

## Estudios que evalúan un test diagnóstico: interpretando sus resultados

Felipe Salech<sup>1,a</sup>, Victoria Mery<sup>1,b</sup>, Francisco Larrondo<sup>1,b</sup>, Gabriel Rada<sup>1,2,3</sup>.

### *Studies about diagnostic tests: interpreting the results*

En términos generales, un test diagnóstico es útil si permite diferenciar dos o más condiciones que de otro modo podrían ser confundidas. En otras palabras, para diferenciar entre distintas enfermedades o condiciones clínicas, así como entre la condición de sano y la de enfermo<sup>1</sup>.

En los estudios sobre tests diagnósticos, al igual que en el análisis crítico de cualquier estudio, el primer paso es evaluar su validez, es decir, cuál es la probabilidad de que exista sesgo por características del diseño. Este aspecto ha sido revisado en artículos anteriores de esta serie<sup>2-4</sup>.

Una vez definida la validez del estudio, el siguiente paso será analizar la correcta interpretación de los resultados presentados en ellos, ya que es posible que un estudio cumpla con todas las características que aseguren su validez, sin embargo, si los resultados muestran que carece de capacidad de discriminar entre las condiciones de interés, éste no tendrá utilidad.

#### CONCEPTOS GENERALES

Recordaremos algunos conceptos generales.

*Gold standard (GS):* El rendimiento de todo test diagnóstico se basa en su comparación con un *gold standard (estándar de oro, patrón de oro, patrón de referencia)*. El GS es la técnica diagnóstica que define la presencia de la condición con la máxima certeza conocida. Debido a la falta de consenso en la forma de traducir este concepto, utilizaremos su denominación en inglés.

*Valores posibles de un test diagnóstico:* Algunos tests entregan resultados binarios o dicotómicos, generalmente positivo o negativo (ej: test pack de embarazo). Algunos se expresan como resultados categóricos (ej: alta, moderada y baja probabilidad de un cintigrama V/Q). Otros, en cambio, entregan resultados continuos (ej: glicemia, colesterol, hemoglobina). Estos valores

<sup>1</sup>Unidad de Medicina Basada en Evidencia, Pontificia Universidad Católica de Chile.

<sup>2</sup>Departamento de Medicina Interna, Pontificia Universidad Católica de Chile.

<sup>3</sup>Servicio de Medicina, Hospital Sótero del Río, Santiago de Chile.

<sup>a</sup>Residente. Hospital Clínico Universidad de Chile.

<sup>b</sup>Residente. Pontificia Universidad Católica de Chile. Santiago de Chile.

continuos pueden ser transformados en binarios si se establece un punto de corte a partir del cual se considerarán los resultados como positivos o negativos para la presencia de la condición (ej: glicemia mayor a 125 mg/dl) o como categóricos, si se establecen rangos, como discutiremos más adelante.

FORMAS DE PRESENTAR LAS PROPIEDADES DE UN TEST

Al comparar un test diagnóstico con un GS, se pueden obtener cuatro combinaciones si los resultados del test se expresan en forma binaria:

1. Verdadero positivo: GS positivo, test positivo
2. Verdadero negativo: GS negativo, test negativo.
3. Falso positivo: GS negativo, test positivo.
4. Falso negativo: GS positivo, test negativo.

Esto se puede resumir en una tabla de contingencia de 2 x 2 (Tabla 1).

A partir de la tabla, se pueden calcular distintas formas de expresar el poder de discriminación o rendimiento de un test diagnóstico. Cada una tiene ventajas y desventajas, muchas veces entregando información complementaria. A continuación, revisaremos las más usadas en la literatura médica sobre estudios diagnósticos.

1. Sensibilidad:

Se define como la razón entre los individuos que tienen un resultado del test positivo y aquellos que tienen la condición o enfermedad de interés (los verdaderos positivos sobre el total de GS positivos) (Tabla 2).

En un paciente determinado, si aplicamos un examen altamente sensible (identifica muy bien a los enfermos) y obtenemos un resultado negativo, podemos descartar razonablemente la enfermedad<sup>5</sup>.

2. Especificidad:

Se define como la razón entre los individuos que tienen un resultado del test negativo y aquellos sin

la enfermedad de interés (verdaderos negativos sobre los GS negativos) (Tabla 2).

Si un paciente tiene un resultado positivo en un test altamente específico podemos confirmar la enfermedad.

3. Valor predictivo positivo y negativo:

El valor predictivo positivo se define como la probabilidad que un individuo con un resultado positivo, tenga la enfermedad<sup>6,7</sup>. Por el contrario, el valor predictivo negativo corresponde a la probabilidad que un individuo con un resultado negativo, no tenga la enfermedad (Tabla 2).

Si bien los valores predictivos, a diferencia de la sensibilidad y especificidad, nos entregan información clínicamente relevante (la probabilidad de que la condición esté o no presente dado el resultado del test), ésta sólo es utilizable si nos enfrentamos a pacientes similares a aquellos en que se realizó el estudio. Los valores predictivos varían enormemente dependiendo de la prevalencia o riesgo basal de la condición, por lo que si nuestro paciente tiene un riesgo mayor o menor, no podemos aplicarlos. Lo anterior no ocurre con la sensibilidad y especificidad, ya que su cálculo no depende de la prevalencia de la condición (al menos desde el punto de vista matemático). Esto ha hecho que constituyan una de las formas más frecuentes de expresar el rendimiento de un test.

En resumen podemos decir que:

- La sensibilidad y especificidad no varían con la prevalencia de la condición, pero no nos hablan de la probabilidad que tiene un paciente de presentar la enfermedad de interés.
- Los valores predictivos nos hablan de la probabilidad que tiene un paciente de presentar la enfermedad de interés, pero varían enormemente dependiendo de la prevalencia de la condición.

A lo anterior hay que agregar otra desventaja de utilizar estas medidas de rendimiento tradicio-

**Tabla 1. Tabla de contingencia para tests diagnósticos**

	<i>Gold standard +</i>	<i>Gold standard -</i>	Total
Test +	Verdadero positivo (A)	Falso positivo (B)	A + B
Test -	Falso negativo (C)	Verdadero negativo (D)	C + D
Total	A + C	B + D	

**Tabla 2. Resumen de las propiedades de un test diagnóstico**

Propiedad del test	Pregunta a responder	Fórmula	Comentarios
Sensibilidad	Qué tan bueno es el test detectando posibles enfermos.	$a/(a+c)$	Si un test es muy sensible, un resultado negativo descarta la enfermedad.
Especificidad	En sentido estricto nos dice qué tan bueno es el test en excluir a los sanos. Se entiende mejor puesto al revés (qué tan bueno es en confirmar enfermos).	$d/(d+b)$	Si un test es muy específico, un resultado positivo confirma la enfermedad.
VPP	Si una persona tiene el test positivo qué tan probable es que tenga la condición	$a/(a+b)$	Los valores predictivos varían mucho con la prevalencia. No utilizar, si nuestro paciente es diferente a los del estudio.
VPN	Si una persona tiene el test negativo qué tan probable es que no tenga la condición	$d/(c+d)$	Pueden calcularse tantos LR como valores posibles tiene un test. En el caso de un test dicotómico tendrá un valor positivo y uno negativo.
LR	Qué tanto más probable es encontrar determinado valor del test en alguien enfermo comparado con alguien sano		También se puede calcular de la siguiente forma: Sensibilidad / (1-especificidad)
LR +	Qué tanto más probable es encontrar el test positivo en alguien enfermo que en alguien sano	$(a/a+c)/(b/b+d)$	También se puede calcular de la siguiente forma: 1-sensibilidad / (especificidad)
LR -	Qué tanto más probable es encontrar el test negativo en alguien enfermo que en alguien sano	$(c/a+c)/(d/b+d)$	

nales. Para calcularlas, necesariamente necesitamos utilizar valores binarios (sí o no, positivo o negativo, presente o ausente), limitando su capacidad diagnóstica.

A continuación revisamos algunas formas de expresar el rendimiento del test, que intentan dar solución a las limitaciones de las tradicionales.

#### 4. Probabilidad pre test y post test

En un artículo previo se introdujeron los conceptos generales acerca de las probabilidades pre y post test, cómo se modifican de acuerdo a los resultados de un examen y cómo se incorporan a la toma de decisión<sup>3</sup>.

Todo paciente en que sospechemos una enfermedad, tendrá una probabilidad de presentarla. Esta dependerá de la prevalencia de la enfermedad en la población, de las características del paciente (edad, género, raza), de los signos y síntomas presentes, etc. Así, antes de realizar

cualquier test diagnóstico, el clínico (explícita o implícitamente) le asigna a su paciente una probabilidad “pre test” de presentar la enfermedad. Una vez realizado el test diagnóstico, esta probabilidad aumentará o disminuirá, dependiendo del resultado del test. A esta nueva probabilidad la llamaremos probabilidad “post test”.

Una forma de aproximarse a la probabilidad pre test en un paciente determinado, es utilizar la prevalencia de la enfermedad en el estudio que estamos analizando (total de pacientes con el GS positivo, o  $A + C$  en nuestra tabla de contingencia, sobre el total de pacientes del estudio o  $A+B+C+D$ ). Si nuestro paciente es similar a la población del estudio, sería razonable utilizar este valor.

Formas más precisas de estimar la probabilidad pre test, corresponden a estudios observacionales en la población de interés o estadísticas locales<sup>8</sup>.

Pocas veces contaremos con estudios del problema de interés, con las características exactas

que presenta nuestro paciente y en nuestra población específica (por ejemplo, la población que consulta al Servicio de Urgencia en que trabajo). A pesar de eso, utilizando la mejor evidencia de que se disponga, y complementándola con juicio clínico y experiencia, habitualmente se logrará una buena estimación.

La propiedad del test que nos permite cuantificar la magnitud y el sentido del cambio de nuestra probabilidad pre test según sea su resultado, es el *likelihood ratio* (razón de probabilidad o cociente de verosimilitud). Dado que no existe consenso acerca de la forma de traducir este término y a su diseminada utilización en inglés, también hemos decidido mantenerlo sin traducir.

5. *Likelihood Ratio* (LR)

Se define como la razón entre la probabilidad de tener determinado resultado del test en la población con la condición *versus* tener el mismo resultado en la población sin la condición. Es decir, la proporción de test positivos en los individuos con la condición en estudio dividido

por la proporción de test positivos en los individuos sin la condición en estudio (Tabla 2). En términos sencillos nos indica la magnitud y el sentido del cambio de la probabilidad pre a post test según sea el resultado del test diagnóstico<sup>1,9</sup>.

Si tomamos un test que tiene sólo dos valores posibles, positivo o negativo, tendremos un valor de LR (+), que representa la magnitud del cambio en caso de presentar un test positivo, y un LR (-), que representa la magnitud del cambio en caso de presentar un resultado negativo.

Si bien con un poco de matemática se puede hacer el cálculo, una forma más práctica de traducir el LR de un test en un cambio objetivo de la probabilidad pre a post test de un paciente determinado, es utilizando el Nomograma de Fagan (Figura 1a)<sup>10</sup>. Si hemos estimado la probabilidad pre test en determinado paciente, y conocemos el LR del test diagnóstico, basta con unir (con una regla) los puntos correspondientes de las 3 columnas del nomograma. La columna izquierda del nomograma representa la probabilidad pre test, la del centro el LR, y la de la derecha, la probabilidad post test<sup>1,9</sup>.

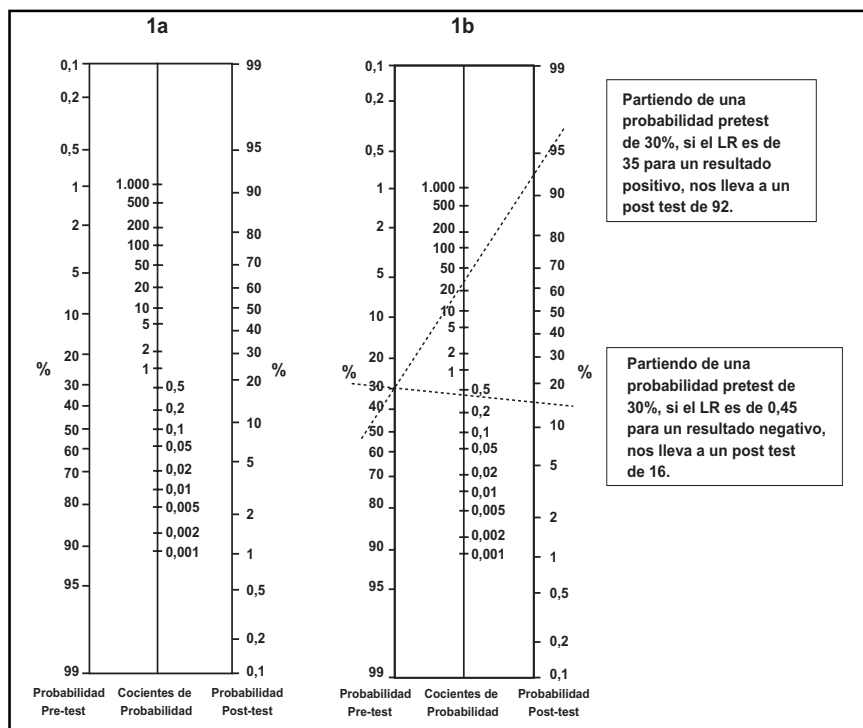


Figura 1. 1a. Nomograma de Fagan. 1b. Ejemplo de utilización del nomograma, basado en el ejemplo 1.

Como guía práctica, cuando un test tiene LR mayores a 10 o menores a 0,1, los cambios en las probabilidades serán en la mayoría de los casos, suficientes para confirmar (superar el umbral terapéutico) o descartar la condición de interés (superar el umbral diagnóstico o de estudio adicional)<sup>3</sup>.

Como describimos antes, algunos test tienen sólo 2 valores posibles (positivo y negativo), sin embargo, la mayoría tienen más de 2 valores, llegando a infinitos posibles valores en un test con resultados continuos.

Una de las principales ventajas del LR que lo diferencia de las otras propiedades del test, es que podemos obtener un LR distinto para cada valor del test, o para un rango de valores. Por ejemplo, si tenemos una enfermedad hipotética, en que el examen que la detecta tiene un valor de 0 en sujetos sanos, y a medida que el valor es mayor, la probabilidad de tener la enfermedad va aumentando; entonces, podemos calcular el LR para distintos rangos y así estimar cual será el aumento en la probabilidad con distintos valores. Así, un valor del test entre 1-10 podría tener un LR de 2, lo cual nos aumentaría un poco la probabilidad, un valor entre 11-20 tendría un LR mayor (digamos un LR de 5), y por tanto nos aumentaría un poco más la probabilidad. Finalmente, un valor de 50 tendría un LR tan alto (10 o más) que en la mayoría de los casos confirmaría la enfermedad. Es imposible hacer lo mismo con la sensibilidad/especificidad o con los valores predictivos.

Los LR permiten resumir y complementar, en un solo valor, dos propiedades de los test diagnósticos, la sensibilidad y la especificidad, y dado que su cálculo se hace a partir de ellos, su valor es independiente de la prevalencia de la condición en la muestra seleccionada.

#### EJEMPLOS DE *LIKELIHOOD RATIO*

A fin de clarificar estos conceptos, presentamos algunos ejemplos de LR.

*Ejemplo 1:* Imagine que está atendiendo a un paciente con un cuadro sugerente de meningitis tuberculosa y decide realizar un test de PCR para micobacterias en líquido cefalorraquídeo. Una revisión sistemática reporta que para esta técnica el LR de un valor positivo es de 35 y el de un valor

negativo es 0,45<sup>11</sup>. Si usted estimó una probabilidad pre test de 30% (por ejemplo, en base a su experiencia y la prevalencia de TBC en su región), un resultado positivo nos llevará a una probabilidad post test de 92%, certeza suficiente para iniciar tratamiento (supera el umbral terapéutico). Por el contrario, un resultado negativo disminuiría esta probabilidad a sólo 16% que no es suficiente para descartar el diagnóstico de meningitis tuberculosa (Figura 1b).

*Ejemplo 2:* Se presenta un paciente con cuadro dudoso de trombosis venosa profunda, y usted decide realizar un Dímero D. Una revisión sistemática de estudios diagnósticos reporta que para esta técnica los LR positivo y negativo respectivamente son 1,6 y 0,12<sup>12</sup>. Si usted consideró una probabilidad pre test de 10%, un resultado negativo nos llevará a una probabilidad post test cercana al 1%, certeza suficiente para descartar el diagnóstico. Por el contrario, un resultado positivo prácticamente no modificará la probabilidad (15%), por lo que se requerirán más estudios para confirmar la condición.

*Ejemplo 3:* En muchas ocasiones un test diagnóstico tiene más de dos posibles resultados. En estos casos debe presentarse el LR asociado a cada uno de ellos por separado. Por ejemplo, un estudio evaluó, entre otras cosas, el rol de la ferritina para el diagnóstico de anemia ferropriva<sup>13</sup>. Un valor de ferritina plasmática entre 45 y 100 mg/L mostró un LR de 0,54 por lo que rara vez modificará la probabilidad en forma importante. El LR asociado a un valor entre 35-45 fue 1,8, para 25-35 fue 2,5 y para 15-25 fue 9,3. Como se puede apreciar, valores de ferritina más bajos van cambiando en forma más importante la probabilidad. Finalmente un valor <15 se asoció a un LR de 55, por lo que prácticamente siempre confirmará el diagnóstico de anemia ferropriva.

*Ejemplo 4:* Para enfatizar la importancia de la probabilidad pre test, analicemos lo que ocurre con los test de *screening*, en donde ésta es generalmente muy baja. Imaginemos una mujer sana de 45 años que se realizó una mamografía en el contexto de un chequeo general, que informó una lesión BIRADS 4 (sospecha de malignidad). En una revisión sistemática el LR para este resultado en particular fue de 125 (lo cual sería excelente de acuerdo a la regla de oro que mencionamos anteriormente)<sup>14</sup>. Dado que la pro-

babilidad pre test para esta paciente de tener un cáncer de mama es de sólo 0,003%, la probabilidad post test sería de 0,37%. En otras palabras, a pesar que el LR asociado a este test es muy elevado, la baja probabilidad pre test hace que éste no sea suficiente para hacer el diagnóstico definitivo, por lo que se requerirán otros estudios para confirmar la presencia de cáncer.

#### PRECISIÓN DE LOS RESULTADOS:

Al igual que en un estudio de terapia, todo resultado en un estudio de test diagnóstico debe ser informado con su intervalo de confianza (el concepto de precisión de los resultados ha sido discutido en un artículo previo de esta serie)<sup>15</sup>.

El intervalo de confianza es el rango de valores dentro del cual se encuentra el valor verdadero (que no puede ser conocido de modo exacto) con un grado prefijado de certeza. Habitualmente se utiliza el "intervalo de confianza de 95%", que quiere decir que dentro de ese intervalo se encontrará el verdadero valor en 95% de los casos.

Cuanto más estrecho es el intervalo, mayor confianza tendremos para utilizar el resultado.

Los conceptos entregados en este artículo se resumen en la Tabla 2.

#### CONCLUSIÓN

Un test diagnóstico es útil en la medida que permite diferenciar dos o más condiciones que de otro modo podrían ser confundidas. Así, el test diagnóstico ideal es aquel que es capaz de detectar la mayor cantidad de pacientes con la condición, excluyendo a la vez a la mayor cantidad de pacientes sin ella.

Los resultados de un test diagnóstico nos ayudan a modificar la probabilidad de presentar o no una determinada condición en un paciente determinado. Existen diversas maneras de presentar las propiedades de un test, cada una con ventajas y desventajas.

El uso de los LR ayuda mejor a los clínicos en el proceso diagnóstico, al hacer explícito el cambio entre probabilidad pre y post test. Así, todos los estudios de test diagnóstico deberían entregar su valor, o al menos los datos que permitan su cálculo.

#### REFERENCIAS

1. JAESCHKE R, GUYATT G, LIJMER J. Diagnostic Tests. En: Guyatt G, Drummond R, ed. Users' guides to the medical literature. Essentials of evidence-based clinical practice. Chicago: Editorial: JAMA Press 2002; 187-217.
2. PANTOJA T, LETELIER LM, NEUMANN I. El análisis crítico de la información publicada en la literatura médica. *Rev Méd Chile* 2004; 132: 513-5.
3. CAPURRO D, RADA G. El proceso diagnóstico. *Rev Méd Chile* 2007; 135: 534-8.
4. VALENZUELA L, CIFUENTES L. Validez de estudios de tests diagnósticos. *Rev Méd Chile* 2008; 136: 401-4.
5. ALTMAN D, BLAND J. Diagnostic tests 1: sensitivity and specificity. *BMJ* 1994; 308: 1552.
6. ALTMAN D, BLAND J. Diagnostic tests 2: Predictive values. *BMJ* 1994; 309: 102.
7. LOONG T. Understanding sensitivity and specificity with the right side of the brain. *BMJ* 2003; 327: 716-9.
8. GUYATT GH, PATTERSON C, ALI M, SINGER J, LEVINE M, TURPIE I, MEYER R. Diagnosis of iron-deficiency anemia in the elderly. *Am J Med* 1990; 88: 205-9.
9. DEEKS J, ALTMAN D. Diagnostic tests 4: likelihood ratios. *BMJ* 2004; 329: 168-9.
10. FAGAN T. Normogram for Bayes's theorem. *N Engl J Med* 1975; 293: 257.
11. PAI M, FLORES L, PAI N, HUBBARD A, RILEY L, COLFORD J. Diagnostic accuracy of nucleic acid amplification tests for tuberculous meningitis: a systematic review and meta-analysis. *Lancet (infectious diseases)* 2003; 3: 633-43.
12. STEIN P, RUSSELL H, KALPESH P. D-dimer for the exclusion of acute venous thrombosis and pulmonary embolism. *Ann Intern Med* 2004; 140: 589-602.
13. GUYATT GH, OXMAN AD, ALI M, WILLAN A, McILROY W, PATTERSON C. Laboratory diagnosis of iron-deficiency anemia: an overview. *J Gen Intern Med* 1992; 7: 145-53.
14. KERLIKOWSKA K, SMITH-BINDMAN R, LJUNG B, GRADY D. Evaluation of abnormal mammography results and palpable breast abnormalities. *Ann Intern Med* 2003; 139: 274-84.
15. CANDIA R, CAIOZZI G. Intervalos de confianza. *Rev Méd Chile* 2005; 133: 1111-5.