Sesgos en los estudios sobre pruebas diagnósticas

V. Abraira

Unidad de Bioestadística Clínica. Hospital Ramón y Cajal. Madrid.
Nodo de la red R_MBE (G03/90)

Aunque el diagnóstico desempeña un papel central en la actividad clínica, los médicos reciben escaso entrenamiento formal en la utilización de pruebas diagnósticas durante su formación académica. Además la calidad de la investigación publicada sobre evaluación de pruebas diagnósticas es pobre. En esta nota se repasan los aspectos de diseño de esos estudios y la cuantificación empírica del impacto de los sesgos relacionados con ellos.

Palabras clave: sesgos, diseño, pruebas diagnósticas.

Although the diagnosis plays a central role in the clinical practice, physicians receive scarce formal training in the use of diagnostic tests during their academic education. Furthermore, the quality of the investigation published on diagnostic test evaluation is poor. In this note, the aspects of these studies' design and the empiric quantification of the impact of the biases related with them are reviewed.

Key words: bias, design, diagnostic tests.

El diagnóstico desempeña un papel central en la actividad clínica: es la primera intervención clínica sobre el paciente y su resultado condiciona el desarrollo de la práctica clínica posterior, sin olvidar el impacto social que puede tener por el efecto de asignar a los ciudadanos etiquetas de falta de normalidad[1]. Tanto es así, que la habilidad para realizar un diagnóstico es una de las cualidades más valoradas por los propios clínicos, e incluso, a veces, se señala el diagnóstico como la actividad más característica de los médicos, actividad que ningún otro profesional puede hacer[2]. Si bien ello contrasta con el escaso entrenamiento formal en la utilización de pruebas diagnósticas que reciben los médicos durante su formación académica[3].

Es además una actividad difícil. Un síntoma de la dificultad conceptual del diagnóstico es que la calidad de la investigación publicada sobre diagnóstico es, por decirlo de forma suave, manifiestamente mejorable[4], como la creciente realización de revisiones sistemáticas sobre pruebas diagnósticas sigue poniendo de manifiesto. Existe una gran variabilidad en el diseño de los estudios sobre evaluación de pruebas diagnósticas y muchos de ellos incluyen sesgos potenciales que afectan a su validez y utilidad y hay también falta de uniformidad en la propia descripción del proceso y sus resultados. De hecho, no es infrecuente que una revisión sistemática no pueda responder a una pregunta concreta sobre diagnóstico, no por falta de artículos que evalúen la prueba, sino precisamente por la baja calidad del diseño y de la presentación de los resultados de los artículos encontrados[5].

Como consecuencia de todo ello (dificultad conceptual, escaso entrenamiento durante la formación y variabilidad y pobre calidad de las publicaciones), los clínicos tienen grandes dificultades para la interpretación de los artículos sobre evaluación de pruebas diagnósticas y, lo que es más relevante, para la aplicación de sus resultados para la elección de las pruebas a realizar a sus pacientes[6,7].

En una nota previa[8] se vieron los índices usados para presentar los resultados de la evaluación de la validez de las pruebas diagnósticas. En esta nota, que la complementa, se repasan los aspectos de diseño de esos estudios y la cuantificación empírica del impacto de los sesgos relacionados con ellos, intentado contribuir a una mayor difusión de las claves que facilitan la interpretación de dichos estudios. El diseño óptimo consiste en seleccionar un grupo de pacientes representativos de aquéllos en los que se pretende usar la prueba y aplicarles a todos ellos, simultáneamente, la prueba en evaluación y otra prueba de referencia, aceptada como patrón para hacer el diagnóstico

Correspondencia: V. Abraira.
Unidad de Bioestadística Clínica. Hospital Ramón y Cajal.
Ctra. Colmenar, km. 9,100.
28034 Madrid.
Correo electrónico: victor.abraira@hrc.es

correcto. Por ejemplo, en la nota sobre los índices[8] se comentó la evaluación de la concentración plasmática de péptido natriurético tipo B para diagnosticar, en ancianos, la disfunción ventricular izquierda; en el artículo comentado se usó el ecocardiograma como prueba de referencia. Ambas pruebas deben interpretarse de modo enmascarado, es decir, cada una se debe interpretar sin que el investigador que lo haga sepa el resultado de la otra. De modo similar a lo que ocurre con el tratamiento[9], este ideal de diseño está en contradicción con la buena práctica clínica, en la que rara vez se solicitan las pruebas simultáneamente, más bien al contrario, las pruebas deberían solicitarse de modo secuencial y cada una de ellas solicitarse e interpretarse en función de toda la información disponible en cada momento, incluyendo los resultados de las pruebas previas. Seguramente de esta contradicción surgen los defectos observados en la literatura. Los que más impacto tienen sobre la estimación de la validez de la prueba tienen que ver con la selección de los pacientes, la falta de independencia en la comparación con el patrón de referencia y la falta de enmascaramiento en la interpretación de las pruebas.

SELECCIÓN DE LOS PACIENTES

Dado que en la práctica clínica los problemas de diagnóstico se plantean entre enfermedades o estados de salud que comparten síntomas, una prueba diagnóstica es verdaderamente útil si permite distinguir entre trastornos que de otra forma podrían confundirse, por tanto la validez de una prueba debe establecerse en ese escenario, es decir, en un estudio que incluya un espectro de pacientes lo más parecido posible al del medio en el que la prueba se pretenda usar en el futuro, típicamente una muestra consecutiva de pacientes. Sin embargo, una tentación muy extendida en estos estudios es el diseño caso-control, en el que se seleccionan dos muestras, una de pacientes que se sabe que tienen la enfermedad y otra de individuos que no la tienen. Se ha demostrado que este diseño introduce la mayor sobreestimación del rendimiento de la prueba. Usando como índice de validez la *odds ratio* diagnóstica, que es un modo de sintetizar en un solo índice la sensibilidad y la especificidad, este diseño lo sobreestima[10] multiplicándolo por un factor de 3.

INDEPENDENCIA ENTRE LA PRUEBA Y EL PATRÓN DE REFERENCIA

Muy frecuentemente las pruebas usadas como referencia o patrón de oro son invasivas; ése es justamente uno de los motivos para desarrollar nuevas pruebas, disponer de pruebas menos agresivas, o más baratas, o más fáciles que los patrones de oro. En consecuencia, suele haber problemas para realizar estas pruebas a individuos no enfermos. Por ejemplo, para evaluar la validez de la mamografía en el diagnóstico del cáncer de mama, una buena prueba de referencia es la biopsia, de hecho es la que se suele usar, aunque obviamente hay problemas, tanto éticos como de factibilidad, para realizar biopsias a mujeres con mamografías negativas. Como consecuencia, en muchos de estos estudios, en particular en todos los que se realizan en condiciones reales de asistencia, la prueba de referencia se realiza a la mayor parte de los pacientes con resultado positivo de la prueba y sólo a una pequeña parte de los que tienen resultado negativo, dando lugar al denominado sesgo de referencia o de verificación parcial. Otros autores resuelven el problema aplicando a los pacientes con resultado negativo en la prueba en evaluación otro patrón de referencia diferente, por ejemplo, seguimiento en el tiempo. Ambas soluciones, si bien frecuentemente son las únicas disponibles, incumplen la asunción de independencia entre pruebas y darían lugar a una sobreestimación del rendimiento diagnóstico. Lijmer et al[10] encuentran que cuando se usan diferentes patrones de referencia, el índice de rendimiento global se sobreestima multiplicándose por dos, aunque, sorprendentemente, no encuentran sobreestimación producida por el sesgo de verificación parcial.

ENMASCARAMIENTO EN LA INTERPRETACIÓN DE LAS PRUEBAS

Siguiendo con el ejemplo de la mamografía, parece claro que una imagen dudosa será interpretada de modo distinto, seguramente mejor, si se conoce el resultado de la biopsia. Por ello, para evaluar la validez de la mamografía, ambas pruebas deben interpretarse sin que se conozca el resultado de la otra. Esta exigencia es tanto más importante cuanto mayor componente de interpretación subjetiva tengan las pruebas en cuestión. Lijmer et al[10] encuentran que efectivamente la falta de enmascaramiento sobreestima el índice de rendimiento global en un 30%.

Aunque en los libros de texto también se señalan otros posibles sesgos en el diseño de los estudios de evaluación de pruebas diagnósticas, tales como el diseño retrospectivo, o la inclusión de pacientes de forma no consecutiva, el estudio de Lijmer no encuentra diferencias en la estimación de los índices entre los artículos que los tienen y los que no (una vez corregidos por los efectos ya comentados de los otros sesgos). Una explicación posible para estos hallazgos sorprendentes es que "los defectos nunca vienen solos" y, en el análisis, unos defectos están acaparando el efecto de otros relacionados, por ejemplo, los diseños retrospectivos suelen ser caso-control y éstos no incluyen pacientes de forma consecutiva. También se podría tratar de alguna limitación de la propia investigación de Lijmer, que sería conveniente replicar, aunque no se puede descartar que haya un exceso de celo en la lógica metodológica en los libros de texto.

BIBLIOGRAFÍA

1. Pérez Fernández M, Gervas J. El efecto cascada: implicaciones clínicas, epidemiológicas y éticas. Med Clin (Barc). 2002;118:65-7.
2. The Editors. Diagnosis, diagnosis, diagnosis. BMJ. 2002; 324:0-doi:10.1136/bmj.324.7336.0/g.
3. Latour J. El diagnóstico. Quaderns de salut pública i administració de serveis de salut, 21. Valencia: Escola Valenciana d'Estudis per a la Salut; 2003.

4. 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.
5. 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.
6. Reid MC, Lane DA, Feinstein AR. Academic calculations versus clinical judgments: practicing physicians' use of quantitative measures of test accuracy. Am J Med. 1998;104:374-80.
7. Zamora J, Urrueta I, Pijoán JI, et al. Variabilidad en la interpretación de los índices de validez de las pruebas diagnósticas. XXIII Reunión de la Sociedad Española de Epidemiología: Las Palmas de Gran Canaria; 2005.
8. Abraira V. Índices de rendimiento de las pruebas diagnósticas. SEMERGEN. 2002;28:193-4.
9. Abraira V. ¿Qué es el análisis por intención de tratar? SEMERGEN. 2000;26:393-4.
10. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999;282:1061-6.