QUADAS-2: instrumento para la evaluación de la calidad de estudios de precisión diagnóstica

QUADAS-2: an instrument for the evaluation of the quality of diagnostic precision studies

Agustín Ciapponi

Resumen

En 2003 se desarrolló la herramienta QUADAS para valorar la calidad de los estudios de pruebas diagnósticas incluidos en una revisión sistemática, aunque la misma puede aplicarse en otros contextos. De la retroalimentación proporcionada por sus usuarios surgieron áreas de mejora, lo que condujo al desarrollo del instrumento QUADAS-2, que consta de cuatro dominios: 1) selección de los pacientes, 2) prueba índice, 3) prueba de referencia, 4) flujo y tiempos. Cada dominio es evaluado en términos de su riesgo de sesgos y los primeros tres dominios también se evalúan por su aplicabilidad. .

Abstract

In 2003, the QUADAS tool was developed to assess the quality of studies of diagnostic tests included in a systematic review, although the tool can be applied in other contexts. Feedback provided by its users determined the suggestion for areas to be improved, which led to the development of the instrument QUADAS-2, which analyzes four domains: 1) patient selection, 2) index test, 3) reference test, 4) flow and times. Each domain is evaluated in terms of their risk of bias, and the first three domains are also evaluated for their applicability.

Palabras clave: estudios de precisión diagnóstico, sesgos, calidad metodológica. **Key words:** diagnostic accuracy studies, bias, metodological quality.

Ciapponi A. QUADAS-2: instrumento para la evaluación de la calidad de estudios de precisión diagnóstica. Evid Act Pract Ambul. 2015;18(1):22-30.

Cómo leer un Artículo

Introducción

Las revisiones sistemáticas de estudios de precisión diagnóstica a menudo se caracterizan por resultados marcadamente heterogéneos debido a las diferencias metodológicas de los estudios incluidos, por lo que es fundamental evaluar adecuadamente aspectos que determinan la validez interna y externa de cada uno de ellos, y por ende su calidad.

Desde su publicación en 2003, el instrumento QUADAS ha sido ampliamente utilizado, por ejemplo, por la Agencia para la Investigación y la Calidad Sanitaria (AHRQ), la Colaboración Cochrane y por el Instituto Nacional de Excelencia Clínica del Reino Unido (NICE).

La herramienta original incluía 14 ítems que permitían evaluar la probabilidad de sesgos, las fuentes de variación (aplicabilidad) y la calidad del reporte. Cada elemento podía ser puntuado como "sí", "no" o "incierto"[1].

Presentamos aquí la herramienta mejorada, el QUADAS-2, desarrollada sobre la base de la experiencia obtenida con la herramienta original y sobre nuevas evidencias sobre las fuentes de sesgos y de variación en los estudios de precisión diagnóstica. Su elaboración fue descripta en el estudio de Whiting y col.[2], que sintetizamos a continuación.

El desarrollo del QUADAS-2 se basó en el enfoque de cuatro etapas propuesto por Moher y col[3]: la definición del foco, la revisión de la base de las evidencias, una reunión de consenso cara a cara y el perfeccionamiento de la herramienta a través de una prueba piloto. La principal decisión consistió en definir la "calidad" de este tipo de estudios tanto por la probabilidad de sesgos[§], como por las preocupaciones respecto de su aplicabilidad a la pregunta de investigación que aborda la revisión. Además, también fueron incluidos espacios para documentar juicios explícitos sobre el riesgo de sesgo con el objetivo de hacer la herramienta más informativa y transparente.

Descripción de la herramienta QUADAS-2

La herramienta completa QUADAS-2 está disponible en inglés (http://www.bris.ac.uk/quadas) y la proporcionamos aquí en español (ver el anexo 1). Fué diseñada para evaluar la calidad de los estudios primarios de precisión diagnóstica, complementando el proceso de extracción de datos de una revisión sistemática.

El instrumento considera cuatro dominios clave: la selección de pacientes, la prueba índice, la prueba de referencia y, finalmente, el flujo de los pacientes a través del estudio y los momentos de realización de la prueba índice y la de referencia (flujo y tiempos).

Su aplicación se completa en cuatro fases: 1) definición de la pregunta de la revisión, 2) adaptación de la herramienta y producción de una guía de revisión específica, 3) revisión del diagrama de flujo publicado para el estudio primario (o su construcción si no fue reportado) , 4) evaluación de los sesgos y la aplicabilidad. Cada dominio se evalúa en términos de su riesgo de sesgo, y en los primeros tres dominios se consideran además las preocupaciones acerca de su aplicabilidad. La herramienta incluye preguntas orientadoras para facilitar la valoración de los aspectos mencionados, remarcando algunas cuestiones de diseño de los estudios primarios.

Fase 1: Revisión de la pregunta

Los revisores primero comunican su pregunta de investigación en términos de pacientes, pruebas índice, pruebas de referencia y la condición diana o problema de salud a ser diagnosticado. Debido a que la precisión de una prueba puede depender del punto del camino diagnóstico (diagnostic pathway) en que va a utilizarse, es necesario describir el ámbito y el uso previsto de la prueba índice, la presentación del paciente y las pruebas diagnósticas previas.

Fase 2: Adaptación del QUADAS-2 a una revisión sistemática

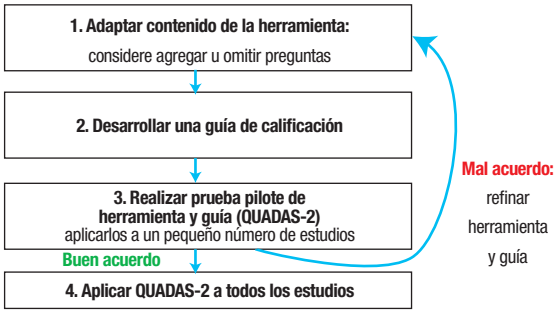
La herramienta QUADAS-2 debe ser adaptada a cada revisión sistemática mediante el agregado o la omisión de las preguntas que correspondan y la elaboración de una guía específica sobre cómo evaluar cada pregunta orientadora para juzgar el riesgo de sesgo (ver figura 1). El primer paso es considerar si alguna pregunta orientadora no se aplica a la revisión o si el núcleo de dichas preguntas no cubren adecuadamente las cuestiones específicas de la revisión. Por ejemplo, para una revisión de una prueba índice objetiva, puede ser apropiado omitir la pregunta orientadora sobre el cegamiento (quien interpreta la prueba diagnóstica que está siendo evaluada y quien interpreta los resultados de la prueba de referencia).

[§] Servicio de Medicina Familiar y Comunitaria. Hospital Italiano de Buenos Aires, Argentina. Coordinador Centro Cochrane Argentino IECS - Instituto de Efectividad Clínica y Sanitaria. agustin.ciapponi@hospitalitaliano.org.ar

§ El sesgo se produce cuando errores sistemáticos o limitaciones en el diseño o la realización de un estudio distorsionan sus resultados.

**La evidencia proporcionada por un estudio puede tener una aplicabilidad limitada si, frente a la pregunta de una revisión (o una pregunta de manejo clínico), los sujetos estudiados tenían diferentes características demográficas o clínicas, si la prueba índice fue aplicada o interpretada de forma diferente, o si la condición diana fue definida de manera diferente.

Figura 1: proceso de adaptación del QUADAS-2 a una revisión sistemática.

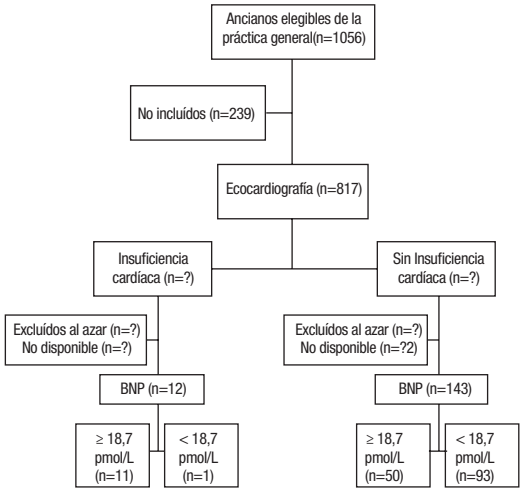


Los revisores deben evitar complejizar la herramienta mediante la adición de demasiadas preguntas orientadoras. Una vez acordado el contenido de la herramienta, debe desarrollarse la guía de calificación específica para dicha revisión. Al menos dos personas deben realizar en forma independientemente una prueba piloto de la herramienta. Si el acuerdo es bueno, la herramienta ya puede utilizarse para evaluar todos los estudios incluidos pero si el acuerdo es pobre, será necesario un mayor refinamiento.

Fase 3: Diagrama de Flujo

A continuación, los revisores deben revisar el diagrama de flujo publicado para el estudio primario, o trazar uno si este no hubiera sido reportado o si el diagrama publicado fuera inadecuado. El diagrama de flujo facilitará los juicios sobre los riesgos de sesgos y debería proporcionar información sobre el método de reclutamiento de los participantes (por ejemplo, el uso de una serie consecutiva de pacientes con síntomas específicos sospechosos de la condición diana o de los pacientes casos y los controles), el orden de ejecución de la prueba y el número de sujetos sometidos a la prueba índice y a la de referencia. Un diagrama dibujado a mano es suficiente, ya que este paso no necesita reportarse como parte de la evaluación QUADAS-2. En la figura 2 se presenta un diagrama de flujo de un estudio primario sobre el uso de los niveles de péptidos natriuréticos tipo B para diagnosticar la insuficiencia cardiaca, basado en un diseño de cohorte[4].

Figura 2: ejemplo de un diagrama de flujo del estudio.



Fase 4: Valoración sobre los sesgos y sobre la aplicabilidad

Probabilidad de sesgo

La primera parte de cada dominio se refiere a las valoraciones y consta de tres secciones: 1) información utilizada para apoyar la evaluación de la probabilidad de sesgo, 2) preguntas orientadoras, 3) el juicio sobre el riesgo de sesgo. El registro de la información utilizada para llegar a la decisión final tiene por objeto transparentar la calificación y facilitar la discusión entre los evaluadores independientes. Las preguntas orientadoras, facilitadoras de las evaluaciones son contestadas por "sí", "no", o "incierto" y están formuladas de tal manera que la respuesta "sí" indica bajo riesgo de sesgo.

El riesgo de sesgo se juzga como "bajo", "alto" o "incierto" utilizando las guías elaboradas en la fase 2. Si las respuestas a todas las preguntas orientadoras de un dominio son "sí", entonces la probabilidad de sesgo puede ser juzgada como "baja". Si se contesta "no" a cualquier pregunta orientadora, existe la posibilidad de sesgo y debería calificarse como "alto". La categoría de "incierta" debe utilizarse únicamente cuando los datos son insuficientes para emitir un juicio.

Aplicabilidad

Las secciones de aplicabilidad están estructuradas de una manera similar a las secciones de sesgo, pero no incluyen preguntas orientadoras. Los revisores registran la información soporte del juicio de aplicabilidad y evalúan luego su preocupación acerca de que el objetivo del estudio no coincida con la pregunta de la revisión. Las cuestiones acerca de la aplicabilidad son calificadas como preocupación "baja", "alta" o "incierta". Los juicios sobre aplicabilidad deben referirse a la fase 1, donde se registró la pregunta de la revisión. Una vez más, la categoría de "incierta" se debe utilizar únicamente cuando se presentan datos insuficientes.

Las siguientes secciones explican brevemente las preguntas orientadoras y el riesgo de sesgo o inquietudes sobre cuestiones de aplicabilidad para cada dominio.

Descripción del instrumento

Dominio 1: Selección de pacientes

Riesgo de sesgo: ¿Podría la selección de pacientes haber introducido sesgos?

1. ¿Se enroló una muestra consecutiva o aleatoria de pacientes?
2. ¿Se evitó un diseño de casos y controles?
3. ¿Se evitaron exclusiones inapropiadas?

Idealmente, un estudio debe enrolar una muestra consecutiva o aleatoria de pacientes elegibles con sospecha de enfermedad para evitar la posibilidad de sesgo. Los estudios que hacen exclusiones inadecuadas (por ejemplo, excluyendo pacientes "de difícil diagnostico") pueden dar lugar a una sobreestimación de la precisión diagnóstica. Por ejemplo, en una revisión de los anticuerpos anti-citrulina para el diagnóstico de artritis reumatoide[5], se encontró que algunos estudios habían incluido participantes consecutivos con diagnóstico confirmado. En estos estudios, la prueba mostró una mayor sensibilidad que en los estudios que incluyeron pacientes con sospecha de enfermedad, pero sin diagnóstico confirmado. Los estudios que reclutaron participantes con enfermedad conocida y un grupo de control sin la enfermedad pueden exagerar de manera similar precisión diagnóstica. Asimismo, cuando son

excluidos los pacientes con "banderas rojas" de la condición diana, también puede dar lugar a una subestimación de la precisión diagnóstica.

Aplicabilidad: ¿Hay preocupación de que la aplicación o la interpretación de la prueba que está siendo evaluada no coincidan con la pregunta de la revisión? Pueden existir preocupaciones acerca de la aplicabilidad de la prueba diagnóstica que está siendo evaluada si los pacientes incluidos en el estudio fueron diferentes a las contemplados por la pregunta de la revisión, por ejemplo, respecto de la gravedad de la condición clínica en cuestión, de sus las características demográficas, de la presencia de diagnósticos diferenciales o condiciones comórbidas, del ámbito del estudio y de las pruebas anteriormente aplicadas a dichos individuos. Por ejemplo, los tumores más grandes son más fáciles de ver que los más pequeños en los estudios por imágenes, al igual que los infartos de miocardio más grandes generan niveles más altos de enzimas cardíacas y son más fáciles de detectar que los pequeños, lo que promueve que la sensibilidad sea sobredimensionada.

Dominio 2: Prueba índice

Riesgo de sesgo: ¿Podría la realización o la interpretación de la prueba índice haber introducido sesgos?

1. ¿Fueron interpretados los resultados de la prueba índice sin conocimiento de los resultados de la de referencia? Vale destacar que el conocimiento de los resultados de la prueba de referencia puede influir en la interpretación de resultados de la prueba índice que está siendo evaluada. La probabilidad de sesgo se vincula con el componente de subjetividad que pueda tener la interpretación de la prueba índice y con el orden de administración de dichas pruebas diagnósticas.

2. Si se utilizó un umbral para definir la positividad o la negatividad de la prueba índice, ¿fue especificado previamente? La selección del umbral de la prueba para optimizar la sensibilidad y/o la especificidad puede dar lugar a una sobreestimación de sus resultados debido a un sobreajuste de su capacidad predictiva. En estos casos es probable que el desempeño de dicha prueba en la vida real (o en cualquier otra prueba independiente) termine siendo más pobre que el observado durante el estudio de investigación.

Aplicabilidad: ¿Hay preocupación de que la conducción de la prueba índice o su interpretación no coincidan con la pregunta de la revisión? Tanto las variaciones en las tecnologías, como la ejecución o la interpretación de los resultados pueden afectar a las estimaciones de la precisión diagnóstica de una prueba. Si los métodos de la prueba índice varían respecto de los especificados en la pregunta de la revisión, pueden existir preocupaciones acerca de la aplicabilidad de sus hallazgos. Por ejemplo, una mayor frecuencia del transductor ecográfico ha demostrado mejorar la sensibilidad para la evaluación de pacientes con trauma abdominal.

Dominio 3: Prueba de referencia

Riesgo de sesgo: ¿Podría la realización o la interpretación de la prueba de referencia haber introducido sesgos?

1. ¿Es probable que la prueba de referencia valore correctamente la condición diana? Las estimaciones de la exactitud de la prueba se basan en el supuesto de que la prueba de referencia es 100% sensible y en que los desacuerdos entre la prueba de referencia y la prueba índice se deben a problemas de clasificación incorrecta de la prueba índice.

2. ¿Fueron interpretados los resultados de la prueba de referencia sin conocimiento de los resultados de la prueba índice? La posibilidad de sesgo se relaciona con la influencia potencial del conocimiento previo de los resultados de la prueba índice sobre la interpretación de la prueba de referencia.

Aplicabilidad: ¿Hay preocupación de que la condición diana, clasificada como tal a través de la prueba de referencia, difiera de la población a la cual estaba referida la pregunta?
La prueba de referencia puede estar libre de sesgos, pero la condición diana diagnosticada puede diferir de la condición diana especificada en la pregunta de la revisión. Por ejemplo, en la definición de infección del tracto urinario, la prueba de referencia se basa generalmente en el urocultivo. Sin embargo, el umbral por encima del cual el resultado se considera positivo puede variar.

Dominio 4: Flujo y tiempos

Riesgo de sesgo: ¿Podría el flujo de pacientes haber introducido sesgos?

1. ¿Hubo un intervalo apropiado entre la prueba índice y la prueba de referencia?

Idealmente, los resultados de la prueba índice y de la de referencia deberían recolectarse al mismo tiempo. Si se produce un retraso, o si el paciente recibe algún tratamiento o es dejado a su evolución temporal entre la aplicación de ambas, la mejoría o el empeoramiento de su condición clínica pueden generar errores de clasificación. La ventana de tiempo entre ambas variará entre las diferentes condiciones clínicas. Por ejemplo, un retraso de unos días puede no ser problemático ante enfermedades crónicas, pero si ante condiciones agudas.

A la inversa, una prueba de referencia que implica un tiempo de seguimiento puede requerir un período mínimo para evaluar si la condición diana está presente. Por ejemplo, para evaluar la sensibilidad o la especificidad de las imágenes por resonancia magnética para el diagnóstico temprano de la esclerosis múltiple, se requiere de un seguimiento mínimo de aproximadamente diez años para definir si cada paciente evolucionó o no hacia dicha enfermedad.

2. ¿Fue aplicada en todos los individuos la misma prueba de referencia?

El sesgo de verificación ocurre cuando sólo una parte del grupo de estudio recibe la confirmación del diagnóstico a través la prueba de referencia, o si a algunos pacientes se les realizó la confirmación diagnóstica a través de una prueba de referencia diferente.
Dicho de otro modo, si los resultados de la prueba que está siendo evaluada influyen en la decisión de realizar o no la prueba de referencia o en cuál prueba de referencia utilizar, la estimación de la precisión diagnóstica puede estar sesgada. Por

ejemplo, en un estudio que había evaluado la precisión del dímero D para diagnosticar tromboembolismo pulmonar, el centellograma de ventilación-perfusión (prueba de referencia 1) solo fue realizado en los individuos a quienes el resultado del dimero D había sido positivo; mientras que el seguimiento clínico (prueba de referencia 2) sólo había sido implementado en quienes el resultado del Dímero D había sido considerado negativo. Esta decisión puede derivar en una clasificación errónea de algunos resultados, y por lo tanto de la determinación de la sensibilidad y la especificidad del método en cuestión.

3. ¿Fueron incluidos todos los pacientes en el análisis?

Todos los participantes incluidos en el estudio deben ser incluidos en el análisis, más allá de que su seguimiento haya sido completo o no ya que aumentan las probabilidades de sesgo si el número de pacientes incluidos difiere del número de pacientes incluidos en la tabla de "2 × 2", especialmente si los pacientes perdidos difieren sistemáticamente de los que completaron el seguimiento.

Incorporando QUADAS-2 en las revisiones sistemáticas de precisión diagnóstica

Debe remarcarse que QUADAS-2 no debe ser utilizado para generar un "puntaje sumario" de la calidad global de estudio que está siendo evaluado[6,7]. Por ejemplo, si luego de evaluar un estudio, valoramos que no genera preocupaciones sobre sobre sus riesgos de sesgos o sobre la aplicabilidad de sus resultados, sólo emitiremos como conclusión que no genera preocupaciones respecto de ambos dominios.

Como mínimo, las revisiones sistemáticas deben resumir los resultados de la evaluación QUADAS-2 para todos los estudios que fueron incluidos en la misma en formato tabular (ver el ejemplo de la tabla 1 y el Anexo 2 para construir su propia tabla) y formato gráfico (ver el ejemplo de la figura 3 y el Anexo 3: planilla de cálculo que le permitirá generar sus propios gráficos).

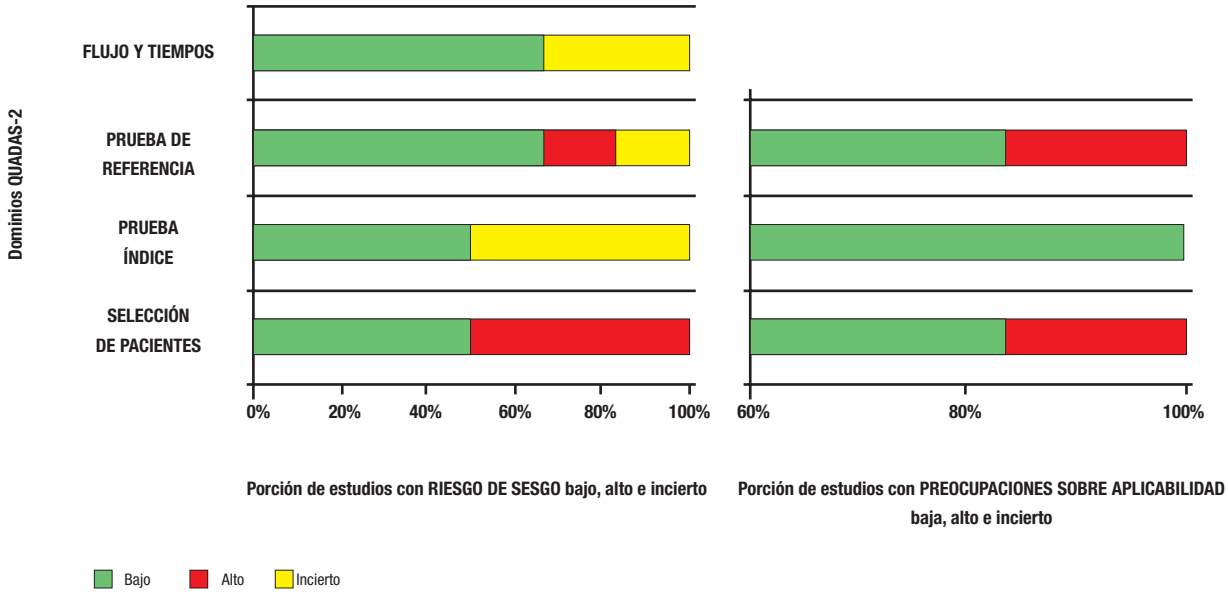
Los revisores pueden optar por restringir el análisis primario, incluyendo sólo los estudios con baja probabilidad de sesgo o

Tabla 1: presentación tabular sugerida para resultados del QUADAS-2

Estudio	Probabilidad de sesgos				Preocupación sobre la aplicabilidad de los resultados		
	Selección de los individuos	Prueba índice	Prueba de referencia	Flujo y tiempos	Selección de los pacientes	Prueba índice	Prueba de referencia
Estudio 1	☺	☺	☺	☺	☹	☺	☺
Estudio 2	☺	☺	☺	☺	☹	☺	☺
Estudio 3	☹	☹	☺	☺	☹	☺	☺
Estudio 4	☹	?	☺	☺	☹	☺	☺

☺ Probabilidad baja. ☹ Probabilidad alta. ? Probabilidad incierta.

Figura 3: Gráfica sugerida para QUADAS-2



que generan poca preocupación respecto de su aplicabilidad, para todos o para determinados dominios.
Si bien podría considerarse la posibilidad de ser más restrictivo en los criterios de inclusión de la revisión incorporando sólo trabajos de un mínimo de calidad en uno o varios dominios, existe consenso de preferir la inclusión de toda la evidencia disponible, y luego comunicar la calidad de todas las investigaciones halladas, investigando luego las posibles razones de la heterogeneidad de los hallazgos.

Los análisis de subgrupos o de sensibilidad permiten evaluar cómo varía la precisión de la prueba índice entre los estudios con diferentes grados de calidad en algunos o en todos los dominios. Por otro lado, los resultados de la evaluación de cada dominio pueden ser utilizados como insumo de análisis de meta-regresión para investigar la asociación de los resultados de cada dominio con los de la precisión estimada.

El sitio web de la Universidad de Bristol dedicado al QUADAS (www.bris.ac.uk/quadas) contiene, además de la herramienta QUADAS-2, un banco de preguntas orientadoras adicionales, una guía más detallada para cada dominio, ejemplos de evaluaciones QUADAS-2 completas y recursos descargables que incluyen: 1) una base de datos para la carga de la información extraída durante el proceso de la revisión; 2) una planilla de cálculo que ayuda a producir representaciones gráficas de los resultados, 3) plantillas de texto para resumir en tablas los resultados de la investigación.

Conclusión

La evaluación cuidadosa de la calidad de los estudios incluidos es esencial para garantizar la calidad de cualquier tipo de revisión sistemática y las que incluyen estudios de precisión diagnóstica no son una excepción.
El proceso riguroso y basado en la evidencia científica utilizado en la construcción de la herramienta QUADAS-2 permitió ofrecer características adicionales y mejoradas con respecto a la ampliamente utilizada QUADAS, incluyendo la distinción entre sesgo y aplicabilidad, la identificación de cuatro dominios clave apoyados por preguntas orientadoras que facilitan la evaluación del riesgo de sesgo, su calificación y las preocupaciones sobre la aplicabilidad incluso para estudios en los que la prueba de referencia se basa en el tiempo seguimiento.
QUADAS-2 constituye una mejora considerable sobre la herramienta original y seguramente contribuirá a desarrollar una base de sólidas evidencias para las pruebas y procedimientos diagnósticos. Sería de gran utilidad realizar comentarios y ofrecer retroalimentación respecto a su uso a través del sitio web QUADAS.

Recibido el 01/08/2014 y aceptado el 01/09/2014

Referencias

1. Whiting P, Rutjes AW, Reitsma JB, y col. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol. Nov 10 2003;3:25.
2. Whiting PF, Rutjes AW, Westwood ME, y col. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. Oct 18 2011;155(8):529-536.
3. Moher D, Schulz KF, Simera I, y col. Guidance for developers of health research reporting guidelines. PLoS Med. Feb 2010;7(2):e1000217.
4. Smith H, Pickering RM, Struthers A, y col. Biochemical diagnosis of ventricular dysfunction in elderly patients in general practice: observational study. BMJ. Apr 1 2000;320(7239):906-908.
5. Whiting PF, Smidt N, Sterne JA, y col. Systematic review: accuracy of anti-citrullinated Peptide antibodies for diagnosing rheumatoid arthritis. Ann Intern Med. Apr 6 2010;152(7):456-464; W155-466.
6. 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.
7. Juni P, Witschi A, Bloch R, y col. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA. Sep 15 1999;282(11):1054-1060.