Validación de pruebas diagnósticas

S. Bellmunt-Montoya

VALIDACIÓN DE PRUEBAS DIAGNÓSTICAS

Resumen. *El diagnóstico es parte fundamental de nuestra práctica diaria. Llegan a nuestras manos multitud de trabajos sobre pruebas diagnósticas y es imprescindible saber discriminar entre aquellos que nos pueden ser útiles y aquellos que sólo producen 'ruido de fondo'. Este trabajo pretende ofrecer una serie de herramientas e informaciones que pueden ayudar en el proceso de lectura crítica de estos estudios, para así seleccionar los que pueden ser útiles. [ANGIOLOGÍA 2007; 59: 433-8]*
Palabras clave. *Diagnóstico. Estudios de validación.*

Introducción

Prueba diagnóstica es toda aquella exploración que tiene como objetivo diferenciar entre salud y enfermedad o discernir entre diversas categorías de estas dos entidades. Cualquier exploración ha de ser validada antes de aplicarse, confirmando que sus resultados se ajustan al máximo a la realidad. Cuando leemos un artículo de validación hemos de saber interpretar los datos para valorar la calidad de la información que nos transmite. Existen herramientas que nos pueden guiar en el proceso de evaluación de los artículos que hablan de validación de pruebas diagnósticas [1-6]:

- *CASP* (*Critical Appraisal Skills Programme*): http://www.redcaspe.org/herramientas/index.htm. Su versión en español se denomina CASPe.
- *GATE (Graphic Appraisal Tool for Epidemiology):* http://www.health.auckland.ac.nz/population-health/epidemiology-biostats/epiq/GATEDx040106.xls.
- *Bossuyt et al* [1]: este texto puede ser consultado en Internet en la dirección: http://www.consort-statement.org/Initiatives/newstard.htm.

Para desarrollar este tema de una manera más práctica y amena, utilizaremos los puntos sugeridos por la herramienta CASPe, para así ir desgranando cada uno de los aspectos que nos ayudarán a evaluar los datos de un estudio de validación.

¿Son válidos los resultados del estudio?

Las tres primeras preguntas son de eliminación (si no se cumplen las expectativas, no vale la pena seguir leyendo el artículo) y las dos siguientes nos pueden dar argumentos (o inconvenientes) para seguir leyendo.

Preguntas de eliminación

1. ¿Existió una comparación con una prueba de referencia adecuada?

¿Es correcto el patrón oro?

No siempre se puede aplicar la prueba de referencia (patrón oro) o no siempre se puede aplicar a todos los

Aceptado tras revisión externa: 17.10.07.

Servicio de Angiología y Cirugía Vascular. Corporació Sanitària Parc Taulí. Sabadell, Barcelona, España.

Correspondencia: Dr. Sergio Bellmunt Montoya. Servicio de Angiología y Cirugía Vascular. Corporació Sanitària Parc Taulí. Parc Taulí, s/n. E-08208 Sabadell (Barcelona). E-mail: 31497sbm@comb.es

pacientes. Pueden existir limitaciones de diversa índole, siendo las principales las económicas y las éticas.

Si se utiliza un 'sucedáneo' como referencia estamos poniendo en riesgo la validez del estudio, es decir, no podemos confirmar que la prueba está midiendo lo que realmente ha de medir y con la exactitud adecuada.

La disponibilidad de dicha prueba de referencia puede condicionar el tipo de diseño del estudio. Teniendo en cuenta que el diagnóstico es un proceso puntual en el tiempo, el diseño más utilizado para validar una prueba es un estudio de corte. A pesar de ello, existen otros diseños que nos pueden aportar datos de interés: los estudios de cohortes y de casos controles.

El estudio de corte es el más habitual, ya que recrea la situación clínica real del diagnóstico. En este tipo de estudio se aplica la prueba que se quiere validar y la de referencia a una muestra representativa de la población diana. El hecho de que la probabilidad pretest (prevalencia) sea la de la población real, permite realizar una estimación de la exactitud de la prueba (sensibilidad y especificidad), además del comportamiento –valores pronósticos positivo (VPP) y negativo (VPN)–.

El estudio de cohortes se plantea cuando no puede realizarse la prueba de referencia sobre la muestra a todos los pacientes, tanto por razones económicas como éticas. Para ello, se divide la muestra en dos grandes grupos según el resultado de la prueba a validar (enfermos y no enfermos) y se siguen en el tiempo para observar la evolución de la enfermedad y confirmar la veracidad del diagnóstico. Los valores estadísticos que se utilizan son los propios de cualquier estudio de cohortes: el riesgo relativo (RR).

En el estudio de casos y controles se escoge un grupo de enfermos y otro de no enfermos que se clasificarían según el resultado de la prueba de referencia. Es entonces cuando aplicaremos sobre todos ellos la prueba a validar. Es muy importante la selección de los sujetos que componen cada grupo: los casos han de reflejar el tipo de pacientes sobre los que se efectuará la prueba a evaluar en situación clínica real y los controles han de incluir a pacientes con entidades que son diagnóstico diferencial de la patología del estudio. La proporción entre casos y controles ha de ser equivalente a la prevalencia de la enfermedad en la población del estudio. Si ello no fuera así, no podremos determinar el comportamiento de la prueba. El parámetro a evaluar es el cociente de posibilidades –*odds ratio* (OR)–.

2. ¿Incluyó la muestra un espectro adecuado de pacientes?

¿Están adecuadamente descritos los pacientes y cómo se seleccionaron?

Al plantear cualquier estudio de validación, inicialmente hemos de definir la población diana, que es aquel grupo de sujetos en los que se sospecha la enfermedad y sobre la que se aplicaría la exploración en situación clínica real. Suele ser una combinación de pacientes con la enfermedad que queremos diagnosticar y de pacientes con cuadros que son diagnóstico diferencial de ésta. Es evidente que cualquier prueba puede diferenciar entre sujetos con un grado avanzado de la enfermedad y sujetos sanos, por lo que la importancia radica en que esta prueba ha de saber discernir a los enfermos, en sus diferentes grados, de entre todos los pacientes habituales sobre los que se realiza la prueba. La muestra es un subgrupo de pacientes de la población diana seleccionados para el estudio y sobre los que se testará la prueba a validar. La selección de los sujetos de la muestra determinará la prevalencia de la enfermedad en el grupo, es decir, la proporción de sujetos de la muestra que padecen la enfermedad. Esta prevalencia es también llamada 'probabilidad pretest' de padecer la enfermedad por tratarse de la probabilidad de que un sujeto de la muestra extraído al azar padezca la enfermedad.

Un sesgo de selección se produce cuando la muestra incluye pacientes inequívocamente enfermos (con estadios de la enfermedad avanzados) y sujetos totalmente sanos (sin ninguna enfermedad dentro del espectro del diagnóstico diferencial de la pa-

tología del estudio). La prueba a validar etiquetará más fácilmente a los sujetos de la muestra aumentando artificialmente la sensibilidad y la especificidad al disminuir los falsos positivos (FP) y negativos (VN) –se acierta más fácilmente–.

3. ¿Existe una adecuada descripción de la prueba?

¿Se define con claridad qué es un resultado positivo y qué es un resultado negativo?
¿Se especifica la reproducibilidad de la prueba (este puede constituir un punto clave en pruebas que dependen del observador, como las técnicas de imagen)?

Se ha de procurar que cualquier dato fruto de un estudio de validación sea lo más transparente y objetivo posible. Cualquier resultado, tanto cualitativo como cuantitativo, no ha de verse influenciado por la subjetividad del observador. Por ello, es imprescindible definir de forma estricta lo que denominaremos resultado positivo y resultado negativo.

Una manera de determinar la consistencia de estos datos es midiendo su reproducibilidad en medidas repetidas, tanto por un mismo observador (intraobservador) como entre diferentes observadores (interobservadores).

La reproducibilidad en variables cuantitativas se mide mediante el coeficiente de correlación intraclase y en variables cualitativas mediante los índices de acuerdo y el índice kappa.

Es del todo incorrecto evaluar la reproducibilidad mediante los coeficientes de correlación y la *t* de Student para muestras apareadas.

Preguntas detalladas (¿vale la pena continuar?)

4. ¿Hubo evaluación 'ciega' de los resultados?

¿Las personas que interpretaron la prueba conocían los resultados del patrón oro (y viceversa)?

Con esta pregunta se pretende detectar un posible sesgo de información: cuando el explorador que evalúa la prueba conoce el resultado de la exploración de referencia o el contexto que rodea a la muestra de pacientes seleccionados. El contexto se refiere a las características del paciente que pueden influir, consciente o inconscientemente, en el diagnóstico. Es evidente que cualquier prueba utilizada habitualmente en un contexto clínico, si se utiliza en la población general detectará peor a los enfermos (bajará su sensibilidad) y etiquetará mejor a los sanos (mejorará su especificidad). Todo estudio de validación ha de detallar los mecanismos que se han empleado para evitar este sesgo.

5. ¿La decisión de realizar el patrón oro fue independiente del resultado de la prueba problema?

Considerar si:
– Se incluyeron preferentemente los resultados positivos en la prueba a evaluar.
– Se utilizaron diferentes patrones oro en los positivos y en los negativos.

Cuando el resultado de la exploración del estudio condiciona la realización de la prueba de referencia, se puede incurrir en un sesgo de verificación. Ello suele suceder cuando la prueba de referencia es molesta o agresiva para el paciente. De esta manera, se etiquetan más fácilmente a sujetos como enfermos y peor a los sanos. La sensibilidad aumenta al disminuir los FN y disminuye la especificidad al aumentar los FP.

Por ejemplo, cuando al validar el eco-Doppler de troncos supraaórticos (TSA) en pacientes de nuestra consulta, sólo realizamos una arteriografía de TSA a pacientes con sospecha de estenosis > 60% por ecografía.

¿Cuáles son los resultados?

*6. ¿Se pueden calcular los cocientes de probabilidad (***likelihood ratios***)?*

¿Se han tenido en cuenta los pacientes con resultado 'no concluyentes'?

¿Se pueden calcular los cocientes de probabilidad para distintos niveles de la prueba, si procede?
Hemos de definir diferentes conceptos, que nos ayudarán a entender los diferentes procesos de validación. Estos conceptos se generan cuando aceptamos que una prueba diagnóstica no siempre acierta y puede tener diferentes errores en sus determinaciones respecto a la realidad:

- *Verdaderos positivos (VP):* pacientes realmente enfermos en los que la prueba ha acertado.
- *Verdaderos negativos (VN):* pacientes sanos en los que la prueba ha acertado.
- *Falsos positivos (FP):* son los etiquetados como enfermos y que realmente están sanos. La probabilidad de cometer este fallo se llama error α.
- *Falsos negativos (FN):* son los etiquetados como sanos y que en realidad están enfermos. La probabilidad de cometer este fallo se llama error β.

Una vez asimilados estos conceptos, que nos han hablado de pacientes en números absolutos, podemos agruparlos y definir nuevos parámetros que nos expresan probabilidades o proporciones:

Sensibilidad. Representa la proporción de pacientes con la enfermedad que son etiquetados como enfermos por la prueba. Dicho de otro modo, representa la proporción de enfermos correctamente identificados por la prueba. El valor será entre 0 (mínimo) y 1 (máximo).

$$Sensibilidad = VP / VP + FN$$
$$Sensibilidad = VP / Total\ enfermos$$

Especificidad. Representa la proporción de pacientes sin la enfermedad que son etiquetados como sanos por la prueba. Dicho de otro modo, la proporción de sanos correctamente identificados por la prueba. El valor será entre 0 (mínimo) y 1 (máximo).

$$Especificidad = VN / VN + FP$$
$$Especificidad = VN / Total\ sanos$$

Valor pronóstico positivo. Es la probabilidad de estar realmente enfermo una vez la prueba ha dado positivo.

$$VPP = VP / VP+FP$$
$$VPP = VP / Total\ positivos$$

Valor pronóstico negativo. Se trata de la probabilidad de estar realmente sano una vez la prueba ha dado negativo.

$$VPN= VN / VN + FN$$

Valor global. Proporción total de sujetos etiquetados correctamente por la prueba:

$$Valor\ global = VN + VP / Total\ de\ sujetos$$

Razón de verosimilitud (likelihood ratio) *o cociente de probabilidad.* Ya hemos relacionado la probabilidad pretest con la prevalencia de la enfermedad en la población diana/muestra. Los cocientes de probabilidad determinarán cuánto aumenta o disminuye esta probabilidad pretest según el resultado de la prueba, ofreciéndonos el valor de la probabilidad postest, que es la probabilidad de padecer o no la enfermedad estudiada según el resultado de la prueba, es decir, nos ofrece los VPP y VPN. Si solamente supiéramos los valores de sensibilidad y especificidad, podríamos conocer el VPP en diferentes situaciones clínicas, dependiendo de la probabilidad pretest de cada población/muestra.

Cociente de probabilidad de un resultado positivo (LR+). Se calcula dividiendo la sensibilidad por el complemento de la especificidad (1 – especificidad), siendo este último concepto la probabilidad de los sanos de dar positivo en el test. El CP+ se refiere al número de veces que es más probable un resultado positivo en enfermos que en sanos. Si conocemos la probabilidad pretest y el CP+ podremos calcular el VPP. Para poder hacerlo, hemos de trabajar con razones de OR. Antes de seguir, explica-

remos cómo se transforman probabilidades (P) en OR y viceversa:

$$OR\ de\ (P) = P / (1 - P)$$
$$P = OR / (OR + 1).$$

La fórmula que aplicaremos es:

$$OR\ del\ VPP = (OR\ de\ la\ probabilidad\ pretest) \times (CP+).$$

Aplicando las transformaciones de probabilidad a OR:

$$VPP / (1 - VPP) = [P / (1 - P)] \times [S / (1 - E)].$$

Una vez obtenida la OR del VPP, lo transformaremos en probabilidad de la forma ya indicada.

Cociente de probabilidad de un resultado negativo. Se calcula dividiendo el complemento de la sensibilidad (1 – sensibilidad) por la especificidad. Al contrario que la razón para resultados positivos, mejor cuanto menor sea el resultado.

Para poder comparar sus resultados con la razón para resultados positivos, podemos definir la razón para resultados negativos como su inverso = especificidad / (1 – sensibilidad). De esta forma, ambas razones de verosimilitud, positiva y negativa, seguirán la misma escala de puntuación, indicándonos cómo influye el resultado de la prueba descartando el diagnóstico. La utilidad de la razón de verosimilitud de un resultado negativo (expresada como E / 1 – S) es que permite determinar el VPN, también en forma de OR, con la siguiente fórmula:

$$VPN / (1 - VPN) = [(1 - P) / P] \times [E / (1 - S)].$$

La interpretación global de los valores absolutos de las razones de verosimilitud es:

> 10. *Excelente:* es un resultado prácticamente concluyente y acaso influirá de una manera decisiva en la probabilidad posprueba.

5-10. *Buena:* este resultado provocará un cambio moderado desde la probabilidad preprueba a la posprueba.

2-5. *Regular:* el cambio provocado será pequeño.

1-2. *Pobre:* prácticamente no influirá en la variación de la probabilidad pretest, siendo el valor 1 el que determinará que la prueba no aporta información alguna.

7. ¿Cuán precisos son los resultados?

Busca o calcula los intervalos de confianza de los cocientes de probabilidad.

El cálculo de estos valores siempre ha de ir acompañado de una medida de dispersión. La más adecuada, por ser la más inteligible, es su intervalo de confianza (IC). El más utilizado es el IC del 95%, que determina que existe un 95% de posibilidades que los resultados reales hallados sobre la población diana se encuentren dentro de este IC. Como es bien sabido, el IC viene determinado por la dispersión de los valores obtenidos y por el número de pacientes incluidos: a valores más homogéneos y poco dispersos, el IC será más estrecho; y a mayor número de pacientes, el IC también será más pequeño.

¿Son los resultados aplicables en tu medio?

8. ¿Serán satisfactorios en tu medio o población local la reproducibilidad de la prueba y su interpretación?

Considera si tu medio parece ser muy diferente al del estudio.

Hemos de determinar si las condiciones que se presentan en el estudio de validación son extrapolables a nuestro entorno, tanto la población, la tecnología aplicada, la destreza de los exploradores, los condicionantes económicos, etc.

Un estudio de validación con unos muy buenos resultados a cargo de un equipo de investigación no significa que automáticamente todo el mundo pueda asumir dichos resultados.

9. ¿Es aceptable la prueba en tu medio?

Considera la disponibilidad de la prueba, los riesgos/molestias de la prueba y los costes.

Recibimos mucha información de publicaciones de equipos que trabajan en países con un funcionamiento del sistema sanitario sustancialmente diferente al nuestro, tanto en la financiación como en las indicaciones o la cultura de la propia población. Ello puede determinar que muchas de las exploraciones no sean asumibles en nuestro medio o, por contrario, podamos disponer de técnicas mejores que no están al alcance de los autores.

Todo ello condicionará la importancia relativa que le hemos de otorgar a dicho estudio, básicamente en su aplicabilidad real.

10. ¿Modificarán los resultados de la prueba la decisión sobre cómo actuar?

Desde la perspectiva de la práctica, si la actitud no va a cambiar, la prueba es (como mínimo) inútil. Considera el umbral de acción y la probabilidad de enfermedad antes y después de la prueba.

Llegados a este punto, y tras haber evaluado cada uno de los aspectos explicados, el lector ha de tener claro si la información obtenida puede modificar nuestra práctica habitual. Podemos decidir que la información no nos va a hacer cambiar la actitud en caso de resultados negativos, resultados positivos pero sin significación estadística (IC no concluyente) o en caso de que la metodología utilizada no nos haya dado garantías.

Bibliografía

1. 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: 41-4.
2. Delgado M, Llorca J, Doménech JM. Estudios para pruebas diagnósticas y factores pronósticos. Barcelona: Signo; 2005.
3. Fernández E, García AM. Búsqueda y lectura crítica de artículos científicos. Barcelona: Signo; 2006.
4. Greenhalgh T. How to read a paper. Papers that report diagnostic or screening tests. BMJ 1997; 315: 540-3.
5. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. JAMA 1994; 271: 703-7.
6. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Epidemiología clínica. Ciencia básica para la medicina clínica. México: Médica Panamericana; 1998.

THE VALIDATION OF DIAGNOSTIC TESTS

Summary. *Diagnosis is a fundamental part of our daily practice. We are confronted by a huge number of studies on diagnostic tests and it is essential to be able to distinguish between those that can be of use to us and those that only produce 'background noise'. The aim of this study is to offer a series of tools and guidelines that may help us to read these reports with a critical attitude, so that we are in a position to choose the ones that may be of most use to us. [ANGIOLOGÍA 2007; 59: 433-8]*

Key words. *Diagnosis. Validation studies.*