como aceptables para detectar a los enfermos y descartar a los sanos, no son "medidas calibradas" ya que no toman en cuenta los acuerdos obtenidos al azar. Para analizar tal efecto se propone el cálculo de la sensibilidad y especificidad esperadas al azar y el posterior análisis del índice *kappa* ponderado como medida de su calidad.[14]

Tabla 1. Resumen de fórmulas para el análisis de indicadores del desempeño de las pruebas diagnósticas

Tabla de contingencia 2 × 2				
		Prueba de Oro		Total
		Enfermo	Sano	
Test	+			
	–			
Total				

Indicador	Fórmula	Interpretación
Sensibilidad	$S = \frac{a}{a+c} * 100$	Eficiencia para detectar a los pacientes enfermos.
Especificidad	$E = \frac{d}{b+d} * 100$	Eficiencia para detectar a los pacientes sanos.
Valor Predictivo Positivo	$VPP = \frac{a}{a+b} * 100$	Poder predictivo de enfermedad cuando el test es positivo.
Valor Predictivo Negativo	$VPN = \frac{d}{c+d} * 100$	Poder predictivo de no enfermedad cuando el test es negativo.
Índice de Validez	$IV = \frac{a+d}{a+b+c+d} * 100$	Proporción de individuos clasificados correctamente en enfermos o sanos.
Razón de verosimilitud positiva	$LR+ = \frac{a/(a+c)}{b/(b+d)}$	¿Cuánto más probable es encontrar un resultado positivo en pacientes realmente enfermos *vs*. sanos?
Razón de verosimilitud negativa	$LR- = \frac{c/(a+c)}{d/(b+d)}$	¿Cuánto más probable es encontrar un resultado negativo en pacientes enfermos *vs*. sanos?

Pero ¿qué información aportan realmente estos indicadores de validez?

Para calcular los valores esperados de sensibilidad y especificidad se estiman los valores esperados en cada casilla de la tabla de contingencia 2 × 2 tal y como se haría para un análisis de frecuencias por Ji-cuadrado. El valor esperado al azar para la casilla "a" será entonces:

$$E(a) = \frac{(a+b) * (a+c)}{T}$$

Siendo

E (a): el valor esperado al azar;

(a+b): el subtotal esperado para la fila donde se encuentra la casilla de interés (a);

(a+c): el subtotal esperado para la columna donde se encuentra la casilla de interés (a).

En la tabla 2 se muestran los valores observados y esperados para el estudio de *Tenezaca Sari y cols.*[13]

Tabla 2. Valores esperados y observados para el estudio de *Tenezaca Sari y colaboradores*[13]

Secuencia de Murphy	Observados		Subtotal	Esperados		Subtotal
	Enfermo	Sano		Enfermo	Sano	
Sí	200	6	206	192,26	13,74	206,00
No	94	15	109	101,74	7,26	109,00
Subtotal	294	21	315	294,00	21,00	315,00

Al calcular la sensibilidad y especificidad esperadas quedarían del siguiente modo:

$$Se = \frac{192{,}26}{294{,}0} * 100 = 65{,}4\ \%$$

$$Ee = \frac{7{,}26}{21{,}0} * 100 = 34{,}6\ \%$$

Siendo:

Se: sensibilidad esperada;

Ee: especificidad esperada.

Los resultados denotan que no existe mucha diferencia respecto a la sensibilidad esperada al azar y la observada en la práctica, lo que de antemano limita ya la validez de este estudio. Se observa, además, cierta diferencia entre la especificidad esperada y observada (34,6 % vs. 71,43 % respectivamente).

Se calcula el coeficiente de concordancia (*kappa*) entre lo esperado y lo observado mediante las fórmulas siguientes [adaptado de 14]:

$$ks = \frac{|S - Se|}{(100 - Se)} = 0{,}08$$

$$ke = \frac{|E - Ee|}{(100 - Ee)} = 0{,}56$$

Siendo:

S y *Se*: sensibilidades observadas y esperadas respectivamente;

E y *Ee*: especificidades observadas y esperadas respectivamente;

ks y *ke*: coeficientes *kappa* ponderados para sensibilidad y especificidad respectivamente.

Con los valores de *kappa* obtenidos y empleando la escala de interpretación de este parámetro, aportada por Landis y Koch (cuadro 1) puede concluirse que el indicador de sensibilidad apenas se diferencia con el esperado, con una *ks* evaluada como mala. Por su parte, el indicador especificidad obtiene valores clasificados de moderados.

Cuadro 1. Escala de Landis y Koch para evaluar el grado de concordancia entre dos observaciones [Adaptada de 15]

Coeficiente kappa	Interpretación
< 0,40	Pobre o débil
0,41 – 0,60	Moderada
0,61 – 0,80	Buena
> 0,80	Muy buena

Los resultados de *kappa* ponen en duda la conclusión inicial del estudio, puesto que, según lo obtenido, la secuencia de Murphy sería relativamente aceptable para detectar la ausencia apendicitis aguda, más no para la presencia de esta enfermedad. Igualmente se puede inferir que los valores de sensibilidad y especificidad, por sí solos, no denotan calidad de la prueba diagnóstica al ser afectados significativamente por la distribución al azar de los pacientes en la tabla de contingencia formada en el estudio realizado.

Valores predictivos: ¿indicadores de seguridad?

La seguridad de un test diagnóstico se relaciona con la capacidad que tiene el mismo para predecir la presencia o ausencia de la condición investigada. En realidad, en la práctica clínica es lo que realmente interesa al especialista: si el test resultara positivo, con qué probabilidad se puede afirmar que realmente se encuentra enfermo el paciente, o lo contrario, si el test fuese negativo. Los indicadores de "seguridad" que se emplean comúnmente son los valores predictivos negativo (VPN) y positivo (VPP) (ver fórmulas de cálculo en la tabla 1).[6,12]

La interpretación de los valores predictivos de una prueba diagnóstica es muy sencilla y fácil de aplicar. En la investigación de *Tenezaca Sari y cols.,*[13] se obtuvo un valor predictivo positivo de 97,09 %, lo que sugiere que de cada 100 casos que resultaron positivos al aplicar la secuencia de Murphy, 97 presentaron realmente apendicitis aguda (pocos falsos positivos). También mostró un valor predictivo negativo de 13,76 % que indica que de cada 100 casos que se reportan como negativos para el test evaluado, solo 13 a 14 pacientes realmente resultan no tener esta enfermedad (muchos falsos negativos). Con estos resultados, el test predice muy bien si un paciente presenta la condición estudiada, pero no permite descartarla con seguridad.

Los resultados anteriores no consideran que los valores predictivos dependen de la prevalencia de la condición analizada. Para evaluar el efecto de esta nueva variable se empleará el teorema de Bayes, que tiene en cuenta la probabilidad *a priori* de presentar la prevalencia de la enfermedad de interés. Aplicando sus

postulados, el cálculo de los valores predictivos quedaría denotado por las fórmulas siguientes:[16]

$$VPP = \frac{S * P}{S * P + (1 - E) * (1 - P)}$$

$$VPN = \frac{E * (1 - P)}{(1 - S) * P + E * (1 - P)}$$

Donde:

VPP: valor predictivo positivo;

VPN: valor predictivo negativo;

E: especificidad;

S: sensibilidad;

P: prevalencia de la enfermedad.

Con estas fórmulas y con los valores de sensibilidad y especificidad reportados en el estudio se calcularon VPP y VPN, dependiendo de la prevalencia de la enfermedad. Sus resultados se muestran en la figura 1.

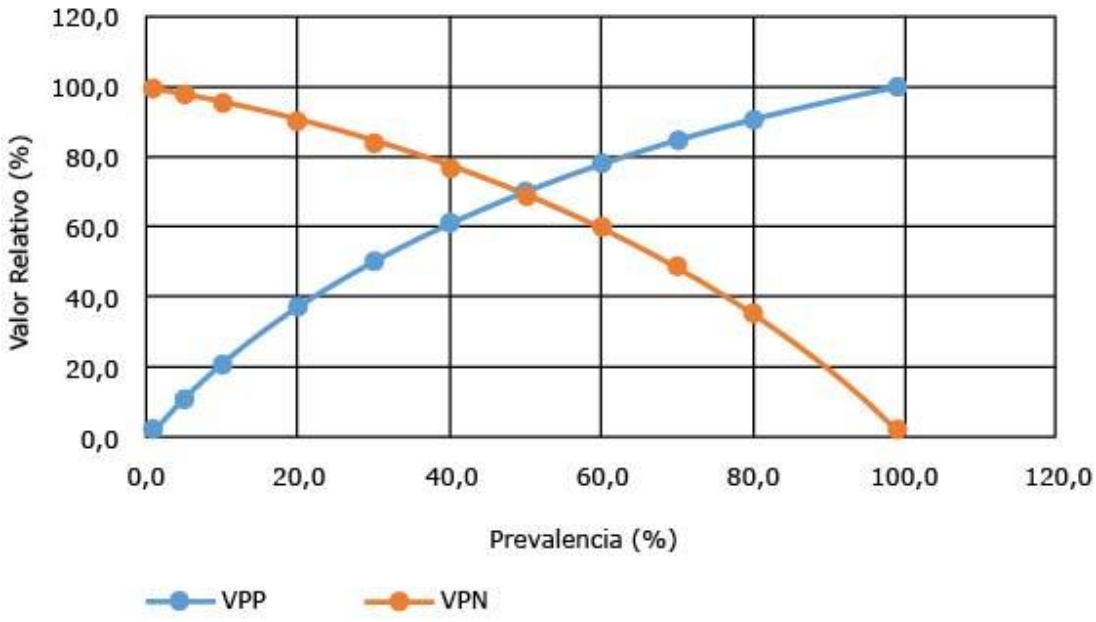

Fig. 1. Efecto de la prevalencia de apendicitis aguda sobre los valores predictivos de la secuencia de Murphy.

Los valores predictivos de la secuencia de Murphy varían de forma inversa cuando se modifica la prevalencia en la población de estudio; es decir, a elevadas prevalencias se obtienen buenas predicciones de la presencia de la enfermedad, mientras que la predicción para individuos sanos es baja. Este resultado se invierte en poblaciones con prevalencias inferiores al 20 %.

Se pudiera llegar a la conclusión de que para pacientes preoperatorios con sospecha de apendicitis aguda donde después de una anamnesis, revisión de historia clínica y examen físico riguroso, la probabilidad de que se presente esta condición es elevada, una secuencia positiva predice eficientemente la enfermedad, pero no puede descartar adecuadamente a los sanos cuando resulte negativa (muchos falsos negativos). Por su parte, en un nivel de atención primaria de salud, donde la prevalencia es relativamente baja, la ausencia de la secuencia de Murphy logra descartar eficientemente esta condición cuando no se presenten los signos (alto VPN), pero resultaría en muchos falsos positivos (bajo VPP).

Proporción correcta de aciertos: ¿un verdadero índice de validez?

Este indicador se refiere a la proporción de individuos que son correctamente clasificados en enfermos y sanos por el test evaluado. Es una medida del acuerdo encontrado entre este el test y la prueba de oro.[17] La forma de calcularlo se muestra en la tabla 1.

En múltiples estudios se ha revisado la utilidad real de este indicador, denotándose su dependencia de la prevalencia de la enfermedad y de la sensibilidad y especificidad de la prueba diagnóstica.[17,18] De este modo, aparte de la fórmula presentada en la tabla 1, se maneja otra más general:

$$IV = P * S + (1 - P) * E$$

Donde:

IV: índice de validez;

P: prevalencia;

S: sensibilidad;

E: especificidad.

Con los datos del estudio evaluado hasta el momento, el *IV* cambiaría levemente desde 71,3 % en una población con una prevalencia de 1 % de apendicitis aguda hasta 68,1 % en otra con una proporción de enfermos de 99 %. Esto no es tan alarmante como lo que ocurre con los valores predictivos analizados. Pero, ¿es esto suficiente para considerar adecuado este índice de validez, como lo proponen *Tenezaca Zari y cols.?*[13]

Streiner[18] hace alusión a que el *IV* de forma similar a lo que ocurre con la sensibilidad y especificidad, tampoco tiene en cuenta el acuerdo obtenido al azar y nuevamente se recurre al coeficiente de concordancia *kappa* para evaluar con mayor seguridad este aspecto. Así propone la fórmula siguiente:

$$k = \frac{|No - Ne|}{N - Ne}$$

Donde:

k: coeficiente kappa

No: frecuencia de aciertos observada

Ne: frecuencia de aciertos esperada al azar

N: total de individuos evaluados.

Con esta fórmula y con los datos de la tabla 2 se obtiene un valor de concordancia entre ambas pruebas de *k*= 0,13 que es significativamente diferente de cero (*p= 0,002*), pero es clasificado de bajo según Landis y Koch (tabla 3). Este resultado va en contra de las conclusiones del estudio evaluado, dado el escaso nivel de concordancia de la secuencia de Murphy con la prueba de oro, observándose que el índice de validez no es una medida muy eficaz para valorar el desempeño de un test diagnóstico. En el artículo no viene tabla 3. ***Ichi***

Razón de verosimilitud: ¿es realmente un apoyo?

En el accionar de los profesionales de la salud encargados de diagnosticar y tratar enfermedades, es de vital importancia discriminar adecuadamente entre la condición de estudio y otras con un cuadro de presentación similar. Para ello, todo galeno parte de una probabilidad pretest de que el individuo analizado tenga la enfermedad sospechada, probabilidad que en el mejor de los casos permite al especialista descartar el diagnóstico presuntivo o iniciar tratamiento.

Por lo general el médico se encuentra en situaciones en las que la evidencia obtenida a través el examen físico, la anamnesis y la historia clínica no es suficiente para descartar un diagnóstico o iniciar un proceso terapéutico, encontrándose en una zona de incertidumbre diagnóstica (Fig. 2). En tales casos necesita obtener información adicional que le permita mover esta probabilidad inicial hacia la zona de descarte o a la de tratamiento, lo que se realiza precisamente mediante las pruebas diagnósticas.[5,7,19]

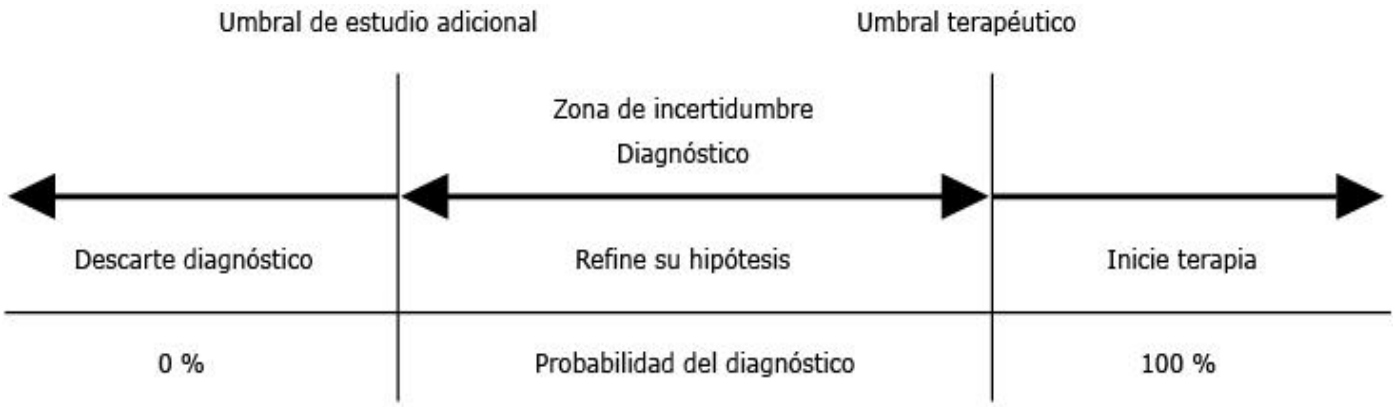

Fig. 2. Zona de incertidumbre diagnóstica, umbrales de prueba complementaria y de tratamiento.[19]

Los *Likelihood Ratio* (LR), cocientes de probabilidades (CP), razón de verosimilitud (RV), o índice de eficiencia diagnóstica (IED), ayudan a evaluar la probabilidad postest en la toma de decisiones terapéuticas. Así LR+ se define como cuánto más probable es que el test resulte positivo en un individuo enfermo que en uno sano, y LR- como cuánto más probable es encontrar un resultado negativo del test en una persona enferma respecto a otra sana.[6,20]

La forma de calcular LR se muestra en la tabla 1, pero otra manera de estimarlas es a partir de los valores se sensibilidad y especificidad, tal y como se muestra a continuación:

$$LR+ = \frac{S}{1-E}$$

$$LR- = \frac{1-S}{E}$$

Donde:

LR+= razón de verosimilitud positiva;

LR-= razón de verosimilitud negativa;

S= sensibilidad;

E= Especificidad.

Aplicando estas fórmulas a los datos, se estiman los siguientes valores para el estudio de Tenezaca Sari y cols.:[13] LR += 2,38; LR-= 0,45. En esencia los resultados muestran que es 2,38 veces más probable encontrar el test positivo en una persona enferma que en una sana, y la probabilidad de encontrar un test negativo en personas enfermas es aproximadamente dos veces menor a lo que se observaría en personas sanas.

Con los resultados de LR podría decirse que el test parece adecuado por la mayor probabilidad de resultar positivo en enfermos que en sanos, y negativo con más frecuencia en sanos que en enfermos. Pero, ¿cuán significativa es esta asociación desde el punto de vista clínico? Un valor de LR= 1 indica que la prueba no tiene ningún poder discriminante, el mismo que se elevará a medida que este indicador se aleje de la unidad. Varios autores registran una escala sencilla para valorar el poder discriminante de LR, teniendo en cuenta los cambios entre la probabilidad pretest y postest. Estos valores se presentan en el cuadro 2.[3,7,21]

Cuadro 2. Interpretación de los valores de LR

Valores de LR	Interpretación
LR+ >10 o LR- < 0,1	La aplicación de la prueba generalmente produce cambios elevados entre la probabilidad pretest y postest, suficientes para tomar una decisión diagnóstica o terapéutica.
LR+ de 5 - 10 y LR- de 0,1-0,2	El cambio de la probabilidad pretest a la probabilidad postest es moderada.
LR+ de 2 - < 5 y LR- de > 0,2-0,5	El cambio entre ambas probabilidades es pequeño o escaso.
LR+ de 1 - < 2 y LR- de > 0,5-1	El cambio es insignificante.

Con estas escalas se nota que la secuencia de Murphy aporta poco a la decisión diagnóstica, pues los cambios entre la probabilidad pretest y postest son bajos.

Otro modo más gráfico de observar los cambios entre la probabilidad pretest y postest, es a través del nomograma de Fagan. Esta herramienta consta de tres columnas o líneas graduadas: la primera referida a la probabilidad pretest, la misma que por lo general se asume como la prevalencia de la enfermedad en la población; la segunda indica el valor de LR+ o LR- obtenidos y la tercera es la probabilidad postest. Con los datos para las dos primeras columnas se puede trazar una línea recta que al interceptar con la tercera columna indica el valor de probabilidad postest, tal y como se muestra en la figura 3.[22]

Se debe aclarar que la probabilidad postest no es otra cosa que el valor predictivo derivado de la aplicación del teorema de Bayes a diferentes prevalencias de la enfermedad. La probabilidad de estar afectado cuando el test resulte positivo se corresponde con VPP y la probabilidad de estar enfermo con un test negativo es 100-VPN.

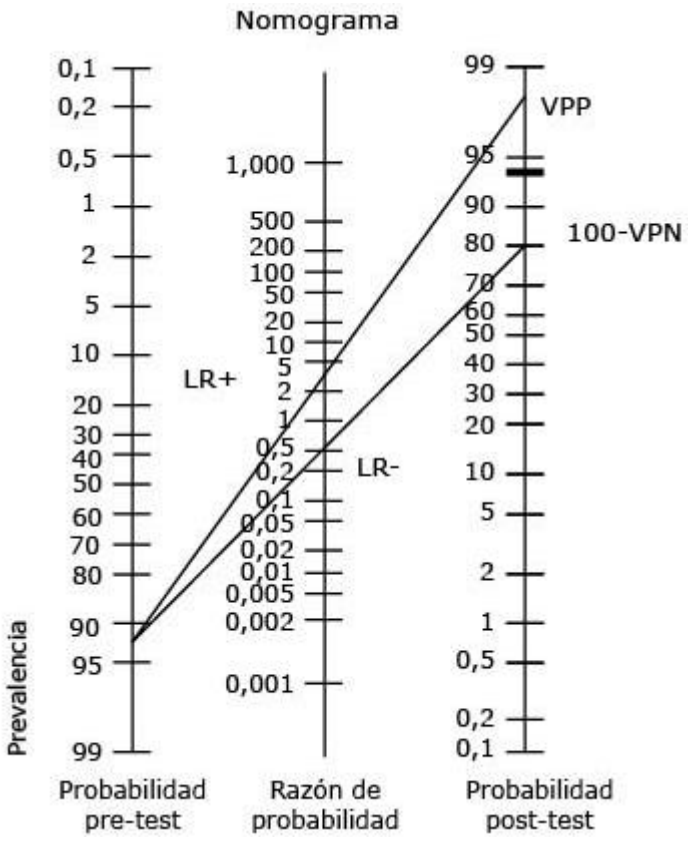

Fig. 3. Normograma de Fagan para estimar la probabilidad postest en el diagnóstico de apendicitis aguda, aplicando la secuencia de Murphy.[22]

Con los resultados de la figura 3 se puede ultimar que en el estudio analizado, una secuencia de Murphy positiva no eleva en gran medida la probabilidad de estar enfermo (*P*pretest= 93,0 %, prevalencia de la enfermedad en la muestra de estudio; *P*postest= 96,9 %). Por su parte, un paciente con un resultado negativo en el test todavía tiene una probabilidad de estar enfermo de más del 80 %, lo que dada la urgencia médica de una apendicitis aguda, es inaceptable. Estos datos

denotan una limitación de las LR en especial para prevalencias extremas donde se requerirían valores de LR+ mucho mayor de 10 y LR- mucho menor de 0,1 para apreciar un cambio significativo en la probabilidad pos-test de estar enfermo.7

¿Son suficientes los indicadores del rendimiento como evidencias de validez de una prueba diagnóstica?

Se han descrito varias metodologías para evaluar los estudios sobre la validación de pruebas diagnósticas, difiriendo la mayoría en la amplitud y la forma en cómo tratan el tema. Estos trabajos concuerdan en al menos tres detalles relevantes: 1) *la validez de la metodología*, 2) *el análisis de los resultados*, 3) *la aplicabilidad real de los resultados*.[1,4,6,23-26] El segundo de estos aspectos se amplió anteriormente, por lo que se analizarán los puntos 1 y 3.

La *validez metodológica* de un estudio sobre procedimientos diagnósticos depende primeramente de la selección adecuada de los pacientes. Si se evalúa el test en una muestra de individuos con cuadros sugerentes de enfermedad respecto a otros sanos, es muy probable que el resultado sea un buen desempeño de la prueba cuando realmente no lo es. Este problema se reconoce como un sesgo de selección, evitable si la muestra también incluye de manera apropiada a pacientes con un espectro clínico muy similar a la patología de interés, que es lo que comúnmente mueve la probabilidad pretest hacia una zona de incertidumbre.[4,5] Debe tenerse en cuenta los diferentes niveles de gravedad de una enfermedad, muchas veces es más fácil detectarla en estadios tardíos o graves (sobreestimándose S y E) y lo contrario en estadios leves.[3]

Otro de los elementos a tomar en cuenta en el diseño de estos estudios se relaciona con la selección del *Gold Standard*. La prueba de "oro" como criterio de verdad, debe ser avalada por la comunidad médica y científica como la más apta para detectar a los individuos sanos y enfermos en todo el espectro posible de presentación. Ambas pruebas (la de "oro" y la de estudio) deben evaluarse por separado y a ciegas, esto es, los dos test deben realizarse siempre y por personas diferentes, que no conozcan los resultados del otro.[3-5,26]

En cuanto a la *aplicabilidad de los resultados,* ésta se enfoca sobre todo a responder tres preguntas fundamentales:[26]

a) ¿Son reproducibles los resultados y la interpretación del test en el medio en el que se empleará? Debe tenerse en cuenta que el espectro de pacientes en los que se va a aplicar debe ser similar a aquel donde se realizó el estudio. Asimismo, los especialistas que aplicarán la prueba deben estar lo suficientemente capacitados y entrenados para evitar el sesgo por variabilidad en la interpretación de los datos obtenidos. Este error se produce cuando el resultado de un test no aporta un valor numérico por un procedimiento objetivo de medición, sino más bien depende de la pericia de un especialista en detectar los signos y síntomas propios de una enfermedad, o en interpretar una radiografía, entre otras. En tales casos se tienen diferentes observadores, cada uno de los cuales muestra diferente Sensibilidad y Especialidad, que cambian constantemente con el tiempo por la influencia del entrenamiento en la detección de la patología.

b) ¿Los resultados del test conducen a un cambio de estrategia en el manejo del paciente? Esto se relaciona con el cambio de probabilidad postest de estar o no enfermo una vez que se aplica la prueba diagnóstica. Como se indicó

previamente, existirán situaciones en las que la probabilidad pretest será tan baja que, independientemente del resultado de la prueba diagnóstica, la probabilidad postest será igualmente baja y se decida no tratar al paciente. Lo contrario ocurriría cuando la probabilidad pretest es muy alta.

c) ¿Estarán mejor los pacientes con el resultado de la prueba? Esta se enfoca más hacia el análisis bioético de la misma. No tendría sentido realizar una prueba si los riesgos superan los beneficios, incluyendo el aspecto económico, especialmente cuando su aplicación conlleva a implementar una estrategia terapéutica sin repercusión sobre la recuperación del paciente.

CONCLUSIONES

Los estudios para validar las pruebas diagnósticas deben seguir una revisión más rigurosa que la establecida hasta el momento en algunas revistas científicas. Los indicadores estadísticos del desempeño de las mismas son un apoyo indispensable durante su evaluación, pero los editores y especialistas en general, deben abordar otros aspectos como la distribución al azar de los pacientes, la prevalencia de la enfermedad a diagnosticar, el diseño metodológico de la investigación y su aplicabilidad. Entre los indicadores matemáticos que ayudarían, se tiene el índice *kappa,* para determinar concordancia entre la prueba de oro y el test evaluado, así como para valorar los índices de sensibilidad y especificidad. Debe exigirse en estos estudios la razón de verosimilitud, pero teniendo especial cuidado en el análisis, cuando la prevalencia de la enfermedad o probabilidad pretest es muy baja o muy alta. Se recomienda que se reporte la probabilidad de presentar la condición de interés antes y después de la prueba diagnóstica.

Conflicto de intereses

Los autores no reportan conflicto de intereses.

Abreviaturas: S: Sensibilidad; E: Especificidad; Se: Sensibilidad esperada al azar; Ee: Especificidad esperada al azar; VPP: Valor Predictivo positivo; VPN: Valor Predictivo Negativo; IV: Índice de Validez; k: Coeficiente Kappa de Cohen; P pretest y P postest: probabilidades pretest y postest respectivamente; LR+ y LR-: razones de verosimilitud positiva y negativa respectivamente.

REFERENCIAS BIBLIOGRÁFICAS

1. Burgos ME, Manterola C. Cómo interpretar un artículo sobre pruebas diagnósticas. Rev. Chil. Cir. 2010 [citado 20 Ene 2016];62(3):301-8. Disponible en: http://www.scielo.cl/pdf/rchcir/v62n3/art18.pdf

2. Ciapponi A. QUADAS-2: instrumento para la evaluación de la calidad de estudios de precisión diagnóstica. EVIDENCIA-Actualización en la Práctica Ambulatoria. 2015 [citado 23 Ene 2016];18(1):22-6. Disponible en: http://www.foroaps.org/files/64fe85009abba8c506e903adf90dbc17.pdf

3. Ochoa Sangrador C, González de Dios J, Buñuel Álvarez JC. Evaluación de artículos científicos sobre pruebas diagnósticas. Evid. Pediatr. 2007 [citado 2 Ene 2016];3:24-9. Disponible en: http://www.aepap.org/EvidPediatr/numeros/vol3/2007_numero_1/pdf/2007_vol3_numero1.24.pdf

4. Vera C, Letelier LM, Carvajal J. Guía para el análisis crítico de estudios que evalúan exámenes diagnósticos. Rev. Chil. Obstet. Ginecol. 2005 [citado 4 Ene 20166];70(3):196-202. Disponible en: http://www.scielo.cl/pdf/rchog/v70n3/art12.pdf

5. Valenzuela L, Cifuentes L. Validez de estudios de tests diagnósticos. Rev. Méd. Chile 2008 [citado 24 Ene 2017];136:401-4. Disponible en: http://dx.doi.org/10.4067/S0034-98872008000300018

6. Donis JH. Evaluación de la validez y confiabilidad de una prueba diagnóstica. Avances en Biomedicina. 2012 [citado 24 Ene 2017];1(2):73-81. Disponible en: http://www.redalyc.org/pdf/3313/331328015005.pdf

7. Bravo Grau S, Cruz JP. Estudios de exactitud diagnóstica: Herramientas para su Interpretación. Rev. Chil. Radiol. 2015 [citado 24 Ene 2017];21(4):158-64. Disponible en: http://www.scielo.cl/pdf/rchradiol/v21n4/art07.pdf

8. Araujo Alonso M. Análisis crítico de estudios de pruebas diagnósticas: I. Medwave 2012;12(7):e5465.

9. Sánchez Pedraza R. Aspectos sobre diseño y tamaño de muestra en estudios de pruebas diagnósticas. Rev. Fac. Med. UN Col. 2001 [citado 24 Ene 2017];49(3):175-80. Disponible en: http://www.bdigital.unal.edu.co/23066/1/19786-65851-1-PB.pdf

10. Cruz Tabuenca H. Sobre el error diagnóstico. Artículo de revisión. Medicina Naturista 2014 [citado 24 Ene 2017];8(1):53-9. Disponible en: https://dialnet.unirioja.es/descarga/articulo/4560700.pdf

11. Carnero-Pardo C. Evaluación de las pruebas diagnósticas. Rev Neurol. 2005 [citado 27 Ene 2017];40(11):641-3. Disponible en: http://www.publicacions.ub.es/refs/Articles/avaluaciopd.pdf

12. Pita Fernández S, Pértegas Díaz S. Pruebas diagnósticas: Sensibilidad y especificidad. Cad. Aten. Primaria. 2003 [citado 27 Ene 2017];10:120-4. Disponible en: https://www.fisterra.com/mbe/investiga/pruebas_diagnosticas/pruebas_diagnosticas.asp

13. Tenezaca-Sari X, Sánchez P, Beltrán L, Tenezaca-Tacuri, A. Validación de la Secuencia de Murphy en el Diagnóstico Clínico de Apendicitis Aguda. Hospital Vicente Corral Moscoso. 2013. Rev. Med. HJCA 2016 [citado 24 Ene 2017];8(2):165-9. Disponible en: http://dx.doi.org/10.14410/2016.8.2.ao.27

14. Cabello López JB, Pozo Rodríguez F. Estudios de evaluación de las pruebas diagnósticas en cardiología. Rev. Esp. Cardiol. 1997 [citado 24 Ene 2017];50:507-19. Disponible en: http://appswl.elsevier.es/watermark/ctl_servlet?_f=10&pident_articulo=496&pident_usuario=0&pcontactid=&pident_revista=25&ty=154&accion=L&origen=cardio&web=www.revespcardiol.org&lan=es&fichero=C500708.PDF&anuncioPdf=ERROR_publi_pdf

15. Cortés-Reyes Édgar, Rubio-Romero Jorge Andrés, Gaitán-Duarte Hernando. Métodos estadísticos de evaluación de la concordancia y la reproducibilidad de pruebas diagnósticas. Rev Colomb Obstet Ginecol. 2010 [citado 24 Ene 2017];61(3):247-55. Disponible en: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0034-74342010000300009&lng=en

16. Fernández Regalado R. El teorema de Bayes y su utilización en la interpretación de las pruebas diagnósticas en el laboratorio clínico. Rev. Cubana Invest. Bioméd. 2009 [citado 24 Ene 2017];28(3):158-65. Disponible en: http://scielo.sld.cu/pdf/ibi/v28n3/ibi13309.pdf

17. Alberg AJ, Park JW, Hager BW, Brock MV, Diener-West M. The Use of "Overall Accuracy" to Evaluate the Validity of Screening or Diagnostic Tests. J. Gen. Intern. Med. 2004 [citado 24 Ene 2017];19:460-5. Disponible en: http://www.scielo.cl/pdf/rmc/v136n3/art18.pdf

18. Streiner DL. Diagnosing Tests: Using and Misusing Diagnostic and Screeing Tests. Journal of Personality Assessment. 2003 [citado 24 Ene 2017];81(3):209-19. Disponible en: http://www.snmmi.org/files/docs/Research_E_Library/General/test%20statistics.pdf

19. Medina MC. Generalidades de las pruebas diagnósticas, y su utilidad en la toma de decisiones médicas. Rev Colomb Psiquiat. 2011 [citado 24 Ene 2017];40(4):787-97. Disponible en: http://www.scielo.org.co/pdf/rcp/v40n4/v40n4a15.pdf

20. Salech F, Mery V, Larrondo, F, Rada G. Estudios que evalúan un test diagnóstico: interpretando sus resultados. Rev Méd Chile. 2008 [citado 24 Ene 2017];136:1203-8. Disponible en: http://www.scielo.cl/pdf/rmc/v136n9/art18.pdf

21. Grimes D, Schulz K. Refining clinical diagnosis with likelihood ratios. Lancet 2005 [citado 24 Ene 2017];365:1500-5. Disponible en: http://www.mwc.com.br/files/L-Epid02-03-Grimes[A784]%20-%20Likelihood%20ratios.pdf

22. Aznar-Oroval E, Mancheno-Alvaro A, García-LozanoT, Sánchez-Yepes M. Razón de verosimilitud y nomograma de Fagan: 2 instrumentos básicos para un uso racional de las pruebas del laboratorio clínico. Rev. Calid. Asist. 2013 [citado 24 Ene 2017];28(6):390-3. Disponible en: http://www.elsevier.es/es-revista-revista-calidad-asistencial-256-articulo-razon-verosimilitud-nomograma-fagan-2-S1134282X13000523

23. Sackett DL, Haynes RB. The architecture of diagnostic research. BMJ 2002 [citado 24 Ene 2017];324:539-41. Disponible en: http://pubmedcentralcanada.ca/pmcc/articles/PMC1122451/pdf/539.pdf

24. Manterola C, Otzen T, Lorenzini N, Díaz A, Torres-Quevedo R, Claros N. Iniciativas Disponibles para el Reporte de Resultados en Investigación Biomédica con Diferentes Tipos de Diseño. Int. J. Morphol. 2013 [citado 24 Ene 2017];31(3):945-56. Disponible en: http://www.scielo.cl/pdf/ijmorphol/v31n3/art29.pdf

25. Altman D, Bossuyt P. Estudios de precisión diagnóstica (STARD) y pronóstica (REMARK). Med. Clin. (Barc.) 2005 [citado 24 Ene 2017];125(Suppl 1):49-55. Disponible en: https://www.researchgate.net/profile/Patrick_Bossuyt/publication/246616449_Estudios_de_precision_diagnostica_STARD_y_pronostica_REMARK/links/54f6d1d40cf2ca5efeff5ad9/Estudios-de-precision-diagnostica-STARD-y-pronostica-REMARK.pdf

26. Rivera S, Letelier LM. Aplicabilidad de un estudio sobre tests diagnósticos. Rev. Méd. Chile. 2011 [citado 24 Ene 2017];139(5):672-5. Disponible en: http://www.scielo.cl/pdf/rmc/v139n5/art17.pdf

Recibido: 4 de septiembre de 2017.
Aprobado: 1ro. de diciembre de 2017.

Dariel Díaz Arce. Unidad Educativa Santana, Ecuador.
Dirección electrónica: ddiaz@santana.edu.ec