AMSTAR-2: herramienta de evaluación crítica de revisiones sistemáticas de estudios de intervenciones de salud

AMSTAR-2: critical appraisal tool for systematic reviews of healthcare interventions

Shea BJ y col. BMJ. 2017;358.j4008

Artículo Especial

Resumen

El número de revisiones sistemáticas (RS) de estudios de intervenciones sanitarias publicadas ha aumentado rápidamente. Estos documentos de síntesis son ampliamente utilizados para la toma de decisiones clínicas y de políticas de salud, pero están sujetos a una serie de sesgos y es importante que los usuarios pueden identificar las RS de mejor calidad. Considerando la relevancia de la actualización de la herramienta Ameasurement Tool to Assess Systematic Reviews (AMSTAR) para la valoración crítica de RS de estudios de intervenciones en salud, el autor resume los aspectos más relevantes de la publicación de Shea y col, traduce el instrumento y su guía de aplicación, y comenta los aspectos salientes del instrumento AMSTAR-2 junto con sus potenciales implicancias en el desarrollo y reporte de RS.

Abstract

The number of published systematic reviews (SR) of studies of healthcare interventions has increased rapidly. These synthesis documents are used extensively for clinical and policy decisions, but they are subject to a range of biases and it is important that users can distinguish high quality SR. Taking in consideration the relevance of the updated Ameasurement Tool to Assess Systematic Reviews (AMSTAR) tool for the assessment of SR of healthcare interventions, the author synthetizes the most relevant features of the article by Shea et al, translates the tool and it´s user´s guide, and comments on the AMSTAR-2 highlights and their potential implications in development and report of SR.

Principales características del instrumento AMSTAR-2

Se han diseñados muchos instrumentos para evaluar los diferentes aspectos de una revisión, pero pocos de ellos permiten una evaluación crítica integral[1-15]. La herramienta Ameasurement Tool to Assess Systematic Reviews (AMSTAR) fue desarrollado para evaluar RS de ensayos aleatorizados[16-18]. AMSTAR-2 permite una evaluación más detallada de las RS que incluyen también estudios no aleatorizados de intervenciones sanitarias, que son cada vez más incorporados en las RS. De hecho, casi la mitad de las revisiones sistemáticas publicadas incluyen ahora estudios de intervención no aleatorizados (EINA)[19-21].

AMSTAR-2 es un cuestionario que contiene 16 dominios (ver Apéndice, figura 1), con opciones de respuesta simples: "si", cuando el resultado es positivo; "no", cuando no se cumplió el estándar o hay información insuficiente para responder; y "si parcial", en casos en que hubo adherencia parcial al estándar. La herramienta actualizada incluye una extensa guía para el usuario (ver Apéndice, Guía del usuario). Aunque no proporciona una calificación global, de las debilidades en los siete dominios considerados críticos, dado que pueden afectar sustancialmente la validez de una revisión y sus conclusiones (ver cuadro 1), surgen cuatro niveles de confianza: alta, moderada, baja y críticamente baja (ver tabla 1).

Cabe aclarar que no siempre los elementos enumerados en el Cuadro 1 serán considerados como críticos; por ejemplo, el riesgo de sesgo puede ser considerado menos importante cuando una revisión se limita a ensayos controlados aleatorizados (ECA) de alta calidad. Si no se realizó la síntesis estadística de los resultados mediante el meta-análisis, tampoco se aplicará el dominio sobre métodos meta-analíticos.

Aplicación

Los autores recomiendan que si se empleará AMSTAR-2 para valorar la calidad de una o más RS para basar decisiones importantes a nivel clínico y de políticas sanitarias, el equipo de evaluación acuerde en forma previa ciertas pautas a tener en cuenta al aplicar la herramienta, incluyendo el contexto de la práctica o la política sanitaria y las preguntas clínicas pertinentes (estructura PICO: población, intervención, comparación, resultados). Por ejemplo, las RS disponibles pueden haber incluido estudios con diferentes comparadores o diferentes períodos de seguimiento, por lo que debe establecerse su relevancia para las preguntas importantes relacionadas con la política en cuestión. Las posibles fuentes de sesgo también deben ser acordadas. Por ejemplo, en estudios observacionales de efectos de inter-

Cuadro 1. Dominios críticos de la herramienta AMSTAR-2

1.	Protocolo registrado antes de la revisión (ítem 2)
2.	Adecuada búsqueda en la literatura (ítem 4)
3.	Justificación de los estudios excluidos (ítem 7)
4.	Riesgo de sesgo de los estudios individuales incluidos (ítem 9)
5.	Métodos meta-analíticos apropiados (ítem 11)
6.	Consideración del riesgo de sesgo en la interpretación de los resultados de la revisión (ítem 13)
7.	Evaluación de la presencia y el impacto probable del sesgo de publicación (ítem 15)

Tabla 1. Valoración de la confianza general en los resultados de la revisión

CONFIANZA	JUSTIFICACIÓN
Alta	Ninguna debilidad crítica y hasta una no crítica: la RS proporciona un resumen exacto y completo de los resultados de los estudios disponibles.
Media	Ninguna debilidad crítica y más de una debilidad no crítica (aunque si son muchas podría justificarse una baja confianza): la RS tiene debilidades, pero no hay defectos críticos, pudiendo proporcionar un resumen preciso de los resultados de los estudios disponibles.
Baja	Hasta una debilidad crítica, con o sin puntos débiles no críticos: la RS puede no proporcionar un resumen exacto y completo de los estudios disponibles
Críticamente Baja	Más de una debilidad crítica, con o sin debilidades no críticos: la RS no es confiable

RS: revisión sistemática

venciones, la confusión por indicación (o gravedad de la enfermedad) puede ser problemática cuando las intervenciones se reservan para ciertos subgrupos de pacientes[22]. Asimismo, estos estudios deberían reclutar nuevos usuarios de una tecnología o fármaco para evitar el sesgo de prevalencia[23]. Si el inicio de una intervención tiende a retrasarse, la elección del comparador puede introducir un sesgo de tiempo inmortal[24]. Los errores de medición clasifican erróneamente la exposición y los resultados y pueden desbalancear los grupos de comparación. Una descripción selectiva de múltiples análisis y resultados también pueden desvirtuar los efectos de una intervención.
Si bien el manual del usuario de AMSTAR 2 proporciona orientación sobre las distintas secciones (ver Apéndice, Guía del usuario), algunos de los juicios de los evaluadores pueden ser complejos (p.ej. si los autores de la revisión han evaluado adecuadamente el riesgo de sesgo en EINA) y puede necesitarse asesoramiento tanto sobre metodología como sobre el contenido. El conocimiento de los contenidos puede ser a veces necesario para determinar si los autores de la revisión han hecho una evaluación adecuada de los elementos PICO (dominio 1 del AMSTAR-2), e identificar posibles factores de confusión.

Acuerdo inter-evaluador
Se midió el acuerdo (mediante kappa) entre evaluadores, quienes tuvieron acceso a la guía del usuario. Los valores variaron sustancialmente a través de los dominios y entre los pares de evaluadores, pero la mayoría de los valores se encontraban en un rango aceptable, con 46 de las 50 puntuaciones kappa en rango de acuerdo moderado o mejor, y 39, acuerdo bueno o mejor.

Usabilidad
Los tiempos para completar el AMSTAR-2 oscilaron entre 15 y 32 minutos. Estas estimaciones no incluyen el tiempo necesario para leer la RS. Los comentarios de los evaluadores incluyeron: que se necesita más tiempo para evaluar los análisis de estudios no aleatorizados y mixtos por cuestiones metodológicas de dichos estudios; que era común que los autores de una revisión mencionaran la presencia o ausencia de sesgo de publicación, pero sin aportar ninguna evidencia; y que los autores de la revisión suelen revelar sus posibles conflictos de intereses, pero no la forma en que lo gestionaron.

Discusión
Debe remarcarse que las respuestas al AMSTAR-2 no deben ser utilizadas para obtener una puntuación global,[25,26] y se recomienda que los usuarios adopten el proceso de calificación basado en los dominios críticos (Cuadro 1), o en alguna variación sobre la base de estos principios.
Los estudios no aleatorizados grandes, a menudo realizados con grandes bases de datos administrativas, son cada vez más utilizados para evaluar el verdadero impacto mundial de una amplia gama de tecnologías y prácticas de salud. Aunque estos estudios a menudo usan métodos sofisticados, la confusión residual o el fracaso para manejar adecuadamente otras fuentes de sesgo puede conducir a estimaciones de efecto inexactas. La inclusión de grandes estudios observacionales en los meta-análisis puede generar estimaciones precisas, pero sesgadas de los efectos de las intervenciones evaluadas[19].
Los dominios del AMSTAR-2 provienen de los instrumentos Cochrane de riesgo de sesgo para estudios aleatorizados y no aleatorizados[27,28]. Sin embargo, AMSTAR-2 actualmente no especifica qué instrumentos de riesgo de sesgo deberían haber utilizado los autores de la revisión para evaluar los estudios no aleatorizados incluidos en una RS. El instrumento ROBINS-I, que es la herramienta más completa para los EINA, fue lanzado en 2016, por lo que es poco realista esperar que los autores de las revisiones iniciadas antes de su lanzamiento lo hayan utilizado[28]. Los evaluadores de RS mediante el AMSTAR-2 deberían cerciorarse de que el instrumento para la valoración del riesgo de sesgo utilizado tenga suficiente capacidad discriminatoria para los dominios especificados. Una revisión realizada por Sanderson y col. identificó 86 herramientas para evaluar la calidad de los estudios observacionales, sin arribar a conclusiones acerca de si alguna de ellas debería ser preferida sobre las otras[29]. Los instrumentos populares de evaluación para estudios individuales, tales como la escala Newcastle Ottawa y la lista de verificación SIGN no sólo se centran en la validez y además se utilizan para generar una puntuación global, algo que no se recomienda[30,31].
AMSTAR-2 ofrece una amplia evaluación de la calidad, incluyendo defectos que puedan haber surgido por una mala conducción de la RS (con efectos inciertos en sus resultados). En este sentido, difiere de otro instrumento de riesgo de sesgo en las RS, el ROBIS)[32]. ROBIS es un instrumento trifásico sofisticado que se centra específicamente en el riesgo de sesgo introducido por la realización de la RS. Cubre la mayoría de los tipos de preguntas de investigación, incluyendo diagnóstico, pronóstico y etiología. Por el contrario, AMSTAR-2 está destinado a ser utilizado para RS de intervenciones sanitarias, pero inevitablemente hay una superposición de dominios entre ambas.

Comentario

Aunque no se ha realizado aun una extensa validación de la herramienta AMSTAR-2, ésta sigue los pasos clave en la realización de una RS de manera exhaustiva, reteniendo 10 dominios de la herramienta validada originalmente y adicionando seis más. Además de ser una lista de cotejo para quienes desarrollan o leen una RS, también puede ser muy útil como herramienta de aprendizaje, para valorar uno de los diseños más frecuentemente utilizados en la toma de decisiones basadas en la evidencia. El desarrollo de AMSTAR-2 se basó en gran medida en el consenso de un grupo de expertos, pero también en una amplia retroalimentación de los usuarios del instrumento original y sigue en proceso de refinamiento permanente. De hecho, los autores animan a los usuarios de AMSTAR-2 a proporcionar información si adaptan al instrumento para ámbitos particulares e informar de su experiencia en www.amstar.ca.
En otras herramientas de valoración de RS, incluyendo al AMSTAR original, no se discriminaba la importancia de los ítems, lo que podía generar valoraciones globales engañosas. Una gran fortaleza de esta herramienta es la definición de dominios críticos a partir de los cuales valorar la confianza general en los resultados de la revisión. Podría ser cuestionable que con sólo alguno de los dominios críticos una RS sea considerada de baja confianza. Por ejemplo, la disponibilidad de protocolo, aunque en ascenso, no es tan frecuente, y muchas RS bien conducidas podrían ser clasificadas como de baja confianza. A favor de esta posición es que si no existe protocolo, estamos expuestos a innumerable cantidad de decisiones como por ejemplo reportar selectivamente resultados, o realizar análisis post hoc, que podrían sesgar las conclusiones. De hecho, la ausencia de protocolo suele asociarse a RS de menor calidad metodológica. Por otro lado, en el ítem 15 se exige la realización de pruebas gráficas o estadísticas para evaluar la posibilidad de sesgo de publicación. Aunque deseable, esto no siempre es posible cuando el número de estudios es menor a diez. En estos casos no aplicables, si los revisores discutieron apropiadamente la probabilidad y la magnitud del impacto del sesgo de publicación, este ítem debiera considerarse como válido aun sin gráficas o pruebas.
Actualmente desde Cochrane Argentina estamos realizando una pru eba de la implementación de esta herramienta en español.

Conclusiones del comentador
AMSTAR-2 ha puesto la vara alta y aunque muchas RS serán clasificadas como de baja confianza, es de esperar que su entrada en escena y futura validación, influyan las nuevas RS y sus actualizaciones, que querrán cumplir con el estándar metodológico presentado.