MEDICINA BASADA EN EVIDENCIA

Revisiones sistemáticas de estudios de tests diagnósticos

Gladys Moreno G[1,2] , Tomás Pantoja C[1,2].

Systematic reviews of studies of diagnostic test accuracy

El proceso diagnóstico cumple un rol fundamental en la práctica clínica de cualquier médico[1]. La información diagnóstica necesaria para dicho proceso proviene de múltiples fuentes, incluyendo los síntomas y signos obtenidos de la historia clínica, los exámenes bioquímicos y de imágenes, la anatomía patológica y los test psicológicos, entre otros. Es esencial que dicha información sea precisa, dado que en base a ella los clínicos tomarán decisiones que finalmente afectarán los resultados en salud de los pacientes. Los estudios que evalúan la precisión de los test diagnósticos entregan esta información, pero deben ser analizados críticamente utilizando una serie de criterios, que han sido discutidos anteriormente en esta misma sección[2].

Sin embargo, muchas veces dichos estudios se publican en revistas altamente especializadas, cuyo acceso al clínico puede ser limitado. Además, y de manera similar a lo que sucede con otro tipo de estudios, la posibilidad de sesgo y la imprecisión de la información pueden ser mayores cuando sólo es aportada por un estudio único que cuando procede de un grupo de estudios que han abordado la misma pregunta y han sintetizado los resultados de distintos estudios. Por ello, resulta útil contar con revisiones sistemáticas (RS), que permitan informar acerca de las propiedades diagnósticas de diferentes tests utilizados en la práctica clínica habitual.

La racionalidad detrás de la realización de RS en estudios de tests diagnósticos es la misma que aquella detrás de su uso en el ámbito terapéutico: obtener estimaciones de la precisión de los tests basadas en toda la evidencia disponible, evaluar la calidad de dicha evidencia (estudios primarios) e intentar dar cuenta de la variabilidad de los resultados entre diferentes estudios[3]. Las revisiones de estudios acerca de la precisión diagnóstica incluyen las mismas etapas de definición de la pregunta, búsqueda en la literatura, evaluación de la elegibilidad y calidad de los estudios y la extracción y síntesis de los datos[4]. Sin embargo, estas RS presentan otros desafíos en cada una de las etapas descritas anteriormente.

Dado que la literatura acerca de los tests diagnósticos se encuentra fragmentada en diferentes tipos de publicaciones y es difícil de localizar, los algoritmos de búsqueda requerirán mayor desarrollo que en el caso de los estudios randomizados (ER) de intervenciones terapéuticas. Por otro lado, los estudios observacionales –que son los utilizados para estudiar este tipo de preguntas– están más expuestos a diferentes tipos de sesgo que los ER, por lo cual en su evaluación crítica deberán incluirse criterios que den cuenta de esta situación. Adicionalmente, la existencia de más de una medida de precisión en cada estudio (por ejemplo, sensibilidad y especificidad), requiere el uso de métodos estadísticos diferentes a los utilizados en las RS de terapia, para obtener una medida única de precisión del test diagnóstico evaluado.

Este artículo aborda algunos de dichos desafíos, especialmente aquellos relacionados con la búsqueda de estudios acerca de la precisión de los test

[1]Unidad de Medicina Basada en Evidencia.
[2]Departamento de Medicina Familiar, Escuela de Medicina, Pontificia Universidad Católica de Chile, Santiago de Chile

Correspondencia a: Dra. Gladys Moreno Gómez. Departamento de Medicina Familiar, Escuela de Medicina, Pontificia Universidad Católica de Chile. Lira 44, primer piso. E mail: gmoreno@med.puc.cl

diagnósticos y las ventajas y desventajas de diferentes formas de combinar los resultados en un metaanálisis. Aquellos aspectos relacionados con la evaluación crítica de los estudios primarios han sido abordados en artículos previos de esta sección[2].

Identificación de los estudios relevantes: la búsqueda

La búsqueda de estudios primarios para ser incluidos en una RS sobre la precisión de tests diagnósticos es más compleja que para las RS de efectividad de intervenciones, debido a la diversidad de las fuentes en dónde se publican este tipo de estudios y la carencia de términos de indexación apropiados. Por ello, la estrategia de búsqueda debe diseñarse cuidadosamente basada en una descripción clara y explícita de la población que recibe el test evaluado, el test propiamente tal, la enfermedad "blanco" y los diseños de estudio. Las fuentes de información deberán incluir necesariamente las bases de datos electrónicas (e.g. MEDLINE, EMBASE), pero además deberán extenderse al uso de otras fuentes como los listados de referencias de los estudios primarios identificados, la consulta a expertos y la "literatura gris".

Respecto a las bases de datos electrónicas, gran parte de la investigación se ha focalizado en el desarrollo de filtros para estudios diagnósticos (principalmente para búsqueda en MEDLINE) y su evaluación[5,6]. Sin embargo, todavía es un área en desarrollo donde no existen algoritmos estándar y la recomendación habitual es contar con un asesor metodológico con conocimiento avanzado de la estructura de las respectivas bases de datos. La Colaboración Cochrane ha iniciado el proceso de producir revisiones sistemáticas en esta área y, por lo tanto, probablemente contará en el futuro con una base de datos de estudios que será posible utilizar, tal como actualmente se encuentra disponible una para estudios y revisiones sistemáticas de la efectividad de intervenciones.

Evaluando críticamente la calidad de los estudios identificados

En el contexto de los estudios que evalúan la precisión de un test diagnóstico, la evaluación de la calidad debiera considerar aspectos del diseño y ejecución de dichos estudios, como la especificación de la población, la descripción de las técnicas diagnósticas evaluadas y su interpretación, y detalles de la forma como se definió y obtuvo el estándar de referencia (*"gold standard"*)[2].

Aunque la literatura relacionada con la evaluación de la calidad de este tipo de estudios es escasa en comparación con la de estudios de efectividad de intervenciones[7-10], dos desarrollos recientes merecen ser mencionados. El grupo STARD (*Standards for Reporting of Diagnostic Accuracy*) publicó en el año 2003 un listado de 25 ítems como una guía para mejorar la calidad del reporte de todos los aspectos de un estudio diagnóstico[11]. En estricto rigor, dicho listado no fue desarrollado como una herramienta de evaluación de calidad, pero muchos de sus ítems han sido incluidos en un desarrollo más reciente denominado QUADAS (*Quality Assessment of Diagnostic Accuracy Studies*). Esta herramienta consiste en 14 ítems que cubren aspectos como el espectro de pacientes, el estándar de referencia, el sesgo por progresión de enfermedad, sesgos de verificación y revisión, sesgo de incorporación, ejecución del test, pérdidas de seguimiento, y resultados intermedios[12].

La manera como se debiera incorporar la evaluación de la calidad en la síntesis es materia de debate[13,14]. Una aproximación relativamente simple es excluir los estudios de baja calidad. Una alternativa menos drástica es incorporar los puntajes de calidad como "pesos relativos" en el análisis estadístico. Sin embargo, la cuantificación de dichos "pesos" es controversial y su racionalidad estadística no es clara. Finalmente, una alternativa favorecida por algunos autores es realizar un análisis de sensibilidad para evaluar la contribución de los estudios de calidad baja a los resultados del metaanálisis[15]. Dicha evaluación compara los resultados del análisis estadístico con la inclusión y exclusión de estudios específicos.

Sintetizando las medidas de precisión del test

El cálculo o no de un estimador que resuma la precisión del test, depende del número y calidad metodológica de los estudios primarios incluidos y el grado de heterogeneidad de sus estimadores de precisión diagnóstica. En la literatura se sugiere una serie de pasos a seguir[15,16]:

1. Presentar los resultados de los estudios individuales
2. Evaluar la presencia de heterogeneidad
3. Evaluar la presencia de un "efecto de punto de corte" implícito

4. Elegir el modelo apropiado de análisis estadístico

La presentación de los estudios individuales debiera incluir información general del estudio (año de publicación, lugar de realización, número de pacientes enfermos y no enfermos, selección de los pacientes, procedimiento detallado de realización del test evaluado, estándar de referencia utilizado, independencia en la realización del test y estándar de referencia) así como un resumen de los resultados. Dada la naturaleza asimétrica de la mayoría de los test diagnósticos (algunos son buenos para descartar una condición y otros para confirmarla), es importante presentar un par de medidas complementarias como, por ejemplo, sensibilidad y especificidad[17].

A pesar de establecer criterios de inclusión relativamente restringidos, la mayoría de las revisiones muestran una importante heterogeneidad en los resultados de los estudios incluidos. Esta puede ser debida al azar o a diferencias en las características clínicas o metodológicas de los estudios. Aunque existen tests estadísticos para evaluar la presencia de heterogeneidad en los resultados más allá del azar, un método básico, pero bastante informativo es el uso de gráficos que muestren los resultados de los estudios individuales y sus intervalos de confianza (Figura 1). Esto permite la evaluación subjetiva de la variación en los resultados del grupo de estudios incluidos en la revisión, tanto en relación a los estimadores puntuales como a la sobreposición de sus intervalos de confianza.

Las estimaciones de la precisión de un test diagnóstico pueden diferir cuando los estudios no utilizan el mismo punto de corte para el test o el *"gold standard"*. La interpretación de muchos tests depende de factores humanos (radiología, anatomía patológica, examen clínico), por lo que diferentes estudios pueden utilizar implícitamente diferentes puntos de corte. Es posible evaluar estadísticamente este "efecto de punto de corte" utilizando coeficientes de correlación entre la sensibilidad y especificidad de los estudios incluidos[18]. La relevancia de este efecto está relacionada directamente al tipo de modelo de análisis estadístico y a la medida resumen de precisión que se utilizará en el metaanálisis.

Al igual que en el cálculo de un estimador único de efecto en los metanálisis de estudios randomizados[19,20], al estimar una medida única de precisión de un test diagnóstico se pueden utilizar modelos de efectos fijos o de efectos aleatorios, dependiendo de la existencia o no de heterogeneidad en los resultados de los estudios[21]. Sin embargo, uno de los aspectos más complejos es la elección e interpretación de la medida que se utilizará como estimador único de la precisión del test. Esto dependerá de la manera como es expresado el resultado del test en los estudios primarios (dicotómico, ordinal o continuo), y de la presencia de heterogeneidad o "efecto de punto de corte". Dentro de las opciones está el uso de estimadores que resumen las medidas habituales de precisión utilizadas en los estudios primarios, como sensibilidad, especificidad o *"likelihood ratios"* (LRs)[17]. Sin embargo, todavía existe controversia respecto a una serie de aspectos metodológicos y estadísticos relacionados con su uso[22,23], por lo cual es frecuente el uso de medidas de mayor complejidad estadística como el *"odds ratio"* diagnóstico (DOR) o las curvas ROC (*receiver operating characteristic*)[18,24]. Recomendamos a los lectores interesados revisar la literatura especializada en el tema[16,25].

En la mayoría de los casos los estudios primarios son heterogéneos y presentan limitada información, por lo que habitualmente es difícil realizar un metaanálisis. En este caso, deben buscarse causas para dicha heterogeneidad, como las características de los estudios, diferentes umbrales y características del test evaluado. Si no es posible o recomendable realizar un metaanálisis, la revisión puede limitarse a un análisis descriptivo cualitativo de la información disponible.

Ejemplos

Presentamos dos ejemplos de RS de test diagnósticos relevantes para la práctica médica general: la precisión del electrocardiograma en el diagnóstico de hipertrofia ventricular izquierda (HVI) en pacientes con hipertensión arterial[26] y la precisión de la historia y el examen físico para identificar pacientes con cefalea que requieren neuroimágenes[27].

En el primer ejemplo, se estudió la precisión de 6 índices electrocardiográficos para el diagnóstico de HVI. La búsqueda de estudios se realizó en bases de datos electrónicas como MEDLINE y EMBASE, referencias de estudios relevantes y revisiones previas, y expertos. Se identificaron 1.761 referencias. Dos revisores evaluaron los resumenes, considerando 152 referencias como potencialmente elegibles y, después de evaluar los artículos en texto completo, se incluyeron 21 estudios con un total de 5.608 pacientes. Tres estudios fueron considerados de alta calidad, 11 de

calidad intermedia, y 7 de baja calidad. La mediana de prevalencia del diagnóstico de HVI fue 33% en atención primaria y 65% en atención secundaria. Los estudios analizaron 12 criterios electrocardiográficos diferentes, pero la revisión se focalizó en los 6 criterios más frecuentemente utilizados. Para todos ellos, la mayoría de los estudios mostró baja sensibilidad y alta especificidad. La mediana de la sensibilidad para los diferentes índices varió entre 10,5% y 21%, mientras que la mediana de la especificidad fluctuó entre 89% y 99%. Asimismo, la mediana del LR negativo fue similar para los diferentes criterios, variando entre 0,85 y 0,91. Por otro lado, se registró mayor variación en la mediana del LR positivo, que fluctuó entre 1,9 y 5,9. Utilizando algunos de los valores promedios, un electrocardiograma negativo para HVI reduciría la probabilidad típica de 33% a 31% en atención primaria y de 65% a 63% en atención secundaria. Los autores concluyen que los criterios electrocardiográficos no debieran ser utilizados para descartar HVI en pacientes con hipertensión.

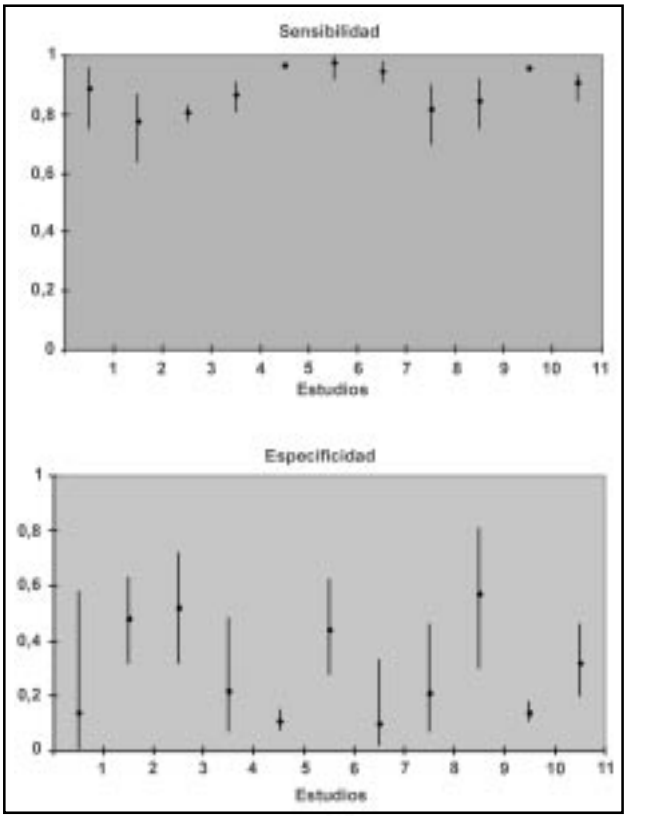

Figura 1. Estimadores puntuales e intervalos de confianza de la sensibilidad y especificidad de 11 estudios acerca de la precisión del test de Lasègue en el diagnóstico de hernia discal en el síndrome de dolor lumbar (© 2002 Devillé et al; licensee BioMed Central Ltd. Disponible en http://www.biomedcentral.com/1471-2288/2/9)

En el segundo ejemplo, se estudió la precisión de una serie de elementos de la historia y el examen físico para identificar (descartar) aquellos pacientes con cefalea que debieran ser sometidos a mayor estudio de imágenes (neuroimágenes). La búsqueda de estudios se realizó en MEDLINE (1966-2005). Los resumenes fueron evaluados independientemente por dos de los autores para decidir los estudios a incluir. Se incluyeron 11 estudios que daban cuenta de 3.725 pacientes con cefalea crónica en ámbitos ambulatorios, hospitalarios y de emergencia. Diez de los 11 estudios fueron de pobre calidad metodológica. La prevalencia de patología intracraneal significativa fue variable dependiendo del contexto del estudio. Se analizaron múltiples variables clínicas, dentro de las cuales algunas estuvieron asociadas a la presencia de una anormalidad intracraneana relevante, de acuerdo al metaanálisis de los LR positivos: cefalea tipo "*cluster*" (LR+ 10,7; IC 95% 2,2-52); hallazgos anormales en el examen neurológico (LR+ 5,3; IC 95% 2,4-12); cefalea poco definida (LR+ 3,8; IC 95% 2,0-7,1); cefalea con aura (LR+ 3,2;IC 95% 1,6-6,6); cefalea agravada por el ejercicio o la maniobra de Valsalva (LR+ 2,3: IC 95% 1,4-3,8); y cefalea con vómitos (LR+ 1,8; IC 95% 1,2-2,6). Considerando la calidad de la evidencia los hallazgos más robustos se relacionaron con hallazgos anormales en el examen neurológico. Ningún elemento clínico fue útil para descartar condiciones patológicas relevantes.

Conclusiones

El proceso diagnóstico juega un rol fundamental en la práctica clínica de cualquier médico y la información válida y confiable acerca de la precisión de los tests diagnósticos es esencial para tomar decisiones que finalmente beneficien a los pacientes. Las revisiones sistemáticas en esta área representan una manera eficiente de disminuir el error y sesgo en la

estimación de la precisión de los test diagnósticos. Sin embargo, todavía existen desafíos metodológicos y prácticos que deberán ser abordados en el corto y mediano plazo para que puedan ser utilizados con mayor confianza en la práctica clínica habitual. Este artículo ha revisado algunos de ellos y ha introducido los conceptos generales de este tipo de revisiones. El trabajo de la Colaboración Cochrane en el área puede representar un importante impulso para esta tarea, similar a aquel que a mediados de los 90 realizó en el área de las revisiones de los efectos de las intervenciones sanitarias.

Referencias

1. Capurro D, Rada G. El proceso diagnóstico. *Rev Med Chile* 2007; 135: 534-8.
2. Valenzuela L, Cifuentes L. Validez de estudios de test diagnósticos. *Rev Med Chile* 2008; 136: 401-4.
3. Deeks JJ. Systematic reviews of evaluations of diagnostic and screening tests. *BMJ* 2001; 323: 157-62.
4. Letelier Lm, Manríquez Jj, Rada G. Revisiones sistemáticas y metanálisis: ¿son la mejor evidencia? *Rev Med Chile* 2005; 133: 246-9.
5. Haynes RB, Wilczynski NL for the Hedges Team. Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical Surrey. *BMJ* 2004; 328: 1040-2.
6. Wilczynski NL, Haynes RB and the Hedges Team. EMBASE search strategies for identifying methodologically sound diagnostic studies for use by clinicians and researchers. *BMC Med* 2005; 3: 7.
7. 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.
8. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 1998; 352: 609-13.
9. Moher D, Cook DJ, Jadad AR, Tugwell P, Moher M, Jones A et al. Assessing the quality of reports of randomised trials: implications for the conduct of meta-analyses. *Health Technol Assess* 1999; 3: 1-98.
10. Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. *Ann Intern Med* 2001; 135: 982-9.
11. Bossuyt P, Reitsma J, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ* 2003; 326: 41-4.
12. Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, 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.
13. Leeflang M, Reitsma J, Scholten R, Rutjes A, Di Nisio M, Deeks J et al. Impact of adjustment for quality on results of metaanalyses of diagnostic accuracy. *Clin Chem* 2007; 53: 164-72.
14. 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.
15. Gatsonis C, Paliwal P. Meta-analysis of diagnostic and screening test accuracy evaluations: methodologic primer. *AJR* 2006; 187: 271-81.
16. Deville WI, Buntinx F, Bouter LM, Montori VM, De Vet HCW, Van Der Windt Dawm et al. Conducting systematic reviews of diagnostic studies: didactic guidelines. *BMC Med Res Methodol* 2002; 2: 9.
17. Salech F, Mery V, Larrondo F, Rada G. Estudios que evalúan un test diagnóstico: interpretando sus resultados. *Rev Med Chile* 2008; 136: 1203-8.
18. Moses LE, Shapiro D, Littenberg B. Combining independent Studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med* 1993; 12: 1293-316.
19. Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Prog Cardiovasc Dis* 1985; 27: 335-71.
20. Dersimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clin Trials* 1986; 7: 177-88.
21. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med* 2001; 20: 2865-84.
22. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificy produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 2005; 58: 982-90.
23. Zwinderman AH, Bossuyt PM. We should not pool diagnostic likelihood ratios in systematic reviews. *Stat Med* 2008; 27: 687-97.
24. Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Stat Med* 2002; 21: 1237-56.
25. Deeks JJ. Systematic reviews of evaluations of diagnostic and screening tests. En: Egger M, Davey Smith G, Altman DG eds. Systematic reviews in health care. Meta-analysis in context. London: BMJ Books; 2001; 248-82.
26. Pewsner D, Juni P, Egger M, Battaglia M, Sundstrom J, Bachmann LM. Accuracy of electrocardiography in diagnosis of left ventricular hypertrophy in arterial hypertension: systematic review. *BMJ* 2007; 335: 711-4
27. Detsky ME, Mcdonald DR, Baerlocher MO, Tomlinson GA, Macrory DC, Booth CM. Does this patient with headache have a migraine or need neuroimaging? *JAMA* 2006; 296: 1274-83.