REVISIÓN METODOLÓGICA

Análisis de la calidad de los estudios de evaluación de pruebas diagnósticas

J. ZAMORA y V. ABRAIRA

Unidad de Bioestadística Clínica. Hospital Ramón y Cajal. Madrid. CIBER Epidemiología y Salud Pública (CIBERESP)

RESUMEN

A pesar de la importancia del diagnóstico en el desarrollo de una buena práctica clínica, la investigación que evalúa el valor de las pruebas diagnósticas es de baja calidad metodológica. Por otro lado, los artículos que describen esta investigación presentan importantes defectos y carencias que hacen difícil la tarea de evaluar críticamente las evidencias disponibles sobre el valor de una prueba. En este artículo se recuerdan los elementos del diseño más importantes de estos estudios de investigación y se describe una herramienta validada para la evaluación de la calidad metodológica de la investigación sobre diagnóstico.

Palabras clave: diseño, validez, sesgos, pruebas diagnósticas.

Introducción

El proceso diagnóstico es una actividad intelectualmente compleja mediante la que se pretende clasificar a los pacientes de acuerdo a la presencia o no de una condición clínica, o se pretende monitorizar la evolución de una patología o graduar la gravedad de la misma. El fin último no es tanto la determinación de si el paciente presenta una determinada enfermedad o no, sino que con el diagnóstico se inicia un proceso de toma de decisiones relacionadas con la obtención de un pronóstico y con la elección de un tratamiento. En muchas ocasiones es la primera intervención clínica sobre el paciente y es de absoluta importancia para el desarrollo de una práctica clínica adecuada. Un error en el diagnóstico puede seguirse del establecimiento de un pronóstico erróneo y de la aplicación de un tratamiento inadecuado con el consiguiente riesgo de provocar un daño a los pacientes.

La obtención de un diagnóstico supone la integración de resultados de numerosas fuentes de información. Típicamente se establece a partir de una combinación de información proveniente de la historia clínica y de la exploración del paciente (por ejemplo a partir de factores de riesgo y la presencia de signos y síntomas) y a partir de los resultados de diversas pruebas diagnósticas (hematológicas, bioquímicas, radiológicas, microbiológicas, histopatológicas, etc.).

Correspondencia: Javier Zamora
Unidad de Bioestadística Clínica
Hospital Ramón y Cajal. Ctra. Colmenar, km. 9,100
28034 Madrid
e-mail: javier.zamora@hrc.es

Estudios de validez de pruebas diagnósticas

Ante la disponibilidad de una nueva prueba diagnóstica se debe proceder al estudio de su validez antes de introducirla en la práctica clínica. Los estudios de validez de pruebas diagnósticas pueden diseñarse con dos objetivos. El primero y más habitual es el de determinar la validez diagnóstica de la prueba que se evalúa. Este estudio supone la comparación de las clasificaciones resultantes de dos procesos independientes de medición. El primero, una prueba de referencia (patrón oro o *gold-standard)* que se asume clasifica de forma válida a los sujetos respecto a la presencia o ausencia de una condición clínica. El segundo proceso es la prueba diagnóstica que se somete a evaluación. El resultado de la comparación se suele expresar con los índices sensibilidad y especificidad, o combinaciones de estos índices tales como los cocientes de probabilidad, o los valores predictivos[1]. Para aquellas pruebas en las que el resultado puede ser utilizado con distintos puntos de corte, el resultado se expresa como una curva ROC que refleja la sensibilidad y especificidad para distintos puntos de corte[2]. El diseño más adecuado para este objetivo de evaluación se corresponde con un estudio observacional, transversal, en el que a una serie consecutiva y representativa de pacientes con sospecha de la enfermedad se les realiza, de forma ciega e independiente, la prueba a evaluar y la prueba de referencia[3], las cuáles son

interpretadas en ausencia de toda información clínica adicional que no estará disponible cuando la prueba se utilice en la práctica. La mayoría de los estudios de evaluación de pruebas diagnósticas en la literatura evalúan este aspecto de la validez diagnóstica aunque pocos de ellos se adhieren a este diseño óptimo.

El segundo objetivo de la investigación sobre pruebas diagnósticas va un paso más allá de conocer el rendimiento diagnóstico de la prueba. Se pretende evaluar el impacto que tendría el uso de una u otra prueba o estrategia diagnóstica sobre el manejo clínico, las decisiones terapéuticas y, en último término, sobre los resultados de salud de los pacientes[4]. Los estudios que afrontan este objetivo, todavía poco numerosos en la literatura, son de diseño complejo y en ocasiones se desarrollan como ensayos clínicos aleatorizados.

En el resto de este artículo, nos referiremos a los estudios de evaluación de la validez de las pruebas diagnósticas.

Análisis de la calidad de los estudios

Existen varias dimensiones sobre las que se puede evaluar la calidad de un estudio. Primeramente, atendiendo a la presencia o no de sesgos en la estimación de la validez de la prueba diagnóstica. Estos sesgos pueden ser debidos a deficiencias tanto del diseño del estudio como de su desarrollo y ejecución. Entre estos elementos, por ejemplo, se encuentran la forma en que se seleccionaron los pacientes, cómo y cuando se aplicaron las pruebas, cómo se interpretaron y cómo fue el flujo de pacientes a lo largo del estudio. Todos estos aspectos se vinculan a la validez interna del estudio. En segundo lugar, se puede evaluar la capacidad de generalización-validez externa, de los resultados del estudio a otros pacientes y entornos, la cual dependerá, entre otros aspectos, de cuál fue el espectro de la enfermedad de los sujetos incluidos en el estudio, del ámbito en el que se desarrolló y de la reproducibilidad, el umbral utilizado y la calibración de la prueba evaluada. Finalmente existe otra dimensión de la calidad que se refiere a la forma en que se escribe el artículo que describe el diseño estudio y sus resultados. La iniciativa STARD para la publicación de estudios de validez diagnóstica está dirigida a editores de revistas y a los autores de artículos y pretende mejorar la calidad de los artículos para permitir a los lectores evaluar los potenciales sesgos del estudio y juzgar sobre su generabilidad[5]. Esta iniciativa es en el área de investigación sobre diagnóstico lo que CONSORT es en el área de los ensayos clínicos[6].

Se han publicado varias revisiones metodológicas que evalúan la calidad de los estudios de diagnóstico, y el panorama es un tanto desalentador[7,8]. En el año 1995, se encontró que de siete criterios de calidad contemplados, sólo uno de ellos (sesgo de verificación) fue correctamente evitado tan sólo en la mitad de los 112 estudios incluidos en la revisión[7]. En otra revisión más reciente del año 2006, de los casi 500 estudios incluidos, sólo uno no presentaba ningún defecto metodológico en su diseño[8]. Esta generalizada baja calidad metodológica y la pobre presentación de los resultados hacen que, con frecuencia, las revisiones sistemáticas sean incapaces de alcanzar conclusiones sobre la validez diagnóstica de una prueba[9,10].

Herramientas para evaluar la calidad de un estudio de evaluación de pruebas diagnósticas

Existen numerosas herramientas de evaluación de la calidad de los estudios de diagnóstico. En un reciente estudio, Whiting y cols., identificaron hasta un total de 91 instrumentos relacionados con la calidad de la investigación en diagnóstico[11]. Entre ellos se encuentran algunas de las escalas y listas de ítems más populares, como la presentada en la serie de guías de usuario de la literatura médica del *JAMA*[12,13] y la guía equivalente del *BMJ*[14]. Este mismo grupo de investigadores ha diseñado otro instrumento, denominado QUADAS (de las siglas en inglés Quality Assessment of Diagnostic Accuracy Studies) que inicialmente fue ideado para la evaluación de la calidad de los estudios primarios incluidos en revisiones sistemáticas[15]. En el desarrollo de esta herramienta, un grupo de expertos utilizó una metodología Delphi para seleccionar y depurar una lista de preguntas relevantes a la hora de determinar la presencia de sesgos metodológicos en el desarrollo e interpretación del estudio. Los 14 ítems seleccionados cubren aspectos relativos al espectro de pacientes utilizados en el estudio, del patrón oro seleccionado, y varios sesgos como el sesgo de progresión de la enfermedad y la paradoja del tratamiento, los sesgos de verificación parcial y de verificación diferencial, el sesgo de incorporación, la calidad de la descripción de las pruebas diagnósticas utilizadas, y el tratamiento dado a las retiradas del estudio y a los resultados indeterminados. El cuestionario incluye estas 14 preguntas de respuesta sí o no (o en su caso no se sabe) y en una segunda versión se contempla la posibilidad de considerar la pregunta «no aplicable» en situaciones concretas. En conjunto se trata de una herramienta de gran utilidad y de extendido uso, sobre todo en la realización de revisiones sistemáticas de pruebas diagnósticas. A pesar de su popularidad, el cuestionario QUADAS no está exento de ciertas críticas. Se ha cuestionado fundamentalmente su reproducibilidad[16,17], sobre todo en los ítems relativos a la presencia de resultados indeterminados o no concluyentes y a la información sobre las pérdidas y retiradas del estudio.

A continuación se muestran los ítems incluidos en el cuestionario QUADAS con una breve explicación del significado de cada uno de ellos[15].

1. ¿Fue el espectro de pacientes representativo de los pacientes que recibirán la prueba en la práctica?

Las variaciones en las características clínicas y demográficas de los sujetos reclutados en el estudio (espectro de la enfermedad) son una contrastada fuente de variación en los estudios de evaluación de pruebas diagnósticas. Por ejemplo, un diseño de casos y controles en

el que se reclutan casos con la enfermedad y controles sanos produce una estimación optimista de la capacidad diagnóstica de la prueba.

2. ¿Se describieron con claridad los criterios de selección?

Este es un aspecto relacionado con la calidad de la escritura del artículo. Si podemos evaluar los criterios de inclusión y exclusión del estudio podremos valorar la aplicabilidad de los resultados a los pacientes de nuestro ámbito en concreto.

3. ¿Es previsible que el patrón de referencia escogido clasifique correctamente el problema a estudio?

Las estimaciones de la validez diagnóstica calculadas en el estudio se basan en la asunción de que el patrón de referencia clasifica correctamente. De esta forma, las discordancias entre la prueba evaluada y la de referencia se achacan a errores de la prueba evaluada.

4. El período transcurrido entre la aplicación de la prueba a estudio y la prueba de referencia, ¿es lo suficientemente corto como para que sea razonable asumir que el problema a estudio no ha evolucionado en ese periodo?

El diseño ideal de estudio de evaluación diagnóstica es un estudio transversal en el que ambas pruebas (evaluada y de referencia) se realizan simultáneamente. Cualquier lapso temporal entre la realización de ambas pruebas puede dar cabida a la evolución de la enfermedad, en uno u otro sentido, con el consiguiente efecto en la estimación del rendimiento diagnóstico.

5. ¿Se verificó el diagnóstico usando una prueba de referencia en toda la muestra del estudio o en una submuestra aleatoria de la misma?

Si la aplicación del estándar de referencia no es independiente del resultado de la prueba a evaluar, pueden aparecer varios mecanismos de sesgos. Un caso típico ocurre cuando a los casos negativos de la prueba diagnóstica no se les somete a la prueba de referencia por ser cruenta (por ejemplo una biopsia). Este modo de operar introduce un sesgo que recibe el nombre de sesgo de verificación parcial o sesgo de referencia.

6. ¿Se aplicó en los pacientes la misma prueba de referencia independientemente del resultado obtenido en la prueba evaluada?

En ocasiones, los investigadores utilizan diferentes pruebas de referencia según el resultado de la prueba evaluada. Un caso habitual es el utilizar el seguimiento clínico como prueba de referencia en aquellos casos que resultaron negativos en la prueba. Este proceder, a veces inevitable, introduce un sesgo denominado de verificación diferencial que suele sobrestimar las propiedades de la prueba evaluada.

7. ¿Eran la prueba de referencia y la prueba a estudio independientes entre sí? (ningún elemento de la prueba a estudio formaba parte de la prueba de referencia).

El sesgo conocido como de incorporación sucede cuando el patrón de referencia incluye, total o parcialmente, los resultados de la prueba evaluada. Este sesgo es distinto de la mera ausencia de enmascaramiento a la hora de interpretar el resultado de la prueba de referencia. Para que exista el sesgo de incorporación hace falta que el resultado de la prueba evaluada forme parte de los criterios de definición del diagnóstico final. Este ítem es sólo aplicable cuando la prueba de referencia es una combinación de otras pruebas.

8. La descripción de la utilización de la prueba evaluada ¿es suficiente para permitir su replicación?

La presencia de esta descripción, y la de la pregunta siguiente, permitirán identificar posibles factores que expliquen variaciones en la estimación de la validez de una prueba diagnóstica debidas a diferencias en la ejecución de las pruebas. Por otro lado, una adecuada descripción es imprescindible también para juzgar la aplicabilidad de las pruebas en nuestro entorno.

9. La descripción de la utilización de la prueba de referencia ¿es suficiente para permitir su replicación?

Ver comentarios a la pregunta anterior.

10. ¿Se interpretó la prueba evaluada sin conocer los resultados de la prueba de referencia?

Esta y la siguiente pregunta se refieren al enmascaramiento en la interpretación de las pruebas. Es razonable pensar que la interpretación de una prueba, tanto más cuanto más subjetiva sea, puede verse influenciada por el conocimiento del resultado de la otra prueba. La ausencia de esta interpretación ciega de las pruebas produce una sobrestimación del rendimiento diagnóstico. En ocasiones puede que esta pregunta (o la siguiente) no sean aplicables, dependiendo del orden de ejecución de las pruebas y dependiendo también de si la prueba es una prueba totalmente objetiva como las pruebas de laboratorio.

11. ¿Se interpretó la prueba de referencia sin conocer los resultados de la prueba evaluada?

Ver comentarios a la pregunta anterior.

12. La información clínica disponible en la interpretación de los resultados de las pruebas ¿es la misma que estará disponible cuando se use la prueba en la práctica?

De forma similar a lo expuesto en la pregunta 10, cualquier información clínica o demográfica que sea de utilidad en la interpretación de los resultados de las pruebas diagnósticas puede influir en sus resultados. Como norma general, para la interpretación de las pruebas sólo debería emplearse la información que estará disponible cuando se aplique la prueba en cuestión en la práctica. Al igual que en la pregunta anterior, en el caso de pruebas totalmente objetivas esta pregunta puede considerarse no aplicable.

13. ¿Se informó de los resultados no interpretables o no concluyentes?

Si los casos en los que falta información de cualquiera de las pruebas del estudio son excluidos del análisis,

los resultados podrían estar sesgados. La dirección y la magnitud del sesgo dependerán de si la causa de estos resultados no interpretables está asociada al estado verdadero del paciente o al resultado de la prueba y, por supuesto, de la frecuencia de los mismos.

14. ¿Se explican las pérdidas y retiradas del estudio?

La justificación de esta pregunta es compartida con la pregunta anterior. Este caso puede considerarse un caso particular de pérdidas de resultados.

Evidencia empírica del sesgo introducido por defectos metodológicos

En el caso de los ensayos clínicos que analizan el efecto de intervenciones terapéuticas, se ha demostrado una sobrestimación del efecto debida a la presencia de defectos metodológicos[18,19]. De forma similar, en el área de investigación sobre diagnóstico, varios autores han demostrado empíricamente la existencia de una asociación entre la calidad metodológica y la estimación del rendimiento diagnóstico. Lijmer analizó 218 evaluaciones de pruebas diagnósticas contenidas en 11 metanálisis y estudió el efecto que sobre estas evaluaciones tenían 6 criterios metodológicos y 3 características de la redacción de los artículos[20]. Entre sus hallazgos más importantes se encuentra el que los estudios que reclutan pacientes con diseños tipo casos enfermos y controles sanos (y por tanto que incluyen muestras no representativas de la población en la práctica) y los estudios en los que se usan diferentes estándares de referencia, se sobreestiman las capacidades diagnósticas de las pruebas que evalúan. Posteriormente Rutjes y cols., reprodujeron el estudio ampliando el análisis hasta 15 elementos metodológicos de calidad y de la redacción del artículo y recogiendo datos de 31 meta-análisis que incluían 487 estudios primarios[8]. Los resultados fueron consistentes con el anterior estudio. Como hallazgos adicionales, los autores encontraron que la inclusión no consecutiva de pacientes y la extracción retrospectiva de datos sobreestiman ligeramente el valor diagnóstico de la prueba mientras que la selección de pacientes por remisión a la prueba en evaluación en lugar de por la presencia de signos y síntomas, lo infraestima.

La calidad de los ensayos clínicos, tanto en su diseño y ejecución como en la comunicación de sus resultados, se ha visto incrementada notablemente en los últimos años. Confiemos que con la aparición de iniciativas como la iniciativa STARD y con la extensión del uso de herramientas como el QUADAS, la evolución de la calidad de los estudios de diagnóstico siga una trayectoria similar.

Bibliografía

1. Abraira V. Índices de rendimiento de las pruebas diagnósticas. *SEMERGEN* 2008; 28: 193-4.
2. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; 143: 29-36.
3. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. *J Clin Epidemiol* 2003; 56: 1118-28.
4. Sackett DL, Haynes RB. The architecture of diagnostic research. *BMJ* 2002; 324: 539-41.
5. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM y cols. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD Initiative. *Ann Intern Med* 2003; 138: 40-4.
6. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *Ann Intern Med* 2001; 134: 657-62.
7. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. *JAMA* 1995; 274: 645-51.
8. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, Van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ* 2006; 174: 469-76.
9. Mijnhout GS, Hoekstra OS, Van Tulder MW, Teule GJ, Deville WL. Systematic review of the diagnostic accuracy of (18)F-fluorodeoxyglucose positron emission tomography in melanoma patients. *Cancer* 2001; 91: 1530-42.
10. Westwood ME, Whiting PF, Kleijnen J. How does study quality affect the results of a diagnostic meta-analysis? *BMC Med Res Methodol* 2005; 5: 20.
11. Whiting P, Rutjes AW, Dinnes J, Reitsma JB, Bossuyt PM, Kleijnen J. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. *J Clin Epidemiol* 2005; 58: 1-12.
12. 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.
13. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. *JAMA* 1994; 271: 389-91.
14. Greenhalgh T. How to read a paper. Papers that report diagnostic or *screening* tests. *BMJ* 1997; 315: 540-3.
15. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol* 2003; 3: 25.
16. Hollingworth W, Medina LS, Lenkinski RE, Shibata DK, Bernal B, Zurakowski D y cols. Interrater reliability in assessing quality of diagnostic accuracy studies using the QUADAS tool. A preliminary assessment. *Acad Radiol* 2006; 13: 803-10.
17. Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PN, Kleijnen J. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. *BMC Med Res Methodol* 2006; 6: 9.
18. Juni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. *BMJ* 2001; 323: 42-6.
19. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995; 273: 408-12.
20. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, Van der Meulen JH y cols. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999; 282: 1061-6.