

Estudios de evaluación de la validez de una prueba diagnóstica: revisión sistemática y metanálisis

Javier Zamora Romero, María Nieves Plana, Víctor Abraira Santos

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

Nefrología 2009;29(Sup. Ext. 6):15-20.

RESUMEN

Cada vez con mayor frecuencia se encuentran en la literatura revisiones sistemáticas y metanálisis que evalúan la validez de las pruebas diagnósticas. Si bien existe un cierto paralelismo con las homólogas revisiones de la eficacia de intervenciones terapéuticas, estas revisiones tienen peculiaridades que es preciso enfatizar. En este artículo se presenta, de forma divulgativa, el proceso de realización de una revisión sistemática de estudios de validez diagnóstica y su posterior metanálisis.

INTRODUCCIÓN

Desde que se propone la utilización de una prueba o estrategia diagnóstica hasta que se incorpora finalmente a la práctica clínica debiera existir un proceso sistemático de evaluación. El marco conceptual para la evaluación de pruebas diagnósticas ha ido evolucionando a lo largo de los años. Ha pasado de considerarse un proceso secuencial en fases¹, mimetizando en mayor o menor medida las fases I a IV del ensayo clínico, a ser una evaluación globalizadora que considera los aspectos técnicos de factibilidad, reproducibilidad, validez e impacto socioeconómico, y el contexto clínico donde se aplicará la prueba². Este último aspecto es de vital importancia para determinar las dimensiones más adecuadas de evaluación de la prueba o estrategia diagnóstica. El rol reservado para la prueba, bien sea el de reemplazar a una existente, incorporarla como una prueba previa de cribado –o *triage*–, o bien utilizarla como una prueba añadida a las que ya se realizan, determinará la dimensión de evaluación más relevante, así como el mejor diseño para determinar sus ventajas^{3,4}.

En los últimos tiempos estamos asistiendo a un periodo de profunda reflexión entre la comunidad científica dedicada a la evaluación de pruebas diagnósticas. Por un lado, la reciente pu-

blicación de la guía de graduación de la calidad de evidencias sobre pruebas y estrategias diagnósticas (GRADE)⁵ y, por otro, la inclusión de Revisiones Sistemáticas de validez de pruebas diagnósticas en la Biblioteca de la Colaboración Cochrane⁶ son dos hitos dignos de destacar.

En el presente artículo expondremos las características de las revisiones sistemáticas y metanálisis de estudios de validez de una prueba diagnóstica. Asumiendo que el lector está más familiarizado con las revisiones sistemáticas de intervenciones terapéuticas, durante los siguientes apartados ocasionalmente trazaremos el paralelismo entre ambos tipos de revisión para enfatizar las peculiaridades de las revisiones de diagnóstico.

REVISIÓN SISTEMÁTICA

El proceso de la revisión sistemática sigue las habituales fases que se inician con el planteamiento del objetivo de la revisión, la búsqueda bibliográfica y selección de artículos a incluir en la revisión, la evaluación de sus características y su calidad metodológica y el posterior análisis estadístico o metanálisis de sus resultados.

En cuanto al objetivo, las revisiones de diagnóstico con frecuencia se limitan a evaluar el rendimiento de una prueba diagnóstica respecto a un patrón de referencia (*gold standard*) sin abordar una comparación explícita con otras pruebas alternativas. El contexto clínico donde se aplicarán los resultados de la revisión a veces no es explícito, lo que dificultará su aplicabilidad. Es relativamente fácil utilizar los resultados de una revisión de tratamiento para la toma de decisiones para un paciente en concreto, mientras que en el caso del diagnóstico, esto es más complejo. Por ello, con frecuencia las evidencias arrojadas por la revisión sistemática se utilizan para la toma de decisiones a nivel de gestores y decisores con un punto de vista más cercano a la evaluación de tecnologías sanitarias que a la práctica clínica.

Como en las revisiones de tratamiento, otro de los objetivos de la revisión es analizar los factores que afectan al rendimiento diagnóstico de una determinada prueba, como pueden ser factores relacionados con la población estudiada (espectro de

Correspondencia: Javier Zamora Romero
Unidad de Bioestadística Clínica.
Hospital Ramón y Cajal. Madrid.
jzamora.hrc@salud.madrid.org
javier.zamora@hrc.es

la enfermedad, ámbito de estudio, etc.), características de la prueba y con el diseño del estudio.

Búsqueda y selección de artículos

Como en cualquier revisión sistemática, debe realizarse una búsqueda de estudios exhaustiva, objetiva y reproducible de la investigación primaria. Ésta no debe limitarse sólo a las bases de datos electrónicas, sino que debe complementarse con búsquedas manuales en las listas de referencias bibliográficas de los artículos incluidos, en los resúmenes de congresos relevantes, consultas con investigadores, registros de organismos evaluadores de investigación, etc. Además de las bases electrónicas habituales (MEDLINE y EMBASE) existen bases de datos específicas de estudios de diagnóstico, como la base MEDION⁷ que recoge revisiones publicadas de estudios de diagnóstico y de cribado.

La identificación de artículos de diagnóstico presenta más dificultades que la búsqueda de ensayos clínicos. No existe como tal un término MeSH (*Medical Subject Heading*) específico que sea comparable al término «randomized controlled trial». El término «sensitivity and specificity» puede ser el más adecuado, pero no en todas las bases de datos está bien indexado. Muchos de los estudios de diagnóstico se realizan alrededor de la propia práctica clínica sin la existencia de un protocolo registrado y/o aprobado por comités éticos de investigación, con lo que se dificulta su seguimiento. No existe una base de datos centralizada de estudios de diagnóstico equivalente a la de ensayos clínicos. Además, hay estudios que presentan resultados de validez diagnóstica de una prueba sin que éste sea su objetivo principal. Todo esto dificulta el proceso de búsqueda y conlleva que no sea recomendable el uso de filtros metodológicos para intentar restringir y focalizar la búsqueda⁸.

La valoración del sesgo de publicación en los estudios de diagnóstico es mucho más compleja que su equivalente en estudios de tratamiento. Los *funnel plot* y demás métodos utilizados para evaluar dicho sesgo de publicación en las revisiones de tratamiento son discutidos para los estudios de diagnóstico^{9,10}. Por otro lado, dado que habitualmente los estudios de diagnóstico no comparan pruebas diagnósticas entre sí, la publicación no suele estar condicionada por la presencia o no de una significación estadística asociada a dicha comparación.

En cualquier caso, como en las revisiones sistemáticas homólogas de tratamiento, el flujo de los estudios desde la búsqueda inicial hasta la realización del metanálisis debería representarse siguiendo la recomendación de las guías QUOROM¹¹, es decir, mediante una gráfica donde consten los distintos estudios con sus exclusiones descritas de acuerdo a las causas.

Evaluación de la calidad de los estudios

Un aspecto clave en toda revisión sistemática es la evaluación de la calidad metodológica de los estudios incluidos con la finalidad de identificar posibles fuentes de sesgos. Vinculado al

análisis de la calidad metodológica con que se diseñó y realizó el estudio, se encuentra el análisis de la forma en que se comunican los resultados del estudio. Es difícil distinguir entre ambos aspectos de la calidad dado que ambos están íntimamente relacionados. Para uno y otro análisis existen dos herramientas muy populares en la actualidad. Por una parte, 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 las publicaciones para permitir así a los lectores evaluar los potenciales sesgos del estudio y juzgar sobre su generalización^{12,13}. Esta iniciativa es al área de investigación sobre diagnóstico lo que el CONSORT fue a la investigación en tratamiento¹⁴. Por otra parte, se encuentra el cuestionario QUADAS, ideado específicamente para la evaluación de la calidad de los estudios primarios incluidos en revisiones sistemáticas de diagnóstico¹⁵⁻¹⁷. El cuestionario contiene una serie de ítems relacionados con aspectos del diseño y el análisis del estudio que cubren desde el espectro de pacientes incluido hasta la presencia de los sesgos más frecuentes.

El uso que se puede dar a los resultados de este análisis de calidad es un tema de debate. Se propone desde una simple descripción de esta calidad con el objeto de valorar el alcance de las evidencias disponibles, hasta la propuesta más drástica de excluir del análisis a los estudios de más baja calidad. En lo que sí parece haber un cierto consenso es en desaconsejar resumir la calidad en una única puntuación numérica y en la dificultad de incorporar explícitamente la calidad en la ponderación de los estudios primarios a la hora de analizar los datos¹⁸. Una alternativa común es la realización de análisis de sensibilidad comparando los resultados que se obtienen incluyendo y excluyendo determinados estudios en función de su calidad o de determinadas características del diseño.

A partir de las publicaciones de Lijmer¹⁹ inicialmente, y Rutjes²⁰ más tarde, existe un consenso generalizado en recomendar la exclusión de los estudios con diseño de casos y controles, dado que sobrestiman la validez de las pruebas y pueden considerarse fases previas de la evaluación de la validez¹.

Análisis estadístico

En general, el metanálisis es un proceso en dos etapas²¹. En un primer paso se estiman los resultados de cada estudio, aunque en el caso de la evaluación de pruebas diagnósticas cada estudio es resumido no por un índice, como en los estudios de evaluación de tratamientos, sino por una pareja de índices que describen la validez de la prueba. Habitualmente, estos dos índices son sensibilidad y especificidad, o bien los cocientes de probabilidad positivo y negativo²². Debe huirse, en la medida de lo posible, de utilizar los valores predictivos pues, como es bien sabido, dependen de la prevalencia de la condición clínica que se diagnostica y ésta puede ser muy variable de estudio a estudio. Sin embargo, a veces es el único índice que se puede obtener por las características del método de referencia²³. En un segundo paso se deben calcular índices globales

de validez para lo que se han propuesto diversos métodos, que se expondrán más adelante.

Abordar un metanálisis sólo es apropiado en el caso de que exista homogeneidad metodológica en los estudios incluidos en la revisión, es decir, que todos ellos hayan evaluado la misma prueba diagnóstica, usando un mismo método de referencia y en pacientes similares. Para valorar estos aspectos se deben extraer los datos pertinentes de los estudios, utilizando formularios *ad hoc* y en la medida de lo posible por duplicado, para incrementar el grado de objetividad del proceso. Estos datos se referirán a las características del diseño del estudio, su calidad metodológica, las características de los pacientes y del ámbito del estudio, de la prueba evaluada y de referencia y, finalmente, los resultados observados. Es importante destacar que este último aspecto a veces limita la inclusión de un estudio en el análisis. Si no fuera posible extraer los datos de validez diagnóstica de un estudio individual en forma de tabla de clasificación cruzada, el estudio debe ser excluido de la revisión.

Evaluación de la heterogeneidad

Además de la homogeneidad metodológica mencionada, se debe valorar la eventual presencia de heterogeneidad estadística en los resultados. Esta evaluación puede hacerse gráficamente presentando la sensibilidad y especificidad de cada estudio en un *forest plot* (figura 1). En estos gráficos se representan los estimadores de los índices junto con sus intervalos de confianza, y suelen presentarse pareados y ordenados de acuerdo a uno de los índices. Lo habitual es encontrar cierta dispersión por el azar en la selección de las muestras de los estudios, pero otros factores la pueden aumentar y el metanálisis debe explorar e identificar estas posibles fuentes de heterogeneidad. Una evaluación estadística de la «cantidad» de heterogeneidad

presente se puede realizar mediante una prueba de la ji al cuadrado. Es frecuente cuantificar su magnitud mediante el índice de inconsistencia (I^2) y una guía, no exenta de críticas, para su interpretación clasifica el índice en baja (25-50%), media (51-75%) y sustancial (>75%)²⁴.

Una fuente característica de heterogeneidad en los metanálisis de validez de las pruebas diagnósticas es la que surge porque los estudios incluidos pueden haber usado diferentes umbrales para definir qué es un resultado positivo de la prueba. Este efecto es conocido como «efecto umbral». La definición del umbral de positividad a veces puede ser explícita, como en el caso de la proteína C reactiva, o implícita como en el caso de pruebas con interpretación subjetiva (pruebas de imagen) o cuando los resultados de la prueba puedan verse afectados por la calibración del aparato de medida. Desafortunadamente, este efecto está presente con mucha frecuencia en la evaluación de pruebas diagnósticas.

Efecto umbral

Para explorar esta fuente de variación es útil representar en una gráfica las parejas de sensibilidad y especificidad de cada estudio en un plano ROC (figura 2). En este plano, la zona más cercana a la esquina superior izquierda supone un buen rendimiento diagnóstico mientras que la zona central, la diagonal en la que sensibilidad y especificidad son iguales, representa una nula capacidad diagnóstica. Si existiera «efecto umbral», los puntos en el plano mostrarían un patrón curvilíneo. Cambiando el umbral de positividad de una prueba se obtendría una mayor (o menor) sensibilidad con el consiguiente efecto contrario sobre la especificidad. Además de gráficamente, este «efecto umbral» puede evaluarse mediante el coeficiente de correlación de Spearman entre la sensibilidad y la especificidad, ya que de existir dicho efecto mostraría una correlación inversa²⁵.

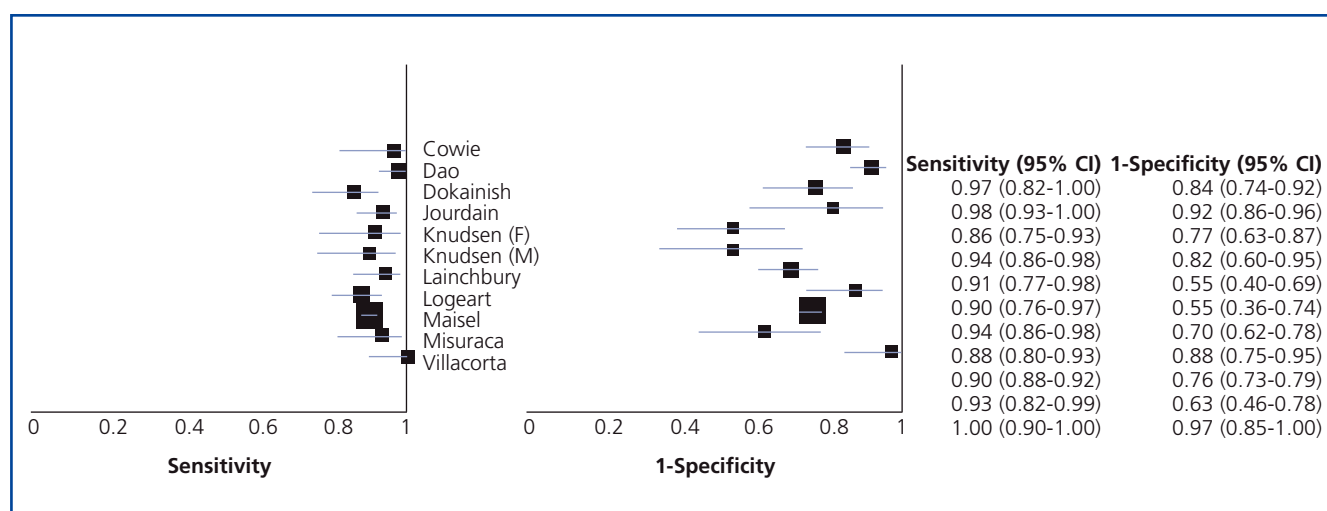


Figura 1. *Forest plot* de sensibilidades y especificidades del BNP para el diagnóstico de insuficiencia cardiaca (datos de la revisión de Latour-Pérez J. Eur J Heart Fail 2005).

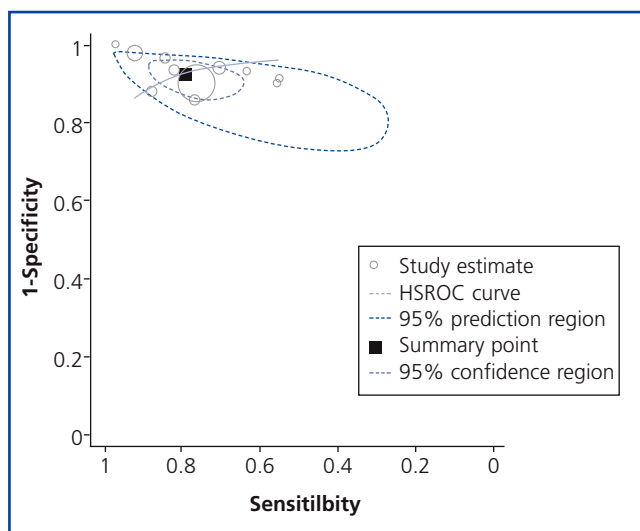


Figura 2. Plano sROC con la estimación de la curva sROC y las regiones de confianza y predicción de los mismos datos de la figura 1.

Los métodos estadísticos más robustos propuestos para el metanálisis tienen en cuenta esta correlación entre sensibilidad y especificidad, y lo hacen mediante la estimación de una curva ROC resumen (sROC) de los estudios incluidos. En presencia de «efecto umbral», no es adecuado aglomerar los índices de sensibilidad y especificidad (o cocientes de probabilidad) independientemente uno del otro ignorando su correlación.

Sin embargo, en limitadas ocasiones los resultados de los estudios primarios son homogéneos y puede descartarse la presencia tanto de efecto umbral como de otras fuentes de heterogeneidad. En esta situación, el resultado global de la revisión podría obtenerse a partir de la combinación ponderada de los índices de los estudios individuales. Como siempre, esta combinación puede hacerse mediante un modelo de efectos fijos o bien mediante un modelo de efectos aleatorios, dependiendo de la magnitud de la heterogeneidad.

Se han propuesto varios métodos para estimar la curva sROC. El primero de ellos, debido a Moses, et al.²⁶, se basa en estimar la regresión lineal entre dos variables creadas a partir de los índices de validez de cada estudio. Estas variables son D y S , que representan, respectivamente, el logaritmo del odds ratio diagnóstico (ORD)²⁷ y una medida indirecta del umbral de positividad de la prueba calculada como:

$$S = \text{logit}(TVP) + \text{logit}(TFP).$$

TVP y TFP son las tasas de verdaderos y falsos positivos, respectivamente, y $\text{logit}(TVP)$ es el logaritmo de TVP dividido por su complementario ($1 - TVP$). Es fácil ver que ambas tasas de resultados positivos cambian con el umbral de positividad de la prueba.

La propuesta de Moses consiste en ajustar el modelo $D = a + bS$. El contraste sobre si hay variación del rendimiento diag-

nóstico (medido por el ORD) con el umbral es equivalente al realizado sobre el parámetro b del modelo. Si $b = 0$ no hay variación y el método da lugar a una curva sROC simétrica mientras que si $b \neq 0$ existe variación del rendimiento con el umbral y la curva sROC es asimétrica y se obtiene deshaciendo la transformación a los ejes originales del plano ROC.

El modelo puede extenderse para analizar el efecto de otros factores sobre el rendimiento diagnóstico (ORD). Estos factores pueden ser relativos al diseño del estudio, características de los pacientes o del test y se incluirían en el modelo anterior como covariables²⁸.

Se han propuesto distintos estadísticos útiles para resumir una curva sROC. El más habitual es el área bajo la curva (ABC) que resume el rendimiento diagnóstico de la prueba en un solo número²⁹: las pruebas perfectas tienen un ABC cercano a 1 y las inútiles cercano a 0,5. Este número puede interpretarse como la probabilidad de clasificar correctamente a dos sujetos seleccionados al azar, uno con la enfermedad y otro sin ella. Finalmente, puede usarse el área para comparar el rendimiento de distintas pruebas diagnósticas. Otro estadístico útil es el índice Q^* , definido como el punto en el que la sensibilidad y la especificidad son iguales. En una curva simétrica este punto es el más cercano al extremo superior izquierdo del plano ROC. Por último, la curva sROC ajustada puede utilizarse para extrapolar una sensibilidad a partir de una especificidad dada o viceversa.

Modelos bivariantes y jerárquicos

El modelo de Moses presenta algunas limitaciones. Por una parte, no tiene en cuenta la distinta precisión con la que se estimaron la sensibilidad y la especificidad en cada estudio, tampoco incorpora la heterogeneidad entre estudios y, por último, la variable independiente del modelo es aleatoria y, por tanto, tiene error de medición³⁰. Para superar estas limitaciones, se han propuesto recientemente modelos de regresión más complejos para estimar la curva sROC.

El primero de ellos es un modelo de efectos aleatorios, bivalente, que parte de la asunción de que los logit de sensibilidad y especificidad siguen una distribución normal bivalente. El modelo contempla la eventual correlación entre ambos índices, modeliza la distinta precisión con la que han sido estimadas la sensibilidad y la especificidad e incorpora una fuente de heterogeneidad adicional debida a la varianza entre estudios³¹.

La segunda propuesta se refiere al modelo conocido como HSROC o modelo jerárquico. Es similar al modelo anterior, salvo que hace explícita la relación existente entre sensibilidad y especificidad a través del umbral. Como el anterior, también tiene en cuenta la heterogeneidad entre estudios^{1,32}.

Ambos modelos, bivalente y jerárquico, permiten obtener estimaciones promedio de sensibilidad y especificidad con sus co-

respondientes regiones de confianza y predicción. Las diferencias entre ambos modelos son pequeñas y recientemente se ha demostrado que, en ausencia de covariables, ambos abordajes son distintas parametrizaciones del mismo modelo³³. Aunque recientemente se ha abogado por la necesidad de utilizar estos métodos³⁴, los resultados proporcionados por estos métodos más sofisticados son muy similares a los obtenidos por el modelo de Moses³⁵.

Existe una gran variedad de paquetes estadísticos que pueden utilizarse para realizar los análisis descritos. Unos son de propósito general, como el SAS y el Stata, que a través de una serie de macros programadas por usuarios facilitan la obtención de los modelos descritos. Las más populares son las macros de Stata denominadas **midas**³⁶ y **metandi**³⁷, y la macro desarrollada para SAS denominada **metadas**³⁸. Otros programas específicos para el metanálisis de estudios de pruebas diagnósticas son el Meta-DiSc³⁹ y el RevMan⁴⁰ en su versión 5. Ambos realizan los análisis básicos mencionados en este artículo, si bien RevMan permite incorporar los resultados obtenidos con los modelos bivariantes.

Aunque nos hallamos en una época muy activa en el diseño de métodos para realizar las revisiones sistemáticas de pruebas diagnósticas, queda mucho camino por andar hasta llegar a equiparar estos métodos con los desarrollados en el ámbito de las revisiones de tratamiento. Estamos seguros de que en un futuro cercano irán apareciendo más desarrollos metodológicos que situarán a la investigación sobre pruebas diagnósticas en el lugar que le corresponde por su importancia en el proceso clínico-asistencial.

BIBLIOGRAFÍA

- Sackett DL, Haynes RB. The architecture of diagnostic research. *BMJ* 2002;324:539-41.
- Van den BA, Cleemput I, Aertgeerts B, Ramaekers D, Buntinx F. The evaluation of diagnostic tests: evidence on technical and diagnostic accuracy, impact on patient outcome and cost-effectiveness is needed. *J Clin Epidemiol* 2007;60:1116-22.
- Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ* 2006;332:1089-92.
- Lijmer JG, Bossuyt PM. Various randomized designs can be used to evaluate medical tests. *J Clin Epidemiol* 2009;62:364-73.
- Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ* 2008;336:1106-10.
- Leeflang MM, Bets-Ossenkopp YJ, Visser CE, Scholten RJ, Hooft L, Bijlmer HA, et al. Galactomannan detection for invasive aspergillosis in immunocompromized patients. *Cochrane.Database.Syst.Rev*.CD007394, 2008.
- MEDION Database. Disponible en: <http://www.mediondatabase.nl/>. [Consultado el 3 de julio 2009.] Ref Type: Computer Program.
- Leeflang MM, Scholten RJ, Rutjes AW, Reitsma JB, Bossuyt PM. Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies. *J Clin Epidemiol* 2006;59:234-40.
- Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol* 2005;58:882-93.
- Song F, Khan KS, Dinnes J, Sutton AJ. Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. *Int J Epidemiol* 2002;31:88-95.
- Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. *Lancet* 1999;354:1896-900.
- Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Ann Intern Med* 2003;138:W1-12.
- Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. *Ann Intern Med* 2003;138:40-4.
- Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials. *BMC Med Res Methodol* 2001;1:2.
- 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.
- 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.
- Zamora J, Abraira V. [Analysis of the quality of studies assessing diagnostic tests]. *Nefrología* 2008;28 Suppl 2:42-5.
- Whiting P, Harbord R, Kleijnen J. No role for quality scores in systematic reviews of diagnostic accuracy studies. *BMC Med Res Methodol* 2005;5:19.
- Lijmer JG, Mol BW, Heisterkamp S, Bossuyt PM, Prins MH, Van der Meulen JH, Bossuyt PM. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;282:1061-6.
- Rutjes AW, Reitsma JB, Di NM, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ* 2006;174:469-76.
- Deeks JJ. Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. *BMJ* 2001;323:157-62.
- Abraira V. Índices de rendimiento de las pruebas diagnósticas. *SEMERGEN* 2008;28:93-194.
- Houssami N, Ciatto S, Macaskill P, Lord SJ, Warren RM,

- Dixon JM, Irwig L. Accuracy and surgical impact of magnetic resonance imaging in breast cancer staging: systematic review and meta-analysis in detection of multifocal and multicentric cancer. *J Clin Oncol* 2008;26:3248-58.
24. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003;327:557-60.
 25. Deville WL, Buntinx F, Bouter LM, Montori VM, de Vet HC, van der Windt DA, Bezemer PD. Conducting systematic reviews of diagnostic studies: didactic guidelines. *BMC Med Res Methodol* 2002;2:9.
 26. 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.
 27. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol* 2003;56:1129-35.
 28. Lijmer JG, Bossuyt PM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. *Stat Med* 2002;21:1525-37.
 29. Walter SD. The partial area under the summary ROC curve. *Stat Med* 2005;24:2025-40.
 30. Gatsonis C, Paliwal P. Meta-analysis of diagnostic and screening test accuracy evaluations: methodologic primer. *Am J Roentgenol* 2006;187:271-81.
 31. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 2005;58:982-90.
 32. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med* 2001;20:2865-84.
 33. Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics* 2007;8:239-51.
 34. Harbord RM, Whiting P, Sterne JA, Egger M, Deeks JJ, Shang A, Bachmann LM. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. *J Clin Epidemiol* 2008;61:1095-103.
 35. Simel DL, Bossuyt PM. Differences between univariate and bivariate models for summarizing diagnostic accuracy may not be large. *J Clin Epidemiol* 2009; doi:10.1016/j.jclinepi.2009.02.007.
 36. Dwamena BA. Midas: A program for Meta-analytical Integration of Diagnostic Accuracy Studies in Stata. Division of Nuclear Medicine, Department of Radiology, University of Michigan Medical School, Ann Arbor, Michigan. 2007. Ref Type: Computer Program.
 37. Harbord RM. Metandi. Stata module for meta-analysis of diagnostic accuracy. Statistical Software Components, Boston College Department of Economics. Revisado el 15 abril 2008. Ref Type: Computer Program.
 38. METADAS: A SAS macro for meta-analysis of diagnostic accuracy studies. User guide version 1.0 beta. Diciembre 2008. Disponible en: <http://srdta.cochrane.org/en/clib.html>. [Consultado 3 julio 2009.] Ref Type: Computer Program.
 39. Zamora J, Abraira V, Muriel A, Khan K, Coomarasamy A. Meta-DiSc: a software for meta-analysis of test accuracy data. *BMC Med Res Methodol* 2006;6:31.
 40. Review Manager (RevMan) [Programa para ordenador]. Version 5.0. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration. 2008. Ref Type: Computer Program.