Bond University

Research Repository

AMSTAR 2

A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both

Shea, Beverley J.; Reeves, Barnaby C.; Wells, George; Thuku, Micere; Hamel, Candyce; Moran, Julian; Moher, David; Tugwell, Peter; Welch, Vivian; Kristjansson, Elizabeth; Henry, David A.

Published in:
BMJ (Online)

DOI:
10.1136/bmj.j4008

Licence:
CC BY

Link to output in Bond University research repository.

Recommended citation(APA):
Shea, B. J., Reeves, B. C., Wells, G., Thuku, M., Hamel, C., Moran, J., Moher, D., Tugwell, P., Welch, V., Kristjansson, E., & Henry, D. A. (2017). AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ (Online)*, *358*, Article j4008. https://doi.org/10.1136/bmj.j4008

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

For more information, or if you believe that this document breaches copyright, please contact the Bond University research repository coordinator.

Download date: 12 Nov 2026

OPEN ACCESS

AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both

Beverley J Shea,[1,2,3] Barnaby C Reeves,[4] George Wells,[3,5] Micere Thuku1,[2] Candyce Hamel,[1] Julian Moran,[6] David Moher,[1,3] Peter Tugwell1,[2,3,7] Vivian Welch,[2,3] Elizabeth Kristjansson,[8] David A Henry[9,10,11]

[1]Ottawa Hospital Research Institute, Clinical Epidemiology Program, Ottawa, Canada
[2]Bruyère Research Institute, Ottawa, Canada
[3]School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa, Ottawa, Canada
[4]School of Clinical Sciences, University of Bristol, Bristol, UK
[5]University of Ottawa Heart Institute, Ottawa, Canada
[6]The Hospital for Sick Children, the Genetics and Genome Biology Program, Toronto, Canada
[7]Department of Medicine, The Ottawa Hospital, Ottawa, Canada
[8]Centre for Research in Educational and Community Services, School of Psychology, Faculty of Social Sciences, University of Ottawa, Canada
[9]Centre for Research in Evidence-Based Practice, Bond University, Gold Coast, Australia;
[10]Dalla Lana School of Public Health, University of Toronto, Toronto, Canada
[11]Institute for Clinical Evaluative Sciences, Toronto, Canada

Correspondence to: B J Shea
bevshea@uottawa.ca

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2017;358:j4008
http://dx.doi.org/10.1136/bmj.j4008

Accepted: 4 August 2017

The number of published systematic reviews of studies of healthcare interventions has increased rapidly and these are used extensively for clinical and policy decisions. Systematic reviews are subject to a range of biases and increasingly include non-randomised studies of interventions. It is important that users can distinguish high quality reviews. Many instruments have been designed to evaluate different aspects of reviews, but there are few comprehensive critical appraisal instruments. AMSTAR was developed to evaluate systematic reviews of randomised trials. In this paper, we report on the updating of AMSTAR and its adaptation to enable more detailed assessment of systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. With moves to base more decisions on real world observational evidence we believe that AMSTAR 2 will assist decision makers in the identification of high quality systematic reviews, including those based on non-randomised studies of healthcare interventions.

SUMMARY POINTS

- Systematic reviews of studies of healthcare interventions effects often include non-randomised studies
- AMSTAR is a popular instrument for critically appraising systematic reviews of randomised controlled clinical trials
- AMSTAR underwent further development to enable appraisal of systematic reviews of randomised and non-randomised studies of healthcare interventions
- The revised instrument (AMSTAR 2) retains 10 of the original domains, has 16 items in total (compared with 11 in the original), has simpler response categories than the original AMSTAR, includes a more comprehensive user guide, and has an overall rating based on weaknesses in critical domains
- AMSTAR 2 is not intended to generate an overall score
- With moves to base more decisions on real world observational evidence, AMSTAR 2 should assist in the identification of high quality systematic reviews

With the rapid increase in biomedical publishing, keeping up with primary research has become almost impossible for healthcare practitioners and policy makers.[1] Consequently, healthcare decision makers rely on systematic reviews as one of the key tools for achieving evidence based healthcare.[2] Systematic reviews provide an opportunity to base decisions on accurate, succinct, credible, and comprehensive summaries of the best available evidence on a topic.[2]

Uncritically accepting the results of a single systematic review has risks. One of us (DM) led efforts to improve standards for reporting of systematic reviews, which led to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement.[3] The reporting guide for systematic reviews of observational (non-randomised) studies is MOOSE (Meta-analysis of Observational Studies in Epidemiology).[4] The quality of reporting of a systematic review may, however, more accurately reflect authors' ability to write in a comprehensible manner rather than the way they conducted their review. This underscores the need for guidelines that evaluate the way in which reviews are planned and conducted.[5 6]

The Cochrane Collaboration Handbook provides a comprehensive guide for review authors, but it does not provide a concise critical appraisal instrument for completed reviews.[5] Several instruments have been designed to evaluate individual studies that are being included in systematic reviews or how certain steps (eg, meta-analysis, testing for publication bias) should be conducted.[7-15] But relatively few instruments assess all important steps in the conduct of a review.[16-21]

AMSTAR (A MeaSurement Tool to Assess systematic Reviews), published in 2007, is one of the most widely used instruments.[22-24] AMSTAR was designed by us and our colleagues as a practical critical appraisal tool for use by health professionals and policy makers who do not necessarily have advanced training in epidemiology, to enable them to carry out rapid and reproducible assessments of the quality of conduct of systematic reviews of randomised controlled trials of interventions. Since publication, several critiques

of the instrument have been published.[25-31] These critiques plus feedback received at workshops and developments in the science of systematic reviews pointed to a need to revise and update the original AMSTAR instrument.

Inclusion of non-randomised studies in systematic reviews

Almost half of published systematic reviews now include non-randomised studies of intervention effects.[4 32-34] There are many concerns about the conduct and reporting of systematic reviews of non-randomised studies.[32 35 36] To summarise, non-randomised studies of healthcare interventions (an important focus of this revision of AMSTAR) are subject to a range of biases that are either not present or are less noticeable in randomised controlled trials, thus requiring different risk of bias assessments. Observational studies are increasingly conducted within large population databases, sometimes with hundreds of thousands or even millions of recipients of healthcare interventions. These generate precise estimates of intervention effects, which may be inaccurate because of residual biases. If these estimates are combined with those from the (generally smaller) randomised controlled trials, the meta-estimates will be weighted towards the observational study estimates. The original AMSTAR instrument did not include an assessment of the risk of bias in non-randomised studies included in a review, which is a key issue given the diversity of designs that such studies may use and the biases that may affect them.

Development of AMSTAR 2

The development and validation of the original AMSTAR instrument (published in 2007) has been described in detail elsewhere.[22-24] Briefly, the original list of items was created from the results of a scoping review of the then available rating instruments. This review identified many over-lapping appraisal items, mainly from two extensively cited reports.[16 17] The lists of items from these reports were combined and reduced by factor analysis. After pilot testing, items were reworded as needed and the reliability and usability of the tool was assessed. A modified version was validated externally and performed well against the global judgments of a panel of content experts.[23] The publications describing the original AMSTAR instrument were widely cited and the instrument has been used and critiqued extensively.[22-31]

We convened an expert group, comprising authors of the original instrument, members with expertise in the conduct of non-randomised studies, development of appraisal instruments, biostatistics, and study designs. The expert group met for a day in Ottawa, Canada and members were presented with the results of updated literature reviews on relevant critical appraisal instruments, the results of surveys of AMSTAR users, recorded experience of participants in AMSTAR workshops at Cochrane Colloquiums in 2015 and 2016, feedback from the AMSTAR website (www.amstar.ca), and published critiques of the original instrument.[16-26] The perspective adopted by the expert group was to increase the value of AMSTAR as a broad critical appraisal instrument designed primarily for systematic reviews of studies of healthcare interventions. The expert group considered that revisions should address all aspects of the conduct of a systematic review, and the challenges of including non-randomised studies. They also thought the revised instrument should function as a teaching aid and as a concise checklist for those conducting reviews. The revisions were not intended to deal with the special requirements of diagnostic test reviews, individual patient data meta-analyses or network meta-analyses, scoping reviews, or realist reviews.[37-41]

We used a nominal group technique to propose and then prioritise specific changes to the instrument and to agree on the draft wording of items. Based on their experience of the instrument and the presentations made at the meeting, participants were asked to record their ideas independently and privately. The ideas were then enunciated in a round-robin format. One idea was collected from everyone, in turn, and presented to the group by the facilitator. This process was continued until all ideas had been listed. Individuals then privately recorded their judgments and rankings. These were aggregated statistically to derive the group judgments. The following changes were agreed on (these are not listed in order of priority as all were considered important enough to mandate modifications to the instrument):

- Simplify the response categories
- Align the definition of research questions with the PICO (population, intervention, control group, outcome) framework
- Seek justification for the review authors' selection of different study designs (randomised and non-randomised) for inclusion in systematic reviews
- Seek more details on reasons for exclusion of studies from the review
- Determine whether the review authors had made a sufficiently detailed assessment of risk of bias for the included studies (whether randomised or non-randomised)
- Determine whether risk of bias with included studies was considered adequately during statistical pooling of results (if this was performed)
- Determine whether risk of bias with included studies was considered adequately when interpreting and discussing the review findings.

A description was formulated for each of the draft items. A small subgroup refined the wording of the items and assembled the draft instrument for testing. Initial pilot testing was performed by group members. Draft versions were presented at workshops held at the Cochrane Colloquiums in 2015 and 2016, where feedback directed further modifications and redrafting of the instrument. The version of the instrument presented here was subject to inter-rater reliability and usability testing.

doi: 10.1136/bmj.j4008 | *BMJ* 2017;358:j4008 | thebmj

Comparison with the original instrument

The supplementary figure provides details of the new instrument (AMSTAR 2). Ten domains were retained from the original tool, with changes to the wording of items based on feedback about the original instrument and experience of testing drafts of the new instrument. Two domains were given more detailed coverage in AMSTAR 2 than in the original instrument: duplicate study selection and data extraction now have their own items (they were combined in the original tool). The possible influence of funding sources is now considered separately for individual studies included in the review and for the review itself. Previously they were combined in one item. We added more detailed and separate considerations of risk of bias for randomised and non-randomised studies. Both sub-items are based on content from the Cochrane risk of bias instruments for randomised and non-randomised (ROBINS-I) studies.[42 43] One domain was removed—consideration of grey literature, previously a separate item, is now handled in the item on literature searching.

In total, four domains were added. Two of these came directly from the ROBINS-I tool—namely, elaboration of the PICO and the way in which risk of bias was handled during evidence synthesis.[43] One of the other new domains—discussion of possible causes and significance of heterogeneity—is an elaboration of content in the original AMSTAR tool. Another new domain—justification of selection of study designs—was part of the adaptation of AMSTAR to deal with non-randomised designs.

The domain specific questions in AMSTAR 2 are framed so that a "Yes" answer denotes a positive result. We removed the "not applicable" and "cannot answer" options in the original AMSTAR instrument because we believe that all domains are relevant to contemporary systematic reviews of healthcare interventions. If no information is provided to rate an item, the review authors should not be given the benefit of doubt and the item should be rated as a "No." We have provided a "partial Yes" response in some instances where we considered it worthwhile to identify partial adherence to the standard.

Rationale for selection of items

Here we summarise our thinking behind the items in AMSTAR 2, which are numbered as in the instrument (see supplementary figure). Supplementary appendix 1 provides a more complete user's guide.

1. Did the research questions and inclusion criteria for the review include the components of PICO?

It is common practice to use the PICO description (population, intervention, control group, and outcome) as a convenient and easily memorised framework for a study question. Sometimes a timeframe should be added if this is critical in determining the likelihood of a study capturing relevant clinical outcomes (eg, an effect of the intervention is only expected after several years).

2. Did the report of the review contain an explicit statement that the review methods were established prior to the conduct of the review and did the report justify any significantdeviations from the protocol?

Systematic reviews are a form of observational research, and the methods for the review should be agreed on before the review commences. Adherence to a well developed protocol reduces the risk of bias in the review. Authors should show that they worked with a written protocol with independent verification.

3. Did the review authors explain their selection of the study designs for inclusion in the review?

For some questions, for instance the effects of policy changes, or for ethical reasons, non-randomised studies may be the only studies addressing the review question. With an expansion of AMSTAR 2 to appraise reviews that include randomised controlled trials or non-randomised studies, or both, it is important that authors justify the inclusion of different study designs in systematic reviews. The authors should indicate that they followed a strategy. When both randomised and non-randomised studies address the same question about the effects of an intervention, we believe that authors should consider whether a review that is restricted to randomised controlled trials will give an incomplete summary of the important effects of a treatment.

4. Did the review authors use a comprehensive literature search strategy?

The importance of adequate literature searching in systematic reviews is well established.[5] This item was carried over with minimal changes to the wording from the original instrument. We have made the response options clearer in AMSTAR 2 and provide more detailed guidance on completion of the item, particularly in relation to the identification of non-randomised studies (see supplementary appendix 1).

5. Did the review authors perform study selection in duplicate?

Best practice requires two review authors to determine eligibility of studies for inclusion in systematic reviews.[5] This involves checking the characteristics of a study against the elements of the research question. In the original AMSTAR, this item covered determining both study eligibility and data extraction. The expert group believed that they were sufficiently distinct processes to merit separate items in AMSTAR 2.

6. Did the review authors perform data extraction in duplicate?

The expert group recognised that data extraction might be more complex for non-randomised studies of healthcare interventions as it usually involves extraction of measures of treatment effects and other associations that have been adjusted for potential confounding, rather than raw outcome data from treated and control groups. A study report may present multiple treatment effects; judgment is therefore needed to select the one that conforms best to the PICO question and is at lowest risk from confounding.

7. Did the review authors provide a list of excluded studies and justify the exclusions?
In the revised instrument we consider excluded and included studies separately. Excluded studies should be accounted for fully by review authors, otherwise there is a risk that they remain invisible and the impact of their exclusion from the review is unknown.

8. Did the review authors describe the included studies in adequate detail?
The revised instrument requires review authors to provide detail about research designs, study populations, interventions, comparators, and outcomes. The detail should be sufficient for appraisers to make a judgment about the extent to which the studies were appropriately chosen (in relation to the PICO) and whether the study populations and interventions were relevant to their questions. This information is needed to determine the extent to which the results of different studies should be combined, help explain heterogeneity, and assist those applying the results.

9. Did the review authors use a satisfactory technique for assessing the risk of bias (RoB) in individual studies that were included in the review?
Biases can be introduced at several stages in the design, planning, conduct, and analysis of a study. This item replaces a less detailed item on "scientific quality." The item specifies domains of bias for randomised and non-randomised studies that should have been considered by reviewers, based on the relevant Cochrane instruments.[42 43] In AMSTAR 2 we ask whether the review authors made an adequate assessment of study level efforts to avoid, control, or adjust for baseline confounding, selection biases, bias in measurement of exposures and outcomes, and selective reporting of analyses or outcomes, or both. The guidance document (see supplementary appendix 1) and the ROBINS-I report provide more detail.[43] We decided not to include assessment of time varying confounding, performance biases, and biases due to missing data, although they are currently included in ROBINS-I.[43] This was because of the complex nature of techniques used to adjust for these potential sources of bias and the frequent lack of data (in contemporary primary studies) to enable assessment of these items. Version 2.0 of the Cochrane risk of bias instrument for randomised controlled trials is now available in draft form, and AMSTAR 2 will be aligned with this in the future.[44]

10. Did the review authors report on the sources of funding for the studies included in the review?
We added a consideration of funding sources in the light of evidence from several sources that the results of industry funded studies sometimes favoured sponsored products, and that industry funded studies were less likely to be published than those that were independently funded.[45-47] Such influences may not be detected as flaws in design or methods (item 9).

11. If meta-analysis was performed, did the review authors use appropriate methods for statistical combination of results?
This is a modified version of an item in the original instrument and is judged separately for randomised and non-randomised studies. Review authors should have stated explicitly in the review protocol the principles on which they based their decision to perform meta-analysis of data from the included studies. This includes the extent to which the studies are compatible (in terms of patients, controls, and interventions) and the value of a single pooled effect (for instance from several compatible but underpowered studies). Where reviewers consider it appropriate to conduct a meta-analysis, the inclusion of non-randomised studies increases the complexity of the analyses and may increase heterogeneity (see supplementary appendix 1).

12. If meta-analysis was performed, did the review authors assess the potential impact of RoB in individual studies on the results of the meta-analysis or other evidence synthesis?
This is a new item that requires reviewers to examine how results vary with inclusion or exclusion of primary studies judged to be at high risk of bias. In cases where review authors have chosen to include only high quality randomised controlled trials there may be little discussion of the potential impact of bias on the results. But where they have included randomised controlled trials of variable quality or non-randomised studies they should assess the impact of study level risk of bias on the results of the review.[48]

13. Did the review authors account for RoB in primary studies when interpreting/discussing the results of the review?
This is a modification of an item from the original instrument. With a greater emphasis on assessing risk of bias, the expectation is that reviewers will make explicit reference to the potential impacts of risk of bias when interpreting and discussing the results of their review and in drawing conclusions or making recommendations.

14. Did the review authors provide a satisfactory explanation for, and discussion of, any heterogeneity observed in the results of the review?
This item is carried over with modified wording from the original instrument. It is important that reviewers investigate possible causes of heterogeneity, including variation in those elements included in the PICO framework (see item 1) and those arising from design and methodological considerations (see item 9). With the inclusion of non-randomised studies, variations in design and analysis may contribute to heterogeneity.

15. If they performed quantitative synthesis did the review authors carry out an adequate investigation of publication bias (small study bias) and discuss its likely impact on the results of the review?
This item is carried over from the original instrument but with modified wording. Publication bias is an

doi: 10.1136/bmj.j4008 | *BMJ* 2017;358:j4008 | the bmj

important problem but it can be difficult for authors to resolve completely. Typically, statistical tests (several are available) or graphical displays are used and if the results are positive they indicate the presence of publication bias. Negative test results are not a guarantee of the absence of publication bias as they are insensitive. A minimum of 10 studies are required to show funnel plot asymmetry.[5] The underlying tendency to selectively publish small positive studies may be compounded by the effects of lower methodological quality of small studies, a greater tendency to selectively report results, and increased clinical heterogeneity when conducted in patient subgroups.[49]

16. Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?

This item is carried over with modified wording from the original instrument and is now separate from consideration of funding of the primary studies included in the review (item 10). As with primary studies, review authors should report their funding sources.[50 51]

Identification of critical domains

All steps in the conduct of a systematic review and meta-analysis are important, but we believe that seven domains can critically affect the validity of a review and its conclusions (box 1). Two of these concern risk of bias, whether it has been assessed adequately and how it can influence the results of a review. The prominence we give to risk of bias is because AMSTAR 2 is going to be used to appraise many systematic reviews that include non-randomised studies.

We recognise that the items listed in box 1 will not always be regarded as critical; for example, risk of bias related items may be considered less important when a review is confined to high quality randomised controlled trials. Other circumstances where the critical nature of items may be questioned are when a review team are using meta-analysis to summarise a known literature base (eg, the output from one or more established clinical trial collaborative groups). In this circumstance the adequacy of the literature search (item 4), listing of excluded studies (item 7), and possibility of publication bias (item 15) may not be considered critical. If a meta-analysis was not performed, the item covering the appropriateness of the meta-analytical methods (item 11) will not apply. However, it is important in this circumstance that appraisers are alert to the possible impact of risk of bias when review authors select individual studies to highlight in a narrative summary.

Flaws in the items that we have identified as critical may not be fatal if further information (eg, directly from the review authors) indicates that the original response option was wrong. This may provide reassurance about the review findings or enable an amendment of the review through additional analyses. We emphasise that our listing is a suggestion and appraisers may add or substitute other critical domains. For example, the failure to include non-randomised studies (item 3) in a review of adverse outcomes of treatment may be a critical flaw, as would the inability to explain large variations in treatment effects across a body of studies (item 14).

Box 1: AMSTAR 2 critical domains

- Protocol registered before commencement of the review (item 2)
- Adequacy of the literature search (item 4)
- Justification for excluding individual studies (item 7)
- Risk of bias from individual studies being included in the review (item 9)
- Appropriateness of meta-analytical methods (item 11)
- Consideration of risk of bias when interpreting the results of the review (item 13)
- Assessment of presence and likely impact of publication bias (item 15)

Applying AMSTAR 2 to systematic reviews

If one or more systematic reviews will be the basis of important practice and policy decisions we recommend that the appraisal team agree on how the AMSTAR 2 items should be applied. This includes the practice or policy context and the questions that should be addressed, based on the relevant PICO components. For example, available systematic reviews may have included studies with different comparators or different follow-up times, and their relevance to the policy relevant questions needs to be established. The likely sources of bias should also be agreed on. For instance, in observational studies of intervention effects, confounding by indication (or disease severity) may be problematic when interventions are reserved for certain subgroups of patients.[52] It is good practice to recruit new users of a technology or drug into studies to avoid prevalence bias.[53] If the start of one intervention tends to be delayed the choice of comparator may introduce immortal time bias.[54] Measurement errors can misclassify exposure and outcomes and may be unbalanced across comparison groups. Selective reporting among multiple analyses and outcomes may give an inaccurate measure of intervention effects.

Supplementary appendix 1 provides guidance on sections of AMSTAR 2. Some of the judgments (particularly whether review authors have adequately assessed risk of bias with individual non-randomised studies) are complex, and advice on both methodology and content may be needed. Content knowledge is sometimes necessary to determine if the review authors have made an adequate assessment of the relevant PICO elements (item 1), and to identify potential confounders.

We strongly recommend that individual item ratings are not combined to create an overall score.[55 56] Rather, users should consider the potential impact of an inadequate rating for each item.

In box 2 we propose a scheme for interpreting weaknesses detected in critical and non-critical items. This is advisory and appraisers should decide which items are most important for the reviews under consideration.

Box 2: Rating overall confidence in the results of the review

- **High**
- *No or one non-critical weakness*: the systematic review provides an accurate and comprehensive summary of the results of the available studies that address the question of interest
- **Moderate**
- *More than one non-critical weakness**: the systematic review has more than one weakness but no critical flaws. It may provide an accurate summary of the results of the available studies that were included in the review
- **Low**
- *One critical flaw with or without non-critical weaknesses*: the review has a critical flaw and may not provide an accurate and comprehensive summary of the available studies that address the question of interest
- **Critically low**
- *More than one critical flaw with or without non-critical weaknesses*: the review has more than one critical flaw and should not be relied on to provide an accurate and comprehensive summary of the available studies

*Multiple non-critical weaknesses may diminish confidence in the review and it may be appropriate to move the overall appraisal down from moderate to low confidence.

Inter-rater reliability of AMSTAR 2

We measured inter-rater agreement with three pairs of raters and three sets of systematic reviews (see supplementary appendix 2). The first pair of raters was involved in the development of AMSTAR 2 (coauthors MT and CH). They individually appraised 20 systematic reviews derived from a rapid search (conducted in 2015 on the terms "systematic review" and "meta-analysis" in the title) using Google Scholar. From the first 200 we selected 20 systematic reviews of any healthcare intervention. The other two pairs of raters were experienced in the appraisal of systematic reviews and were not involved in the development of AMSTAR or AMSTAR 2. They applied AMSTAR 2 during their routine work, performing appraisals of systematic reviews of two topics: interventions to reduce medication errors (14 reviews) and non-pharmacological therapies for Parkinson's disease (20 reviews) (see references in supplementary appendix 2). In both cases systematic reviews had been identified through comprehensive literature searches (details available on request). All raters had access to the user guide (see supplementary appendix 1), applied the instrument individually, and did not try to achieve consensus. In total, six raters applied the instrument to 54 systematic reviews, of which 20 included only randomised controlled trials, 18 included only non-randomised studies of interventions, and 16 included a mixture of both designs.

Supplementary appendix 2 provides summaries of the κ scores for agreement between the three pairs of raters across the three sets of reviews. The values varied substantially across items and between pairs of raters. Most values were in an acceptable range, with 46 of the 50 κ scores falling in the range of moderate or better agreement and 39 displaying good or better agreement. There were no large differences between raters, and those who had been involved in the development of AMSTAR 2 did not have higher levels of agreement than the rater who was not involved. Items 9, 12, and 13 are concerned with measurement of risk of bias and how this is handled during discussion of the meta-analysis and interpretation of the results. The ranges of κ scores for these items were similar to those seen with other items in the instrument (see supplementary appendix 2). For items 9 and 11 the κ values for risk of bias judgments for randomised controlled trials were similar to those for non-randomised studies.

Usability of AMSTAR 2

The completion times for the 20 reviews used by reviewers 1 and 2 ranged from 15-32 minutes. These estimates do not include the time taken to read the reviews. This is almost twice the time taken to complete the original AMSTAR instrument (range 10-15 minutes), when it was applied to systematic reviews that were limited to randomised controlled trials.[57] The comments from the reviewers included: that the removal of the "can't answer" and "not applicable" response options in the original instrument forced them to make judgments; that it takes longer to evaluate the non-randomised and mixed study reviews, but this requires the reviewer to confront important methodological issues; that it was common for review authors to mention the presence or absence of publication bias, but not provide any evidence; and that review authors would disclose their potential competing interests but not how they managed them.

Discussion

AMSTAR 2 is a major revision of the original AMSTAR instrument, which was designed to appraise systematic reviews that included randomised controlled trials.[22-24] The main modifications include simplified response categories; a more detailed consideration of risk of bias with included studies, and how this was handled by review authors in summarising and interpreting the results of their reviews; better alignment with the PICO framework for research questions; a more detailed justification of selection of study designs for inclusion in a review; and more information on studies that were excluded from reviews. In addition, we recommend defining critical domains before starting an appraisal of a systematic review. Identification of weaknesses in these domains should undermine confidence in the results of a systematic review.

We stress that responses to AMSTAR 2 items should not be used to derive an overall score.[55 56] The original AMSTAR instrument was often used for this purpose and this was facilitated by the website (www.amstar.ca). We accept that an overall score may disguise critical weaknesses that should diminish confidence in the results of a systematic review and we recommend that users adopt the rating process based on identification of critical domains (see box 2), or some variation based on these principles.[56]

We envisage that AMTAR 2, like its predecessor, may have a role as a convenient teaching aid and as a brief checklist for those conducting systematic reviews. However, we stress that the instrument does not explain in detail the logic and methods of conducting systematic reviews, and those looking for

doi: 10.1136/bmj.j4008 | *BMJ* 2017;358:j4008 | the bmj

comprehensive advice should consult the Cochrane Handbook.[5]

The consideration of risk of bias in individual studies is equally important for randomised and non-randomised studies of healthcare interventions but is generally better understood with the former. Large non-randomised studies, often conducted in large administrative databases, are increasingly being used to assess the real world impact of a wide range of healthcare technologies and practices. Although such studies often use sophisticated methods, residual confounding or failure to deal with other sources of bias may lead to inaccurate estimates of effect. Inclusion of large observational studies in meta-analyses may generate precise but biased estimates of intervention effects.[32]

The items in AMSTAR 2 that deal with risk of bias identify domains specified in the Cochrane risk of bias instruments for randomised and non-randomised studies.[42 43] These represent a consensus, in each case developed with input from more than 30 experts in methodology. However, AMSTAR 2 does not currently specify which risk of bias instruments review authors should have used to assess non-randomised studies included in a systematic review. The ROBINS-I instrument, which is the most comprehensive tool for non-randomised studies evaluating the effects of healthcare interventions, was released in 2016 and it is unrealistic to expect authors of reviews started before its release to have used it.[43] Presently, AMSTAR 2 leaves it to the review authors and those appraising the review to satisfy themselves that the risk of bias instrument used by review authors has sufficient discriminatory ability for the specified risk of bias domains. A review by Sanderson and colleagues identified 86 tools for assessing quality of observational studies, without a clear preference among them.[58] The authors pointed to the need to agree on critical elements for assessing susceptibility to bias in observational epidemiology. In part this review led to the development of ROBINS-I.[43] Popular appraisal instruments for individual studies, such as the Newcastle Ottawa Scale and the Scottish Intercollegiate Guidelines Network (SIGN) checklist may not focus on validity alone.[59 60] The Newcastle Ottawa Scale appears to lack sensitivity and is sometimes used to generate an overall score, something that is not recommended because it may disguise critical weaknesses in a review.[56 61]

AMSTAR 2, as a critical appraisal instrument for systematic reviews, joins several published instruments designed for this purpose.[3 4 16 17 19 20 25 62] Two prominent examples are concerned with guidelines for reporting systematic reviews, rather than their conduct.[3 4] Two highly cited instruments were the basis for the development of the original AMSTAR tool.[16 17 22] Two published instruments are direct derivatives of the original AMSTAR.[19 25] Another publication includes a checklist used to appraise systematic reviews that are being included in an umbrella review.[20] Overlap between the content of this checklist and the original AMSTAR is considerable.[22]

AMSTAR 2 provides a broad assessment of quality, including flaws that may have arisen through poor conduct of the review (with uncertain impact on findings). In this respect it differs from another instrument, the Risk Of Bias In Systematic reviews (ROBIS).[62] ROBIS is a sophisticated three phase instrument that focuses specifically on the risk of bias introduced by the conduct of the review. It covers most types of research question, including diagnosis, prognosis, and aetiology. In contrast, AMSTAR 2 is intended to be used for reviews of healthcare interventions. Inevitably there is overlap in the items considered by ROBIS and AMSTAR 2; indeed, two investigators (BCR, BJS) were involved in the development of both.

In developing AMSTAR 2 we sought to maintain its familiar and popular stepwise checklist approach and augmented this by the addition and modification of items. AMSTAR 2 will be familiar to users of the original instrument, although more demanding to use for reasons discussed previously. Because AMSTAR 2 is structured around the key sequential steps in the conduct of a systematic review, it may be used as a brief teaching aid or as a checklist by those conducting systematic reviews.

Unlike the original instrument, AMSTAR 2 identifies critical weaknesses (see box 1) that should reduce confidence in the findings of a review, and it asks users to prespecify how this list will vary for the review topic. We understand that there will be debate about membership of this list and propose that users may wish to prespecify a different set of critical items for a specific PICO research question or setting.

We did not perform an extensive validation of the revised AMSTAR 2 tool. In its development, 10 domains were retained from the original validated tool, albeit with some wording changes based on feedback and extensive experience of using it. Two domains were given more detailed coverage: duplicate study selection and data extraction now have their own items (they were combined in the original tool); we have added more detailed, and separate, considerations of risk of bias for randomised and non-randomised studies. The sub-items were derived from widely used Cochrane instruments. One domain was removed; consideration of grey literature, previously a separate item, is now handled in the item on literature searching. In total, four domains were added. Two of these come directly from the ROBINS-I tool—namely, elaboration of PICO in the review and the way in which risk of bias was handled during evidence synthesis.[43] One of the other new domains, discussion of possible causes and importance of heterogeneity, is elaboration of content in the original AMSTAR tool.[22] The final domain, justification of selection of study designs, is justified by adapting AMSTAR to deal with non-randomised designs. We do not think this needs validation because we believe it is obvious that authors of systematic reviews should justify why they have included study designs that are more susceptible to bias.

The levels of agreement achieved by the three pairs of raters varied across items, but they were moderate to substantial for most items. Notably, the agreement between two raters involved in the development of AMSTAR 2 was no higher than that achieved by experienced raters who had not been involved its development. We did not expect perfect agreement, and differences between raters reflect the demanding nature of some item level judgments and should prompt group discussion of their causes and importance, and, if needed, consultation with experts in subject matter and methods.

In developing AMSTAR 2 we relied heavily on the consensus of the expert panel, but we also received extensive feedback from users of the original instrument in the form of direct communications, website comments, and evaluations made at teaching workshops at Cochrane Colloquiums. In the later phases of development of AMSTAR 2 we had access to, and discussed, recently published critiques of AMSTAR.[25-31]

Our experience of releasing and using the original AMSTAR instrument is that judgments need to be made and users may sometimes decide to make modifications to the instrument.[25 26 30] We encourage investigators to provide feedback, and, if they adapt the instrument for particular settings, to report their experience at www.amstar.ca.

We thank for their assistance in the development of the AMSTAR critical appraisal instruments: Lex Bouter, Maarten Boers, Alonso Carrasco-Labra, Jeremy Grimshaw, Ranjeeta Mallick, Jordi Pardo-Pardo, and Larissa Shamseer, and for conducting the reliability studies reported in this manuscript: Brian Hutton, Pauline Barbeau, Fatemeh Yazdi, Vesa Basha, and Roxanne Ward.

Contributors: BJS, DAH, GW, and PT conceived the project. BJS and DAH oversaw the project. BJS and DAH led the working group. All authors contributed to the development of AMSTAR 2 and to writing associated guidance. BJS, DAH, BCR, and PT led the drafting and redrafting of the manuscript. All other authors reviewed and commented on drafts of the manuscript. BJS and DAH are the guarantors

Funding: This work was supported by an operating grant from the Canadian Institutes for Health Research (grant No MOP-130470). BCR is supported in part by the UK National Institute for Health Research Bristol Cardiovascular Biomedical Research Unit.

Competing interests: All authors have completed the ICMJE uniform disclosure form at http://www.icmje.org/coi_disclosure.pdf and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years, no other relationships or activities that could appear to have influenced the submitted work.

Provenance and peer review: Not commissioned; externally peer reviewed.

This is an Open Access article distributed in accordance with the terms of the Creative Commons Attribution (CC BY 4.0) license, which permits others to distribute, remix, adapt and build upon this work, for commercial use, provided the original work is properly cited. See: http://creativecommons.org/licenses/by/4.0/.

1 Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS Med* 2010;7:e1000326. doi:10.1371/journal.pmed.1000326
2 Mulrow CD. Rationale for systematic reviews. *BMJ* 1994;309:597-9. doi:10.1136/bmj.309.6954.597
3 Moher D. Alessandro Liberati, Tetzlaff J, Altman DG, and the PRISMA Group. Preferred Reporting Items for (SR) and Meta-Analyses: The PRISMA Statement. *Ann Intern Med* 2009;6:264-9. doi:10.7326/0003-4819-151-4-200908180-00135
4 Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. *JAMA* 2000;283:2008-12. doi:10.1001/jama.283.15.2008
5 Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. www.handbook.cochrane.org.
6 Dechartres A, Charles P, Hopewell S, Ravaud P, Altman DG. Reviews assessing the quality or the reporting of randomized controlled trials are increasing over time but raised questions about how quality is assessed. *J Clin Epidemiol* 2011;64:136-44.
7 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. doi:10.1186/1471-2288-3-25
8 Wong WC, Cheung CS, Hart GJ. Development of a quality assessment tool for systematic reviews of observational studies (QATSO) of HIV prevalence in men having sex with men and associated risk behaviours. *Emerg Themes Epidemiol* 2008;5:23. doi:10.1186/1742-7622-5-23
9 Verhagen AP, de Vet HC, de Bie RA, et al. The Delphi list: a criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. *J Clin Epidemiol* 1998;51:1235-41. doi:10.1016/S0895-4356(98)00131-0
10 Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J Epidemiol Community Health* 1998;52:377-84. doi:10.1136/jech.52.6.377
11 Murray J, Farrington DP, Eisner MP. Drawing conclusions about causes from systematic reviews of risk factors: The Cambridge Quality Checklists. *J Exp Criminol* 2009;5:1-23doi:10.1007/s11292-008-9066-0.
12 Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res* 2012;21:651-7. doi:10.1007/s11136-011-9960-1
13 Bérard A, Andreu N, Tétrault J, Niyonsenga T, Myhal D. Reliability of Chalmers' scale to assess quality in meta-analyses on pharmacological treatments for osteoporosis. *Ann Epidemiol* 2000;10:498-503. doi:10.1016/S1047-2797(00)00069-7
14 Thompson S, Ekelund U, Jebb S, et al. A proposed method of bias adjustment for meta-analyses of published observational studies. *Int J Epidemiol* 2011;40:765-77. doi:10.1093/ije/dyq248
15 Deeks JJ, Altman DG, Bradburn MJ. Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. Systematic reviews in health care: meta-analysis in context. Wiley & Sons, 2008;285-312.
16 Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. *N Engl J Med* 1987;316:450-5.
17 Oxman AD, Guyatt GH. Validation of an index of the quality of review articles. *J Clin Epidemiol* 1991;44:1271-8. doi:10.1016/0895-4356(91)90160-B
18 Oxman AD, Cook DJ, Guyatt GH. Evidence-Based Medicine Working Group. Users' guides to the medical literature. VI. How to use an overview. *JAMA* 1994;272:1367-71. doi:10.1001/jama.1994.03520170077040
19 Scottish Intercollegiate Guidelines Network. Methodology Checklist 1: Systematic Reviews and Meta-analyses. www.sign.ac.uk/checklists-and-notes.html
20 Aromataris E, Fernandez R, Godfrey CM, Holly C, Khalil H, Tungpunkom P. Summarizing systematic reviews: methodological development, conduct and reporting of an umbrella review approach. *Int J Evid Based Healthc* 2015;13:132-40. doi:10.1097/XEB.0000000000000055
21 Whiting P, Savović J, Higgins JP, et al. ROBIS group. ROBIS: A new tool to assess risk of bias in systematic reviews was developed. *J Clin Epidemiol* 2016;69:225-34. doi:10.1016/j.jclinepi.2015.06.005
22 Shea BJ, Grimshaw JM, Wells GA, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Med Res Methodol* 2007;7:10. doi:10.1186/1471-2288-7-10
23 Shea BJ, Hamel C, Wells GA, et al. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *J Clin Epidemiol* 2009;62:1013-20. doi:10.1016/j.jclinepi.2008.10.009
24 Shea BJ, Bouter LM, Peterson J, et al. External validation of a measurement tool to assess systematic reviews (AMSTAR). *PLoS One* 2007;2:e1350. doi:10.1371/journal.pone.0001350
25 Kung J, Chiappelli F, Cajulis OO, et al. From systematic reviews to clinical recommendations for evidence-based health care: validation of revised assessment of multiple systematic reviews (R-AMSTAR) for grading of clinical relevance. *Open Dent J* 2010;4:84-91.

26 Pieper D, Buechter RB, Li L, Prediger B, Eikermann M. Systematic review found AMSTAR, but not R(evised)-AMSTAR, to have good measurement properties. *J Clin Epidemiol* 2015;68:574-83. doi:10.1016/j.jclinepi.2014.12.009
27 Faggion CMJr. Critical appraisal of AMSTAR: challenges, limitations, and potential solutions from the perspective of an assessor. *BMC Med Res Methodol* 2015;15:63.
28 Teich ST, Heima M, Lang L. Dental Students' Use of AMSTAR to Critically Appraise Systematic Reviews. *J Dent Educ* 2015;79:1031-9.
29 Burda BU, Holmer HK, Norris SL. Limitations of a measurement tool to assess systematic reviews (amstar) and suggestions for improvement. *Syst Rev* 2016;5:58.
30 Wegewitz U, Weikert B, Fishta A, Jacobs A, Pieper D. Resuming the discussion of AMSTAR: What can (should) be made better?*BMC Med Res Methodol* 2016;16:111. doi:10.1186/s12874-016-0183-6
31 Dahm P. Raising the bar for systematic reviews with Assessment of Multiple Systematic Reviews (AMSTAR). *BJU Int* 2017;119:193. doi:10.1111/bju.13754
32 Egger M, Schneider M, Davey Smith G. Spurious precision? Meta-analysis of observational studies. *BMJ* 1998;316:140-4. doi:10.1136/bmj.316.7125.140
33 Renehan AG, Tyson M, Egger M, Heller RF, Zwahlen M. Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies. *Lancet* 2008;371:569-78. doi:10.1016/S0140-6736(08)60269-X
34 Page MJ, Shamseer L, Altman DG, et al. Epidemiology and reporting characteristics of systematic reviews of biomedical research: a cross-sectional study. *PLoS Med* 2016;13:e1002028.
35 Shapiro S. Meta-analysis/Shmeta-analysis. *Am J Epidemiol* 1994;140:771-8. doi:10.1093/oxfordjournals.aje.a117324
36 Fleiss JL, Gross AJ. Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: a critique. *J Clin Epidemiol* 1991;44:127-39. doi:10.1016/0895-4356(91)90261-7
37 Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM. Cochrane Diagnostic Test Accuracy Working Group. Systematic reviews of diagnostic test accuracy. *Ann Intern Med* 2008;149:889-97. doi:10.7326/0003-4819-149-12-200812160-00008
38 Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ* 2010;340:c221. doi:10.1136/bmj.c221
39 Lumley T. Network meta-analysis for indirect treatment comparisons. *Stat Med* 2002;21:2313-24. doi:10.1002/sim.1201
40 Pawson R, Greenhalgh T, Harvey G, Walshe K. Realist review--a new method of systematic review designed for complex policy interventions. *J Health Serv Res Policy* 2005;10(Suppl 1):21-34. doi:10.1258/1355819054308530
41 Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol* 2005;8:19-32. doi:10.1080/1364557032000119616.
42 Higgins JP, Altman DG, Gøtzsche PC, et al. Cochrane Bias Methods Group. Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomized trials. *BMJ* 2011;343:d5928. doi:10.1136/bmj.d5928
43 Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 2016;355:i4919. doi:10.1136/bmj.i4919
44 National Collaborating Centre for Methods and Tools (2017). Appraising the risk of bias in randomized trials using the Cochrane Risk of Bias Tool. Hamilton, ON: McMaster University. (Updated 1 September, 2017) http://www.nccmt.ca/resources/search/280. (accessed 7 Sept, 2017).
45 DeAngelis CD, Fontanarosa PB. Impugning the integrity of medical science: the adverse effects of industry influence. *JAMA* 2008;299:1833-5. doi:10.1001/jama.299.15.1833
46 Lexchin J, Bero LA, Djulbegovic B, Clark O. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. *BMJ* 2003;326:1167-70. doi:10.1136/bmj.326.7400.1167
47 Yaphe J, Edman R, Knishkowy B, Herman J. The association between funding by commercial interests and study outcome in randomized controlled drug trials. *Fam Pract* 2001;18:565-8. doi:10.1093/fampra/18.6.565
48 Bilandzic A, Fitzpatrick T, Rosella L, Henry D. Risk of Bias in Systematic Reviews of Non-Randomized Studies of Adverse Cardiovascular Effects of Thiazolidinediones and Cyclooxygenase-2 Inhibitors: Application of a New Cochrane Risk of Bias Tool. *PLoS Med* 2016;13:e1001987. doi:10.1371/journal.pmed.1001987
49 Nüesch E, Trelle S, Reichenbach S, et al. Small study effects in meta-analyses of osteoarthritis trials: meta-epidemiological study. *BMJ* 2010;341:c3515. doi:10.1136/bmj.c3515
50 Bero L, El-Hachem P, Abou-Haidar H, Neumann I, Schünemann HJ, Guyatt GH. What is in a name? Nonfinancial influences on the outcomes of systematic reviews and guidelines. *J Clin Epidemiol* 2014;67:1239-41. doi:10.1016/j.jclinepi.2014.06.015
51 Committee on Publication Ethics. Code of conduct and best practice guidelines for journal editors. https://publicationethics.org/files/Code%20of%20Conduct_2.pdf
52 Salas M, Hofman A, Stricker BH. Confounding by indication: an example of variation in the use of epidemiologic terminology. *Am J Epidemiol* 1999;149:981-3. doi:10.1093/oxfordjournals.aje.a009758
53 Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. *Am J Epidemiol* 2003;158:915-20. doi:10.1093/aje/kwg231
54 Suissa S. Immortal time bias in pharmaco-epidemiology. *Am J Epidemiol* 2008;167:492-9. doi:10.1093/aje/kwm324
55 Greenland S, O'Rourke K. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. *Biostatistics* 2001;2:463-71. doi:10.1093/biostatistics/2.4.463
56 Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* 1999;282:1054-60. doi:10.1001/jama.282.11.1054
57 Shea, BJ Assessing the Methodological Quality of Systematic Reviews: The Development of AMSTAR, PhD, Vrije Universiteit Amsterdam 2008: page 70. https://research.vu.nl/ws/portalfiles/portal/2927023
58 Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. *Int J Epidemiol* 2007;36:666-76. doi:10.1093/ije/dym018
59 Wells GA, Shea B, O'Connell D, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of observational studies in meta-analyses. www.ohri.ca/programs/clinical_epidemiology/oxford.asp
60 Scottish Intercollegiate Guidelines Network SIGN 50: Methodology Checklist 3: Cohort Studies. www.sign.ac.uk/checklists-and-notes.html
61 McGettigan P, Henry D. Cardiovascular risk with non-steroidal anti-inflammatory drugs: systematic review of population-based controlled observational studies. *PLoS Med* 2011;8:e1001098. doi:10.1371/journal.pmed.1001098
62 Whiting P, Savović J, Higgins JP, et al. ROBIS group. ROBIS: A new tool to assess risk of bias in systematic reviews was developed. *J Clin Epidemiol* 2016;69:225-34.

Supplementary figure: the AMSTAR 2 instrument
Supplementary appendix 1: AMSTAR 2 guidance document
Supplementary appendix 2: Inter-rater reliability of AMSTAR 2 items and references for studies included in the analyses

No commercial reuse: See rights and reprints http://www.bmj.com/permissions Subscribe: http://www.bmj.com/subscribe