Evidence to inform the development of ROBIS, a new tool to assess the risk of bias in systematic reviews

17 September 2013

Penny Whiting, Philippa Davies, Jelena Savović,

Deborah Caldwell, Rachel Churchill

School of Social and Community Medicine, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol, BS8 2PS, United Kingdom

Please cite as:

Whiting P, Davies P, Savović J, Caldwell D, Churchill R. (September 2013) Evidence to inform the development of ROBIS, a new tool to assess the risk of bias in systematic reviews, Available from http://www.robis-tool.info [accessed {date}]

Contents

Tables

Figures

Chapter 1. Background

Systematic reviews are generally considered to provide the most reliable form of evidence for the effects of an intervention.[2] Systematic reviews can be used to address questions on any topic using studies of any design. For example, systematic reviews of the effects of anti-hypertensive agents on blood pressure may include only randomised controlled trials (RCTs). A review on the effects of dietary factors on cancer risk could include both RCTs and observational studies, while a study on the accuracy of biomarkers for renal failure would include diagnostic test accuracy studies. Despite possible differences in objectives and inclusion criteria, all systematic reviews should follow the same basic methodological approach and reporting structure. This includes pre-defining a set of objectives and inclusion criteria, using explicit and reproducible methodology, undertaking comprehensive searches that aim to identify all relevant studies, assessing the quality of included studies, and using a standardised presentation and synthesis of the characteristics and findings of the included studies.[2] As with any study, systematic flaws or limitations in the design or conduct of the review have the potential to bias results. There is potential for bias to arise at all stages of the review process and it is important to consider whether these potential biases result from flaws in the design and conduct of the review when interpreting the results.

The impact of potential flaws in the design and conduct of systematic reviews are becoming better understood. Following the development and adoption of PRISMA[3] producers of systematic reviews such as the Cochrane Collaboration are now becoming more focused on trying to prevent potential biases in their reviews by developing explicit expectations for conduct and reporting. For example, the MECIR guidelines for conduct[4] have been formally adopted by the Cochrane Collaboration. Many items on this list are intended to avoid flaws in review design and conduct that may be associated with potential biases.

There are a number of tools available for undertaking critical appraisal and quality assessment of systematic reviews. Although none have become universally accepted, the most commonly used quality assessment tool is probably the AMSTAR tool.[5] This tool was developed by Beverley Shea and colleagues in 2007. It was systematically developed and

has undergone formal evaluation. [6] It consists of 11 items related to reporting and a number of potential biases, each of which is rated as "yes", "no", "can't answer" or "not applicable". Advances in quality assessment of primary studies included in systematic reviews have moved away from generic quality checklists to a more domain based assessment of the risk of bias (in some cases combined with questions about applicability). The Cochrane Risk of Bias tool[7] was published in 2011 and is designed to assess the risk of bias in RCTs. It includes seven domains, each of which are assessed in terms of the risk of bias with studies rated as high, low or unclear risk of bias for each domain. Work is underway to develop a similar tool to assess the risk of bias in observational studies.[8] For assessing the quality of diagnostic accuracy studies included in systematic reviews, the QUADAS tool, published in 2003[9, 10] was similar in structure to the AMSTAR tool; it consisted of 14 items each of which were rated as "yes", "no" or "unclear". Items were not restricted to risk of bias but also concerned reporting quality and variability across studies. An update to the original QUADAS tool, QUADAS-2, has recently been published.[11] The revised tool is structured as four key domains, each of which is assessed in terms of the risk of bias and applicability of the primary studies to the review question. Items relating to reporting quality have been removed. There is currently no tool available specifically to assess the risk of bias in systematic reviews; all currently available tools have a broader objective of critical appraisal or focus specifically on meta-analyses. This project aims to develop a new tool to assess the risk of bias in systematic reviews, the "ROBIS tool".

Chapter 2. Approach and Scope of ROBIS

Key points

There is a need for a tool to assess the risk of bias in systematic reviews; no tool with this specific aim currently exists.

We suggest adapting an approach proposed by Moher[12] for guideline development, including a face-to-face meeting, to develop ROBIS. This is similar to the approach used to develop QUADAS-2.

We have used a three-phased approach to inform the development of ROBIS:

1. Classification of MECIR standards (Chapter 3)
2. Review of existing quality assessment tools (Chapter 4)
3. Review of studies that have used the AMSTAR tool (Chapter 5)

Conceptual decisions

- Bias should focus on internal validity only "a systematic error or deviation from the truth, in the summary estimates and/or review conclusions"
- Domain based structure supported by signalling questions, similar to QUADAS-2
- Domains rated as high/low/unclear risk of bias
- Signalling questions rated as "yes/no/unclear" or "yes/probably yes/probably no/no/no information"
- Striving for comprehensive tool, avoiding overlap between items

2.1 Introduction to the ROBIS project

The ROBIS project aims to develop a new tool to assess the risk of bias in systematic reviews. We have selected an approach similar to that used to develop QUADAS-2 for the development of ROBIS.[11] This is based on methods for guideline development proposed by Moher et al and involves a series of steps (Table 1).[12] The main focus will be a face-to-face group meeting of experts in the area of systematic reviews. This report summarises the pre-meeting activities in particular items 2 to 4 from Table 1 – rationale and scope of ROBIS, development of the evidence base, and generation of a list of items for consideration. The ROBIS initiative is funded by a grant from the UK Medical Research Council (MRC) and National Institute for Health Research (NIHR) joint Methodology Research Programme. The project is led by Penny Whiting (Kleijnen Systematic Reviews and University of Bristol) and Rachel Churchill (University of Bristol). The project team also includes Jelena Savovic, Philippa Davies, and Deborah Caldwell (University of Bristol). They work closely with the steering group (Appendix 1) who provide advice on the project methods and conceptual decisions. The face-to-face meeting is scheduled to be held in Quebec, alongside the Cochrane Colloquium in September 2013. It will include experts in the area of systematic reviews (Appendix 2) who have been invited to give a spread of expertise across review methods (e.g. searching, synthesis) and review topic areas (e.g. RCTs, diagnostic reviews, prognostic reviews). The project started in May 2013 and is funded for 9 months; we hope to have a final draft of ROBIS available by January 2014.

2.1 Rationale for ROBIS

Although there are a number of tools available for undertaking critical appraisal and quality assessment of systematic reviews none focus specifically on the assessment of risk of bias within a systematic review. Most, including AMSTAR, the most commonly used tool, follow a simple checklist approach. Advances in the area of quality assessment of primary studies included in systematic reviews have moved away from generic quality checklists, to a more domain based assessment of the risk of bias, in some cases combined with questions about applicability. The Cochrane Risk of Bias tool[7] and QUADAS-2 are examples of domain based tools and work is underway to develop a similar tool to assess the risk of bias in observational studies.[8]

Table 1: Proposed stages for the development of ROBIS: adapted from Moher et al. "Reporting Guidance to Developers of Health Reporting Guidelines"

Pre-meeting activities	
Item #	
1	Funding the ROBIS initiative
2	Rationale and scope of ROBIS (Chapter 2)
3	Develop the evidence base - Phase 1: Classification of MECIR items (Chapter 3) - Phase 2: Review of existing quality assessment tools for systematic reviews (Chapter 4) - Phase 3: Review of studies that have used the AMSTAR tool (Chapter 5)
4	Generating a list of items for consideration (Chapter 6)
5	Organization and logistics of ROBIS development meeting
5a	Identify group members
5b	Decide size and duration of the meeting
5c	Book the meeting venue
5d	Develop meeting logistics
5e	Develop meeting agenda (Chapter 7)
5f	Prepare materials to be sent to participants prior to meeting (This report)
5g	Arrange to record the meeting
Face-to-face meeting activities	
7	Present and discuss results of pre-meeting activities and relevant evidence
8	Discuss the rationale for including items in the checklist
9	Generate items for inclusion in checklist
11	Discuss strategy for producing documents; identify who will be involved in which activities; discuss authorship
12	Discuss knowledge translation strategy
Post-meeting activities	
13	Develop ROBIS
14	Pilot ROBIS
15	Develop background document
16	Develop a publication strategy
Post-publication activities	
18	Seeking and dealing with feedback and criticism
20	Website development

2.2 Scope of ROBIS

During their call on 23rd June, the steering group agreed on the following conceptual decisions:

Definition of risk of bias in systematic reviews used for the ROBIS project

For the purposes of the ROBIS tool, bias is defined as the risk of "a systematic error or deviation from the truth, in the summary estimates and/or review conclusions" and is therefore related only to the internal validity of the review. ROBIS will not consider applicability.

Preliminary conceptual decisions taken by the ROBIS steering group

ROBIS will have the following general applications:

- Allow those conducting overviews of systematic reviews to assess the risk of bias in included studies
- Allow consistent and reliable assessment of risk of bias by reviewers with different backgrounds
- Distinguish between reviews at high and low risk of bias

Structure

The ROBIS tool must be relatively short and straightforward to complete. We propose adopting a domain based structure similar to those used in Cochrane Risk of Bias tools and QUADAS-2. Signalling questions will be included to help judge the risk of bias; these questions flag aspects of study design related to the potential for bias and aim to help reviewers judge risk of bias. ROBIS should not incorporate a summary quality score.

Rating

We propose a three phased approach to scoring risk of bias: information used to support the judgment of risk of bias, signalling questions, and judgment of risk of bias. By recording the information used to reach the judgment (*support for judgment*), we aim to make the rating transparent and facilitate discussion among review authors independently completing

assessments. Signalling questions could be answered as "yes," "no," or "unclear" and should be phrased such that "yes" indicates low risk of bias. Alternative ratings systems could be discussed, for example a modification to this system is currently being developed for non-randomised studies so that items are rated as "yes", "probably yes", "probably no", "no", "no information". It was agreed to discuss a move to such a rating system at the face-to-face meeting.

We suggest that risk of bias is judged as "low," "high," or "unclear", as in other similar tools (e.g. Cochrane Risk of Bias tool[7], QUADAS-2[11] and PROBAST (Whiting 2013, personal communication)). If the answers to all signalling questions for a domain are "yes," then risk of bias can be judged low. If any signalling question is answered "no," potential for bias exists. Review authors must then use guidance that they have produced specific to their review to judge risk of bias. The "unclear" category should be used only when insufficient data are reported to permit a judgment.

Comprehensive nature of the tool

When developing ROBIS we need to aim to develop a set of independent criteria that work together, i.e. to ensure that there is no overlap between items.

2.3 Develop the evidence base

We used a three phased approach to provide the evidence to inform the development ROBIS. The results of each of these phases are summarised in the report to facilitate discussion at the face to face meeting.

Phase 1: Classification of Methodological Expectations for Cochrane Intervention Reviews (MECIR) items (Chapter 3)

We reviewed the 80 MECIR conduct items[4] and classified each item as relating to risk of bias, variability/applicability, the reporting quality, or as being a "process" item (i.e. items relating to how the review should be conducted from a practical perspective). For each bias item we developed a suggested "signalling question". This review aimed to identify possible signalling questions for inclusion in ROBIS.

Phase 2: Review of existing quality assessment tools for systematic reviews (Chapter 4)

We reviewed 40 existing tools designed to assess the quality of systematic reviews or meta-analyses. We classified items included in the tool according to 5 areas of bias within systematic reviews (question/inclusion criteria, search, review process, synthesis and conclusions). We also discussed details on tool development, tool structure, item rating, and inter-rater reliability. This review also aimed to identify possible signalling questions for inclusion in ROBIS.

Phase 3: Review of studies that have used the AMSTAR tool (Chapter 5)

We conducted a review of overviews that have used the AMSTAR tool to assess the quality of included systematic reviews. The aim of this review was to provide information on the requirements of users of ROBIS.

2.4 Generate a list of items for consideration for inclusion in ROBIS (Chapter 6)

Based on the results of the three review phases, we identified possible items for inclusion in ROBIS and summarised information on the requirements of ROBIS.

2.5 Face-to-Face meeting of the ROBIS Group

We will hold a one-day face to face meeting to develop a first draft of the ROBIS tool. A group of around 25 methodological experts and reviewers working on systematic reviews have been invited to participate in this meeting and received the evidence report prior to the meeting. During the meeting we will present summaries of the evidence identified. Groups of 4 to 6 participants will discuss the proposed scope of the tool, the domains to be covered by ROBIS and signalling questions within domains. Based on meeting discussion and feedback, the project leads will produce a first draft working version of the ROBIS tool. This will be agreed with the steering group before being circulated to meeting participants.

2.6 Piloting and refinement of the ROBIS tool

Using a modified Delphi method, we will use multiple rounds of piloting to refine successively amended versions of the ROBIS tool. Online questionnaires will be developed

to gather structured feedback for each round. Other forms of feedback (e.g. via e-mail or verbal discussion) will also be accepted. The participants in the face to face meeting (the "ROBIS group") will have the opportunity to comment on all drafts of the ROBIS tool. In addition, we will hold workshops at relevant conferences where we will present the ROBIS tool and give participants the opportunity to pilot the tool and provide feedback. Pairs of reviewers working on the BEST project (Best Evidence Summaries of Topics in Mental Health; http://ccdan.cochrane.org/best-mental-health) will pilot a draft version of the tool on a number of reviews. This will provide data on inter-rater reliability. Once sufficient agreement has been reached, a final version of the tool will be agreed. A background document providing guidance on how to apply the tool will also be developed.

Chapter 3. Phase 1: Classification of MECIR items

Key points

80 MECIR items were coded to identify those relating to risk of bias

- 46 items were classified as relating to sources of bias
- We grouped items into the following five domains: Selection, Searching, Review process, Synthesis, Conclusions
- We proposed signalling questions for each "bias" item

3.1 Objective

To identify Cochrane MECIR conduct items relating to potential sources of bias in systematic reviews and to use these to develop signalling questions for possible inclusion in ROBIS.

3.2 Methods

Cochrane recently completed the MECIR project to establish conduct and reporting standards for Cochrane Reviews. [4] A list of 80 items describing the methodological expectations for the conduct of Cochrane intervention reviews was produced and has been adopted by the Cochrane Collaboration. We reviewed the items recommended by MECIR and classified each item as relating to risk of bias, variability/applicability of the review, or as being a "process" item (i.e. items relating to how the review should be conducted from a practical perspective). This was done independently by two reviewers with disagreements resolved through discussion. For items relating to risk of bias, we then proposed a "signalling question" for possible inclusion in ROBIS. Signalling questions were phrased so that they covered a single item, could be answered as "yes", "no" or "unclear" and so that "yes" indicated absence of bias. We grouped the items into the following five domains:

1. Selection
2. Searching
3. Review process
4. Synthesis
5. Conclusions

3.3 Results

The classification of each individual MECIR item is summarised in Appendix 4. Of the 80 MECIR conduct items, we considered 46 to relate to risk of bias. Items which we considered to be associated with a risk of bias in a systematic review are summarised in Table 1, together with a suggested signalling question for each item.

3.4 Summary

We classified 46 of the 80 MECIR items as relating to risk of bias. We proposed signalling questions for each "bias" item and grouped items into the following five domains: Selection, Searching, Review process, Synthesis, Conclusions.

Table 2: MECIR items classed as "bias" with suggested signalling questions

Item name	Standard	Possible signalling question for bias items
SELECTION		
1. Formulating review questions	Ensure that the review question and particularly the outcomes of interest, address issues that are important to stakeholders such as consumers, health professionals and policy makers.	Were review objectives clearly specified?
2. Pre-defining objectives	Define in advance the objectives of the review, including participants, interventions, comparators and outcomes.	
3. Pre-defining unambiguous criteria for participants	Define in advance the eligibility criteria for participants in the studies.	Were inclusion criteria clearly defined?
4. Pre-defining a strategy for studies with a subset of eligible participants	Define in advance how studies that include only a subset of relevant participants will be handled.	Were criteria for handling studies that include only a subset of relevant participants specified?
5. Pre-defining unambiguous criteria for interventions and comparators	Define in advance the eligible interventions and the interventions against which these can be compared in the included studies.	Was ambiguity in inclusion criteria for interventions and comparators avoided?
6. Clarifying role of outcomes	Clarify in advance whether outcomes listed under 'Criteria for considering studies for this review' are used as criteria for including studies (rather than as a list of the outcomes of interest within whichever studies are included).	Was it clear whether outcomes were specified as inclusion criteria?
7. Pre-defining study designs	Define in advance the eligibility criteria for study designs in a clear and unambiguous way, with a focus on features of a study's design rather than design labels.	Was ambiguity in inclusion criteria for study design avoided?
8. Excluding studies based on publication status	Include studies irrespective of their publication status, unless explicitly justified.	Were studies eligible for inclusion irrespective of publication status?

Item name	Standard	Possible signalling question for bias items
9. Changing eligibility criteria	Justify any changes to eligibility criteria or outcomes studied. In particular, *post hoc* decisions about inclusion or exclusion of studies should keep faith with the objectives of the review rather than with arbitrary rules.	Were studies excluded from the review post hoc for reasons not specified as inclusion criteria?
10. Pre-defining outcomes	Define in advance which outcomes are primary outcomes and which are secondary outcomes.	Were outcomes pre-defined?
11. Pre-defining choices from multiple outcome measures	Define in advance how outcome measures will be selected when there are several possible measures (e.g. multiple definitions, assessors or scales).	Were criteria for selection of outcome measures specified?
12. Pre-defining time points of interest	Define in advance the timing of outcome measurement.	Was timing of outcome measurement pre-specified?
SEARCHING		
13. Searching key databases	Search the Cochrane Review Group's Specialized Register (internally, e.g. via the Cochrane Register of Studies, or externally via CENTRAL). Ensure that CENTRAL and MEDLINE (e.g. via PubMed) have been searched (either for the review or for the Review Group's Specialized Register).	Did the review search an appropriate range of databases?
14. Searching specialist bibliographic databases	Search appropriate national, regional and subject specific bibliographic databases.	
15. Searching for different types of evidence	If the review has specific eligibility criteria around study design to address adverse effects, economic issues or qualitative research questions, undertake searches to address them.	If the review focused on specific types of data, e.g. economic or qualitative questions, were specific searches carried out for these data?
16. Searching trials registers	Search trials registers and repositories of results, where relevant to the topic through ClinicalTrials.gov, the WHO International Clinical Trials Registry	Were trial registers searched?

Item name	Standard	Possible signalling question for bias items
	Platform (ICTRP) portal and other sources as appropriate.	
17. Searching for grey literature	Search relevant grey literature sources such as reports/dissertations/theses databases and databases of conference abstracts.	Were grey literature sources searched?
18. Searching within other reviews	Search within previous reviews on the same topic.	Were previous reviews on the same topic screened?
19. Searching reference lists	Check reference lists in included studies and any relevant systematic reviews identified.	Were reference lists of included studies and relevant reviews screened?
20. Searching by contacting relevant individuals and organisations	Contact relevant individuals and organisations for information about unpublished or ongoing studies.	Were experts and/or relevant organisations contacted for additional studies?
21. Structuring search strategies for bibliographic databases	Inform the structure of search strategies in bibliographic databases around the main concepts of the review, using appropriate elements from PICO and study design. In structuring the search, maximize sensitivity whilst striving for reasonable precision. Ensure correct use of the AND and OR operators.	Was the search structured appropriately?
22. Developing search strategies for bibliographic databases	Identify appropriate controlled vocabulary (e.g. MeSH, Emtree, including 'exploded' terms) and free-text terms (considering, for example, spelling variants, synonyms, acronyms, truncation and proximity operators).	Were search terms appropriate?
23. Using search filters	Use specially designed and tested search filters where appropriate including the Cochrane Highly Sensitive Search Strategies for identifying randomized trials in MEDLINE, but do not use filters in pre- filtered databases e.g. do not use a randomized trial filter in CENTRAL or a systematic review filter in DARE.	Were filters used appropriately?
24. Restricting database searches	Justify the use of any restrictions in the search strategy on publication date,	Were any restrictions on date, publication format,

Item name	Standard	Possible signalling question for bias items
	publication format or language.	or language appropriate?
REVIEW PROCESS		
25. Making inclusion decisions	Use (at least) two people working independently to determine whether each study meets the eligibility criteria, and define in advance the process for resolving disagreements.	Did inclusion assessment involve at least two reviewers? Was the process for resolving disagreements specified?
26. Excluding studies without useable data	Include studies in the review irrespective of whether measured outcome data are reported in a 'usable' way.	Were studies included irrespective of how outcome data were reported?
27. Collating multiple reports	Collate multiple reports of the same study, so that each study rather than each report is the unit of interest in the review.	Was each study rather than report included as the unit of interest?
28. Extracting study characteristics in duplicate	Use (at least) two people working independently to extract study characteristics from reports of each study, and define in advance the process for resolving disagreements.	Did data extraction involve at least two reviewers using a standardised form?
29. Extracting outcome data in duplicate	Use (at least) two people working independently to extract outcome data from reports of each study, and define in advance the process for resolving disagreements.	
30. Obtaining unpublished data	Seek key unpublished information that is missing from reports of included studies.	Were additional sources used to identify data not included in published reports?
31. Assessing risk of bias	Assess the risk of bias for each included study. For randomized trials, the Cochrane 'Risk of bias' tool should be used, involving judgments and supports for those judgments across a series of domains of bias, as described in Chapter 8 of the Cochrane Handbook (version 5 or later).	Was the risk of bias of the included studies formally assessed?
32. Assessing risk of bias in duplicate	Use (at least) two people working independently to apply the risk of bias tool	Did risk of bias assessment involve at least two

Item name	Standard	Possible signalling question for bias items
	to each included study, and define in advance the process for resolving disagreements.	reviewers?
SYNTHESIS		
33. Addressing risk of bias in the synthesis	Address risk of bias in the synthesis (whether qualitative or quantitative). For example, present analyses stratified according to summary risk of bias, or restricted to studies at low risk of bias.	Was risk of bias considered in the synthesis of results?
34. Ensuring meta- analyses are meaningful	Undertake (or display) a meta-analysis only if participants, interventions, comparisons and outcomes are judged to be sufficiently similar to ensure an answer that is clinically meaningful.	If a meta-analysis was conducted were appropriate methods used?
35. Assessing statistical heterogeneity	Assess the presence and extent of between- study variation when undertaking a meta- analysis.	Were differences between studies (heterogeneity) assessed?
36. Addressing missing outcome data	Consider the implications of missing outcome data from individual participants (due to losses to follow up or exclusions from analysis).	Were missing outcome data considered?
37. Addressing skewed data	Consider the possibility and implications of skewed data when analysing continuous outcomes.	Was the possibility and implications of skewed data considered for continuous outcomes?
38. Addressing studies with more than two groups	*If multi-arm studies are included*, analyse multiple intervention groups in an appropriate way that avoids arbitrary omission of relevant groups and double-counting of participants.	Were multi-arm studies analysed appropriately?
39. Comparing subgroups	*If subgroup analyses are to be compared*, and there are judged to be sufficient studies to do this meaningfully, use a formal statistical test to compare them.	Were subgroup analyses compared using formal statistical tests?
40. Interpreting subgroup analyses	*If subgroup analyses are conducted*, follow the subgroup analysis plan specified in the protocol without undue emphasis on particular findings.	Were subgroup analyses pre-specified?

Item name	Standard	Possible signalling question for bias items
41. Considering statistical heterogeneity when interpreting the results	Take into account any statistical heterogeneity when interpreting the results, particularly when there is variation in the direction of effect.	Was heterogeneity taken into account when interpreting the results?
42. Addressing non- standard designs	Consider the impact on the analysis of clustering, matching or other non-standard design features of the included studies.	Was the impact of non-standard design features on the analysis considered?
43. Sensitivity analysis	Use sensitivity analyses to assess the robustness of results, such as the impact of notable assumptions, imputed data, borderline decisions and studies at high risk of bias.	Were sensitivity analyses used to assess the robustness of results?
44. Investigating reporting biases	Consider the potential impact of reporting biases on the results of the review or the meta-analyses it contains.	Was reporting bias assessed?
CONCLUSIONS		
45. Justifying assessments of the quality of the body of evidence	Justify and document all assessments of the quality of the body of evidence (for example downgrading or upgrading if using the GRADE tool).	Were assessments of the quality of the body of evidence justified?
46. Formulating implications for practice	Base conclusions only on findings from the synthesis (quantitative or narrative) of studies included in the review.	Were the review conclusions supported by the results of the review?

Chapter 4. Phase 2: Review of existing quality assessment tools for systematic reviews

Key points

40 existing quality assessment checklists for systematic reviews or meta-analyses were identified

- Only three had been rigorously developed; others were either adapted from existing tools or did not report methods on tool development
- Most tools were generic; 5 targeted reviews of RCTs, 9 others each targeted specific areas including RCTS and non-randomised studies, controlled clinical trials, intervention studies, observational studies, diagnostic test accuracy studies, genetic association studies, health status measurement instruments, scientific and policy research and agronomy
- The number of items in each tool ranged from 4 to 43 (median 10)
- Most tools were simple checklists; three had a more complex structure including one domain based tool
- The majority of tools included a simple rating of yes/no with some also including a not clear/not reported option. Some included a quality scoring system. Four tools included more complex rating systems with 5 to 7 options. Several tools were rated descriptively or did not include a rating system
- IRR, where reported , was fair to high
- We grouped items according to the following domains: selection, searching, review process, synthesis, and conclusions. Most tools included at least 1 item for each domain.

4.1 Objective

To review existing tools designed to assess the quality of systematic reviews of meta-analyses.

4.2 Methods

We conducted a review of existing quality assessment tools for systematic reviews. We included any paper reporting a quality assessment or critical appraisal tool aimed at assessing systematic reviews or meta-analyses. We excluded papers describing tools designed as guidelines for the conduct or reporting of systematic reviews, general lists of items rather than structured tools (i.e.

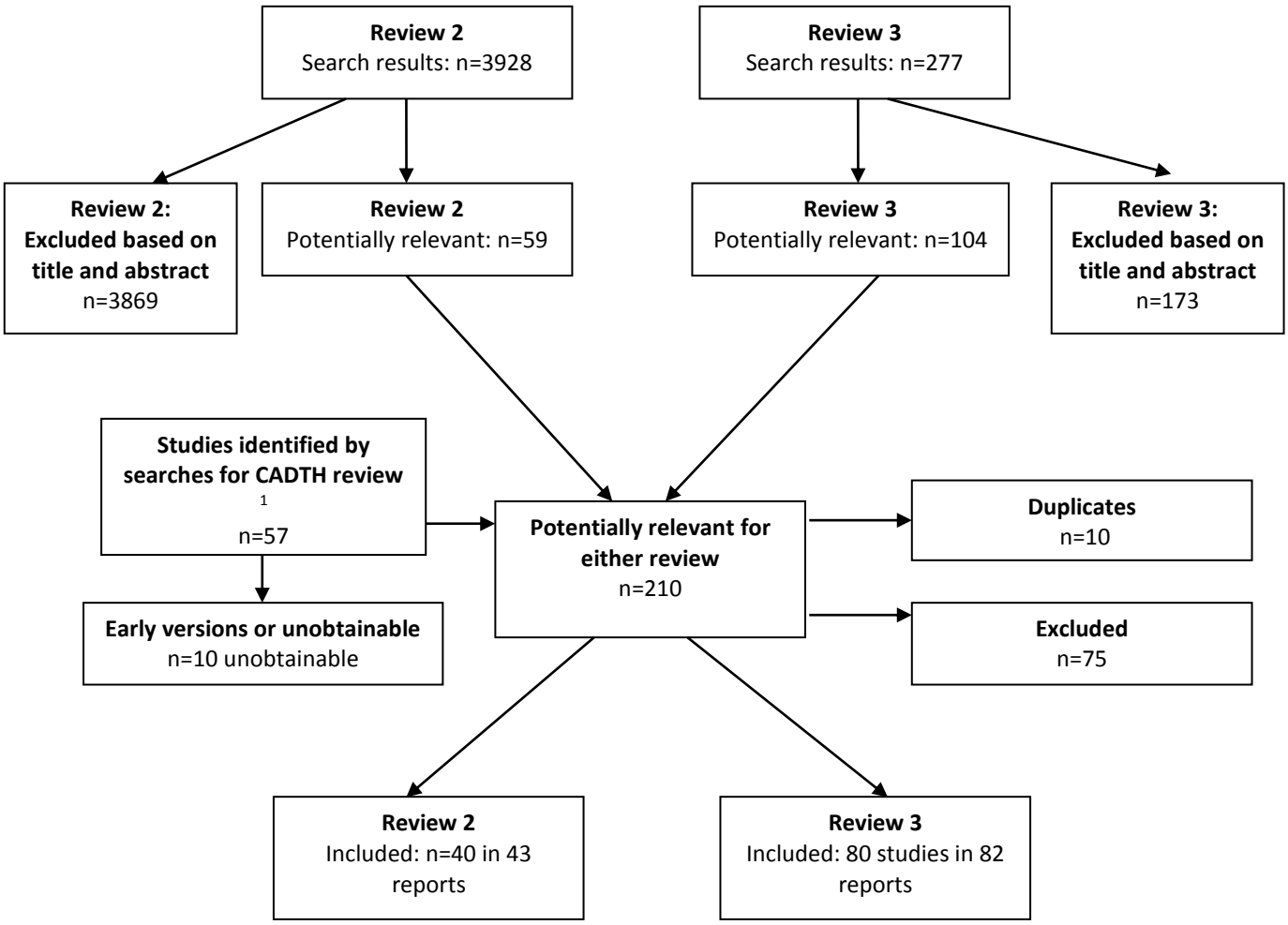
could not be used for evaluative application), conference abstracts and non-English papers. If multiple versions of the same tool were available, the most recent version of the tool was selected. The Canadian Agency for Drugs and Technologies in Health (CADTH) carried out a similar review which was published in 2012.[1] Although the inclusion criteria for the CADTH review differed from this review, the search criteria were similar, allowing us to use their initial pool of studies as a source of potentially relevant articles. The CADTH review identified 57 tools published prior to December 2007; 49 of these were available and were assessed for inclusion in our review. We revised and updated the searches carried out for the CADTH review to identify any tools published since their searches were undertaken. We searched MEDLINE (September 2007 to May 2013), EMBASE (2007 to May 2013), the Cochrane Methodology Register (2007 to 2013), and BIOSIS (2007 to 2013), conducted internet searches using the Google search engine, and contacted expects in the field. Full details of the search strategies are available in Appendix 3.

The results of our searches were screened for relevance independently by two reviewers. Disagreements were resolved through consensus or referral to a third reviewer where necessary. When screening the search results for review 2, studies that appeared relevant for review 3 were also ordered (and vice versa). Similarly, full text inclusion assessment was conducted for reviews 2 and 3 at the same time so that studies ordered for either review were assessed for inclusion in both reviews. We extracted data on the items covered by each of the tools, the general structure of the tool, how items within the tool were rated, whether the tool targeted reviews of specific study designs or topic areas, methods used to develop the tool, any evaluation of the tool, and details on inter-rater reliability. Data on items relating to applicability were not extracted. Individual items were mapped to bias domains used in Phase 1: selection, searching, review process, synthesis, and conclusions. Inclusion assessment and data extraction were performed by one reviewer and checked by a second. We grouped similar items and where possible matched these to the signalling questions proposed as part of Phase 1. We summarised the number of studies assessing each quality item and provided a narrative synthesis of methods used to develop and evaluate the tools.

4.3 Results

The searches identified 3928 records (Figure 2). We included 40 tools reported in 43 publications designed for the quality assessment or critical appraisal of systematic reviews or meta-analyses. Full details of the tools are summarised in Appendix 5.

Figure: Flow of studies through the review process for reviews 2 and 3



Details on tool development

Only three tools could be described as having been rigorously developed; AMSTAR, OQAC and a tool for assessing the quality of meta-analyses.[6, 13, 14] Four tools were adapted from single published tools,[15-18] and ten were adapted from multiple existing tools or guidelines.[19-28] None of the other tools provided details on how the tools were developed.

Study designs targeted by the tools

The majority of tools did not specify what types of review or meta-analysis the tool targeted. However some mentioned a focus of reviews for specific study designs or topic areas. Five tools focused on RCTs,[6, 14, 24, 29, 30] one on RCTs and non-randomised studies,[19] one on controlled clinical trials,[31] one on intervention studies,[32] one on observational studies,[21] one on diagnostic test

accuracy reviews,[33] one on genetic association studies,[34] one on health status measurement instruments,[35] one on "scientific and policy research",[36] and one on agronomy.[37]

Tool Structure

The number of items included in each tool ranged from 4 to 43 (median 10). Most tools were simple checklists but three had more complex structures. One was domain based with four domain questions that were rated as 'Yes', 'Probably Yes', 'Unsure', 'Probably No' and 'No'. Within each domain there were several supportive questions that had specific scoring guidelines – either yes/no/unclear/(not relevant) or specific questions where reviewers were asked to select all answers that applied e.g. "Eligibility criteria were stated and suitably specific for (check all that apply)… (participants, intervention, comparator, outcomes, study designs)".[14] One had three general descriptive questions that started "how…" for example "How were the papers identified?" and then specific questions which could generally be answered as yes/no/unclear for example "Is the topic well defined?".[38] The third tool was structured mainly as a checklist but in addition to the standard checklist questions which were each rated as "Reported, partially reported, not reported" also had a number of additional items, were referred to as "quality standards", which were rated as "yes, unclear or no". These tended to be more specific than the checklist questions. For example, a standard item was "Search Strategy (at least one electronic database was searched and the names of the databases are provided); the supporting quality standard was "at least MEDLINE and EMBASE".

Many tools included single questions covering multiple items; this should be avoided as it makes it very difficult to score items. For example "Have unpublished trials been searched for (contact with investigators and for pharmaceutical companies)?" actually covers three separate components – attempts to locate unpublished data, contacting investigators and contacting pharmaceutical companies. It is difficult to know how to score this item if one out of the three has been done. Other included very broad general questions such as "Is the topic well defined?" and "Whether the search for studies was comprehensive".

Item rating

The most common rating system, used in 13 tools, was based on rating individual items as "yes" or "no" with some tools also including an option for "unclear", "not reported", "can't tell", "not

applicable" or "can't answer".[6, 16, 19, 22, 23, 29, 32, 35, 39-43] A small number of these tools also included options for comments,[41] descriptive answers,[23, 39] or specific answers.[35] Two of these tools incorporated guidelines for grading reviews as high or low risk of bias based on whether they were rated as "yes" for key items.[19, 29] A further tool used a similar rating system but also included a category for "partially" as well as "yes" and "no" and assigned a score of 2 for ratings of "yes" and 1 for ratings of "partially", summing scores to give reviews a summary score.[16] Another tool used a similar rating method but rated items as "specified", "not specified" and "not reported" rather than "yes" and "no".[27] Items on two tools were rated as "adequate", "partial", "none" or "unknown".[17, 31] One of these incorporated a scoring system where "adequate" scored 2 points and "partial" 1 point with scores summed to give a total score.[17] Another tool used a similar rating system, rating items as "reported", "partially reported", and "not reported".[20] One tool rated items as high, low and unclear.[44]

Four tools used more complex rating systems. One tool used a numerical scoring system with different items assigned different numerical scores with scores summed to give a total score.[24] Another used a 7-point scale with items rated from 1 (lowest quality) to 7 (highest quality).[13] One rated items as "well covered", "adequately addressed", "poorly addressed", "not addressed", "not reported", "not applicable".[18] The fourth rated items as "yes", "probably", "unsure", "probably no" and "no".[14]

Five tools were rated descriptively i.e. they required a narrative description of each item,[15, 25, 38, 45, 46] and one tool included a semi-structured rating where the rating varied according to the item.[21] Twelve tools did not specify how items should be rated.[26, 28, 30, 33, 34, 36, 37, 47-51]

Inter-rater reliability

Data on inter-rater reliability was available for 5 tools.[6, 13, 14, 17, 24] The authors of one tool reported that reviewers agreed on 95% of all items in the tool but did not report a formal evaluation of inter-rater reliability.[24] The intraclass correlation coefficient (ICC) for agreement between two scores given by each rater was 0.84 for a modified version of the Sacks tool.[17] Agreement between the score given by each rater and the common score ranged from 0.89 to 0.96.[17] The OQAC tool was reported to have an overall ICC of 0.71 (95% CI 0.59, 0.81), this varied when stratified according to reviewer expertise.[13] Inter-rater agreement for the individual items of

AMSTAR had a mean kappa of 0.70 (95% confidence interval [CI] 0.57, 0.83) ; this corresponds to fair to good agreement.[52] Evaluation of inter-rater agreement for a tool for assessing the quality of a meta-analysis found that weighted kappa measures ranged from 0.30 (summary question B) to 0.45 (summary question D) which was reported to correspond to 'fair' or 'moderate' agreement.[14]

Tool content

Domain 1: Review question and eligibility criteria

All but two of the tools included at least one item relating to framing the review question or eligibility criteria.[21, 37] Three tools contained an item relating to whether the review had mentioned a review protocol.[17, 25, 53] Over half the tools included an item relating to whether the review asked an appropriate or well defined question. Two of these took this a step further to ask whether there was a narrow focus to the question;[22, 42] this is not necessarily associated with risk of bias or a desirable feature of a review. Although over half the tools covered inclusion criteria the majority only considered them in relation to whether they were defined/explicit with only 7 tools including items to assess whether inclusion criteria were appropriate. Only four tools included items to cover specific components of the selection criteria such as population, intervention/index test, study design and outcome. [20, 32] [33, 49] One tool also included an item to assess whether the review was restricted to RCTs and considered this to be positive feature if this was the case.[29] Table 2 provides a summary of the review question and eligibility criteria items covered by the tools with the number of tools covering each item.

Table 3 Number of tools covering each "Selection Domain" question

Question	Number of tools
Was there a review protocol?	3[17, 25, 53]
Did the review ask a well-defined focused question?	24 [6, 15, 18, 20, 22, 23, 25-27, 30, 33, 38-43, 45-47, 49-51]
Was there a narrow focus of the question?	2[22, 42]
Were inclusion and exclusion criteria defined/explicit?	21[6, 14, 16, 20, 22, 23, 25, 27, 29, 30, 32-35, 41, 44, 46, 48, 49, 51, 54]
Were the inclusion criteria appropriate?	7[15, 28, 32, 39, 40, 45, 47, 50]
Were inclusion criteria defined in terms of population/clinical context?	3[20, 32 33]
Were inclusion criteria defined in terms of intervention/index test?	3[20, 32 33]
Were inclusion criteria defined in terms of outcome/reference standard?	3[20, 32 33]
Were inclusion criteria defined in terms of study design?	2[20, 32]
Were inclusion criteria appropriate in terms of population?	1[32]
Were inclusion criteria appropriate in terms of intervention?	1[32]
Were inclusion criteria appropriate in terms of outcome?	2[32, 49]
Were inclusion criteria appropriate in terms of study design?	1[32]
Only RCT/double blind RCTs included?	1[29]

Domain 2: Searching for studies

All but one of the tools included at least one item relating to the literature search.[36] Some tools only assessed a single broad item relating to the literature search such as whether the literature search was "comprehensive" or even just "literature search".[53] Others required a more detailed assessment of individual components of the search such as whether attempts were made to locate unpublished studies and avoid language bias. Others only considered the electronic database component of the search assessing which databases were searched and the date the databases were searched.[48] Table 3 provides a summary of the items covered by the tools with the number of tools covering each item.

Table 4 Number of tools covering each "Searching" Domain question

Question	Number of Tools
Was a comprehensive literature search performed? Is it likely that relevant studies were missed?	23[6, 16, 18, 19, 23, 30, 44, 46, 54 17, 29, 45 39 51 50 32 49 47 40 43 15 38 14]
Literature search/Data identification	2[53 27]
Search Strategy (At least one electronic database was searched and the names of the databases are provided).	1[20]
Are descriptions provided to ensure representativeness of the sample? (*no further details very old review)*	1[26]
Was a two phase search strategy described (identification of search terms and search for studies)?	1[32]
Were details of the search procedures provided?	12[33, 41 22, 25 37 17, 26, 35, 37, 42, 49, 54]
Was the full search strategy presented?	6[14, 33 34 35 21, 49]
Was the search structured appropriately?	2[28 32]
Were multiple bibliographical databases searched?	8[24 14, 39 45 22, 32 35, 48]
Were language restrictions avoided?	8[40 32 6, 21, 22, 39 45 41]
Were reference lists (of included studies and relevant) reviews screened?	8[14 32 6, 22, 39 45 20,24]
Were searches carried out for unpublished studies?	6[40, 41 6, 32, 39, 49]
Were hand-searches conducted?	5[24, 40 14 45 20]
Were grey literature sources searched?	3[6, 14, 21]
Were experts and/or relevant organisations contacted for additional studies?	4[14 39, 49, 24]
Was industry contacted for additional studies?	2[14 49]
Were internet searches carried out?	1[14]
Were in house collections searched?	1[14]
Were age restrictions avoided?	1[41]
Were quality restrictions avoided?	1[41]
Has a search for multiple publications of the same trial or patient data been undertaken?	1[49]

Domain 3: Review Process

All but two of the tools included at least one item related to the review process.[36, 38] The most commonly included item was whether the quality of included studies was formally assessed which is included in 33/40 tools; only three tools included an item on whether the criteria used were appropriate.[16, 32, 54] Whether multiple reviewers were involved in inclusion assessment, data extraction and quality assessment were also frequently covered items. None of the tools made a

distinction between the process of reference screening and inclusion assessment. Sixteen tools also included an item on whether details of included primary studies were reported; this relates more to reporting than to bias and does not strictly map to review process but does not fit in any of our other domains either. Table 4 provides a summary of the items covered by the tools with the number of tools covering each item.

Table 5 Number of tools covering each "Review Process" Domain question

Question	Number of Tools
Did inclusion assessment involve at least two reviewers?	10[14, 19, 40 6, 16, 34, 35, 44 47, 51]
Was inclusion assessment blinded to study results?	2[17, 53]
Was the risk of bias (quality) of the included studies formally assessed?	33[17, 24, 45 30 40 43 14 15, 32 22 19 44 20 27, 34 28, 49 47 50 51 54 39 16 6, 21 42 18 26 23 48 25 38 33]
Were criteria used to assess quality appropriate?	3[32 54 16]
Were all of the trials RCTs?	1[30]
Did risk of bias (quality) assessment involve at least two reviewers?	7[24 32 19 20 47 51 25]
Was agreement between reviewers reported (and acceptable for the quality assessment)?	3[24 20 25]
Did data extraction involve at least two reviewers?	13[17 14 33 32 19 20, 34 35 6 47 51, 53 25]
Was agreement between reviewers reported for data extraction?	4[17, 41 53 25]
Were methods to discuss disagreements in data extraction reported?	1[33]
Was a recognised and agreed upon data extraction tool used?	1[32]
Was data extraction done using a standardised form or were data categories extracted listed?	1[20]
Were data extraction forms pilot tested?	1[41]
Was there a detailed explicit coding book for data extraction?	1[41]
Was data extraction blinded to treatment groups?	2[17, 53]
Were additional sources used to identify data not included in published reports?	7[17 38, 45 34 50 21 48]
Excluded trials listed (and reasons reported)	5[41 53 6 17 49]
Details of included studies reported/tabulated	16[41 33 22 28 6 37 53 25 17 24 17 38 19, 45 34 49]
Is the method used to assess primary studies reproducible?	1[15]
Were inclusion criteria applied in an unbiased way?	1[28]
Is the selection of trials objective and independent of the results (ideally blinded selection)?	1[49]

Question	Number of Tools
Was a description of the methodology used included?	1[18]
Does a theoretical pramework serve as the basis for coding, hypothesis testing and interpretation of results?	1[26]
Are decision rules made explicit at each step of the process?	1[26]

Domain 4: Synthesis

All but one of the tools included at least one item relating to synthesis.[35] Some tools contained single very general items such as "How were the results of the primary studies combined?"[23] whilst others contained a very detailed list of statistical items.[14] The most commonly included item was whether heterogeneity was investigated/assessed, which was included in 19/40 tools; only 9 tools included items to assess whether studies were sufficiently similar to be pooled. Seven tools assessed whether a summary estimate was provided and ten assessed whether methods used to pool data were appropriate. Fifteen tools considered whether reporting/publication bias or missing studies was assessed and 7 tools assessed whether study quality was considered in the synthesis of results. Most other items were each included in one or two tools. Table 5 provides a summary of the items covered by the tools with the number of tools covering each item.

Table 6 Number of tools covering each "Synthesis" Domain question

Question	Number of Tools
Were the statistical methods described?	4[17, 34, 38 49]
What was the overall effect?	3[45 15 39]
How precise were the results?	3[45 15 39]
Has the review question been answered?	1[32]
Were major findings of the review summarised?	1[32]
Were results reported in sufficient detail to enable replication of results by the reviewer?	1[46]
Was there a forest plot/graphical display of study specific results?	2[34 25]
How were the results of the primary studies combined?	1[23]
Were differences between studies (heterogeneity) assessed?	19[38 30 40 14, 22, 32, 33 20 34 28 49 51 31, 37 48 46 25 25 41]
Were the results consistent across studies?	5[45 15 47 21 48]
Were studies sufficiently similar to be pooled?	9[17 41 40 32 36 18 26 45 39]
Were reasons for variation discussed?	1[40]

Question	Number of Tools
Was a sensible strategy used to address statistical heterogeneity in meta-analyses?	1[14]
Was a narrative synthesis presented?	3[19, 27 40]
Was a quantitative (pooled) analysis presented?	7[27 20 31 22 24 41 40]
Were meta-analysis methods reported?	2[32 51]
If a meta-analysis was conducted were appropriate methods used?	10[28 51 31 19, 32 16 6 34 37 30]
If pooling was not performed, were reasons for this reported?	1[24]
Was the power of trials with negative findings discussed?	1[24]
Was reporting (publication) bias assessed?	15[17, 38, 41 14 22 36 20 34 37 6, 31, 46, 48 25]
Were sensitivity analyses used to assess the robustness of results?	5[17 30 34 37 31]
Did the researchers use more than one method of statistical pooling to provide multiple indicators for interpreting the results?	1[41]
How sensitive were the results to the way the review has been conducted?	2[43 50]
Were RCTs discussed separately from other study designs or were only RCTs pooled?	2[24 17]
Was the robustness of the results discussed	1[49]
Were subgroup analyses performed?	3[17 33 31]
Were outcomes related to study characteristics?	1[26]
Are analytic methods used which differentiate whether characteristics affect diagnostic accuracy or test threshold?	1[33]
Were subgroup analyses compared using formal statistical tests?	1[14]
Were subgroup analyses pre-specified?	2[30 49]
Was the rationale for the choice of subgroups given?	1[49]
Was risk of bias (quality) considered in the synthesis of results?	7[40, 43 20, 33, 44 31 16]
Were data analysed on an ITT basis?	1[17]
Were missing outcome data considered?	1[14]
Were cross-over trials mentioned?	1[14]
Were cluster randomised trials mentioned?	1[14]
Were other study designs mentioned?	1[14]
Were comparisons sensible within each meta-analysis?	1[14]
Were outcomes sensible within each meta-analysis?	1[14]
Was double counting of individuals avoided?	2[14 41]
Was the choice of effect size appropriate (e.g. MD vs. SMD)?	1[14]
Was the possibility and implications of skewed data considered for continuous outcomes?	1[14]

Question	Number of Tools
Were methods appropriate to rare events/sparse data?	1[14]
Were cut-points to dichotomize continuous/ordinal outcomes justified?	1[14]
Were time-to-event data appropriately dealt with?	1[14]
Were ordinal data appropriately dealt with?	1[14]
Were indirect comparisons performed appropriately?	2[14 17]
Was the unit of analysis consistent across studies?	1[26]
Were combined tests of significance accompanied with estimates of effect size?	1[41]
Did the researcher examine multiple independent and dependent variables separately through blocking, mediating effects?	1[41]
Were nonparametric measures of effect size used when appropriate, such as with ordinal or dichotomous data?	1[41]
Does the method of pooling sensitivity and specificity take account of their interdependence?	1[33]
Are the estimation of the treatment effect and its CI, and the results of the association tests given?	2[49 41]
When multiple test categories are available, are they used in the summary?	1[33]
Were all important outcomes considered?	2[15 39]
Was the dataset available?	1[37]
Was the statistical programme reported?	1[37]
Were the pooled data appropriate for testing the hypothesis?	1[46]
Was the comparability of the cases and controls assessed?	1[46]

Domain 5: Summarising the findings and reaching conclusions

This domain was the most frequently omitted of the five domains that we have defined. Only 26/40 tools included at least one item covering this domain. The most commonly included item was whether the conclusions were consistent with the review findings. This was addressed by 13 of the tools. Four tools included an item on whether results were appropriately interpreted in the light of risk of bias in included studies. Six tools included one or more items on whether sources of support and/or conflicts of interest were specified. There is debate whether this item is a risk of bias item. Table 6 provides a summary of the items covered by the tools with the number of tools covering each item.

Table 7 Number of tools covering each "Summarising the findings and reaching conclusions" Domain question

Question	Number of tools.
Were the review conclusions supported by the results of the review?	13[14, 16, 19, 20, 29, 30, 38, 40, 43, 45, 50, 51, 54]
Were results appropriately interpreted in the light of risk of bias in included studies?	4[6, 14, 19, 20]
Were the limitations of the meta-analysis identified?	3[19, 30, 41]
Did the researcher consider alternative explanations for the results obtained?	2[41, 46]
Were subgroup analyses interpreted cautiously?	3[14, 45, 50]
Were all important outcomes considered?	1[45]
Were results appropriately interpreted in the light of risk of reporting bias?	1[14]
Were recommendations linked to the strength of the evidence?	1[50]
Were results appropriately interpreted in the light of any multiplicity?	1[14]
If there was "no evidence of effect" was caution taken not to interpret this as "evidence ofno effect"?	1[50]
Were sources of support/conflicts of interest specified?	6[6, 14, 17, 20, 21, 29]

4.4 Summary

We identified 40 existing quality assessment checklists for systematic reviews or meta-analyses. Only three of these had been rigorously developed; others were either adapted from existing tools or did not report methods on tool development. Most tools were generic; 5 targeted reviews of RCTs, 9 others each targeted specific areas including RCTS and non-randomised studies, controlled clinical trials, intervention studies, observational studies, diagnostic test accuracy studies, genetic association studies, health status measurement instruments, scientific and policy research and agronomy. The number of items in each tool ranged from 4 to 43 (median 10). Most tools were simple checklists; three had a more complex structure including one domain based tool. The majority of tools included a simple rating of yes/no with some also including a not clear/not reported option. Some included a quality scoring system. Four tools included more complex systems with 5 to 7 options. Several tools were rated descriptively or did not include a rating system. Inter-rater reliability, where reported, was fair to high. We grouped items according to the following domains: selection, searching, review process, synthesis, and conclusions. Most tools included at least 1 item for each domain.

Chapter 5. Phase 3: Review of studies that have used the AMSTAR tool

Key points

We included 80 overviews that used AMSTAR to assess the quality of included systematic reviews.

- Systematic reviews included in the overviews included a variety of study designs but all either investigated epidemiological associations or interventions; none included diagnostic or prognostic studies
- 57 reviews included RCTs and 32 included observational studies, some also included other designs such as case-studies/series, n of 1 studies and descriptive studies
- The number of systematic reviews included in each overview ranged from 1 to 369 (median 16)
- 17 overviews modified AMSTAR or used a modified version: 3 used R-AMSTAR, 1 used a Chinese translation of AMSTAR, 8 used the original AMSTAR items but modified the scoring, 3 added items to AMSTAR, 1 removed an item from AMSTAR (item on conflict of interest), 2 modified guidelines for scoring items (1 also modified scoring)
- Most overviews only included a narrative discussion of the AMSTAR assessment; 6 overviews used AMSTAR as a criterion for inclusion, 3 on the basis of summary scores; 3 overviews did not report any results of the AMSTAR assessment and 3 only reported details in tables; 5 performed statistical investigations of associations between AMSTAR items and various other features
- 60/80 assigned summary quality scores. The majority assigned 1 point for each AMSTAR item and summed scores to generate a summary score. Some stratified studies as high, medium or low quality based on summary scores

5.1 Objective

To review studies that have used the AMSTAR tool to assess the quality of systematic reviews.

5.2 Methods

Although there is currently no accepted tool to assess the quality of systematic reviews, our preliminary searches demonstrated that the AMSTAR tool is the most commonly used. We carried out a review of studies that used the AMSTAR tool to evaluate how the quality assessment of systematic reviews is conducted in practice. The aim of this review was to provide information on the requirements of users of such a tool. We identified reviews that had used AMSTAR through the following searches:

1. ISI Web of Science citation search of the three key AMSTAR publications[5, 6, 52]
2. Cochrane library
    a. 'AMSTAR' in Search all text
    b. 'Overview' in title, abstract or keywords
3. NHS Centre for Reviews and Dissemination's Database of Abstracts of Reviews of Effects (DARE) http://www.crd.york.ac.uk/CRDWeb/SearchPage.asp - 'AMSTAR' in any field
4. Cochrane Collaboration's Comparing Multiple Interventions Methods Group register of overviews of reviews http://cmimg.cochrane.org/

The results of our searches were screened for relevance independently by two reviewers. Disagreements were resolved through consensus or referral to a third reviewer where necessary. When screening the search results for review 3, studies that appeared relevant for review 2 were also ordered (and vice versa). Similarly, full text inclusion assessment was conducted for reviews 2 and 3 at the same time so that studies ordered for either review were assessed for inclusion in both reviews. For practical reasons, the review was restricted to full-text reports published in English, French or Dutch. We included any review that had used AMSTAR to assess the quality of included systematic reviews. We extracted information about the review topic, number of included SRs, number and types of included study, methods of synthesis, any modifications made to AMSTAR, and how the tool was used within the review. Inclusion was performed by one reviewer and checked by a second. Data extraction was performed by one reviewer. A second reviewer checked 50% of the data extraction. A narrative synthesis was used to combine results.

5.3 Results

The searches identified 277 titles and abstracts (Figure 2). We included 80 overviews reported in 82 publications that used AMSTAR to assess the quality of included systematic reviews.

Details of included overviews

The systematic reviews included in the overviews included a variety of study designs but all either investigated epidemiological associations or interventions; none included diagnostic or prognostic studies. Where reported, 57 reviews included RCTs and 32 included observational studies (.e.g. cohort studies, case-control studies, cross-sectional studies, interrupted time series, before and after studies), some also included other designs such as case-studies/series, [55-62] n of 1 studies [56], and descriptive studies.[57, 63] Of the reviews that included RCTs, 30 also included observational studies and two included qualitative studies.[55, 64] A further review included qualitative studies and unspecified quantitative studies.[65] The number of systematic reviews included in each overview ranged from 1 to 369 (median 16). Five of the overviews conducted a quantitative synthesis [66-70], three in the form of a network meta-analysis[66, 67, 69] and two conducted a meta-analysis of primary studies from the included systematic reviews. [68, 70] The remaining studies provided a narrative description, 26 of these only presented a description of the methodological quality of the studies with no synthesis of the actual findings of the reviews.[71] [65, 72-75] [59, 76-87] [88-94]

Modifications made to AMSTAR

Seventeen overviews modified the AMSTAR tool or used a modified version of the tool,[57, 86, 91, 95-97, 55, 59-61, 68, 71, 72, 85, 98-100] all other reviews used AMSTAR in its published form. Three reviews [61, 96, 98] used the R-AMSTAR criteria[101] and one used the original AMSTAR tool but translated it into Chinese.[71] Eight reviews retained the original AMSTAR items but made modifications to the way in which these were scored,[59, 68, 71, 86, 87, 96, 98, 102] The original AMSTAR tool scores each item as "yes", "no", "can't answer" or "not applicable". Changes to scoring included: adding a "partially" category[86]; collapsing "no" and "can't answer" into a single category[91]; collapsing "not applicable" and "can't answer" into a single category[85]; changing the scoring to 'met,' 'unclear/partly met,' or 'not met' (2 overviews)[97, 99]; changing the "can't answer" and "not applicable" to "not reported" and "unclear"[60]; adding a category of "not reported"[100]; and changing the scoring system to "high", "low" or "unclear"[44]. Three overviews added items to the AMSTAR tool. One assessed whether or not the outcome measures in the reviews were clearly described and integrated in the results [57], one assessed whether the effect of methodological bias analysed[68] and one added two items relating to external validity[55] – one concerning reporting of participants' functional limitations and one on the study setting. Only one overview removed an item from AMSTAR - the

one related to conflict of interest.[95] Two overviews modified the guidelines for scoring individual AMSTAR items and provided a detailed description of changes made[68, 86], one of these also changes the way items were scored.[86]

Incorporation of AMSTAR assessment into the review

The majority of overviews only provided a narrative discussion of the AMSTAR assessment and did not make any further attempts to integrate the results into the overview. Six overviews used AMSTAR as a criterion for inclusion of systematic reviews in the overview. Three of these assigned summary quality scores and only included reviews that scored above a certain AMSTAR score.[96, 103, 104] A further overview also assigned summary scores and selected only the systematic review with the highest AMSTAR score for inclusion and updating.[70] One overview excluded one review as it did not fulfil any of the AMSTAR criteria[99] and another reported only data from systematic reviews that it considered to be reliable based on AMSTAR ratings, although it did not report exact criteria used to make a judgement of what was considered reliable.[105] One review intended to include AMSTAR scores as independent variables in meta-regression analyses but was unable to do so due to insufficient observations.[68] Three reviews reported using AMSTAR but did not report any results of the AMSTAR assessment.[88, 106] Two reviews only reported summary gradings of quality based on AMSTAR in tables with no discussion of AMSTAR ratings in the text;[107, 108] in one of the overviews details on how the grading was made were not reported.[107] A further review provided a detailed breakdown of the AMSTAR rating in a table but did not discuss the assessment further in the text.[109] One overview constructed veritas plots with AMSTAR scores constituting one item on these plots.[77] One overview used linear regression analysis with AMSTAR score as the dependent variable to investigate changes in study quality over time.[78] One overview used ANOVA to compare AMSTAR ratings across various groupings within the overview (e.g. meta-analysis vs. systematic review)[110] and another used students t-test to compare groupings (e. g. systematic vs. other reviews).[111] One overview assessed the association between PRISMA and AMSTAR scores using linear regression.[92] Another performed an empirical comparison of gradings obtained with AMSTAR to those obtained with R-AMSTAR and also of compliance with individual AMSTAR items between Cochrane and non-Cochrane reviews.[79, 112]

Summary quality ratings

Sixty overviews produced some form of summary quality rating. Almost all (49 overviews) assigned systematic reviews a score of one for each AMSTAR item fulfilled and then added these items to produce a summary score. Twenty one of these overviews then stratified studies as high (9-11 or 8-11 points), medium (5-8 or 4-7 points) or low (0-4 or 0-3 points) quality based on their summary quality scores. One overview assigned studies 1 point if an AMSTAR item was rated as yes but also assigned items that were rated as "can't tell" 0.5 points.[78] A further overview adopted a similar scheme but assigned reviews 2 points for AMSTAR items that were met and 1 point for those that were partially met.[86] Three overviews[61, 96, 98] used the R-AMSTAR tool[101] which incorporates a summary quality score assigning systematic reviews a score out of 44. Five overviews assigned reviews a score based on the percentage of items fulfilled.[79, 92, 108, 113, 114] Two of these stratified studies as high (73-100%), medium (27-73%) or low (<27%) quality.[108, 113] One overview graded studies as A, B or C based on AMSTAR ratings but did not report how this was done.[107]

5.4 Summary

We included 80 overviews that used AMSTAR to assess the quality of included systematic reviews. Systematic reviews included in the overviews included a variety of study designs but all either investigated epidemiological associations or interventions; none included diagnostic or prognostic studies. Fifty seven reviews included RCTs and 32 included observational studies, some also included other designs such as case-studies/series, n of 1 studies and descriptive studies. The number of systematic reviews included in each overview ranged from 1 to 369 (median 16). Seventeen overviews modified AMSTAR or used a modified version: 3 used R-AMSTAR, 1 used a Chinese translation of AMSTAR, 8 used the original AMSTAR items but modified the scoring, 3 added items to AMSTAR, 1 removed an item from AMSTAR (item on conflict of interest), 2 modified guidelines for scoring items (1 also modified scoring). Most overviews only included a narrative discussion of the AMSTAR assessment; 6 overviews used AMSTAR as a criterion for inclusion, 3 on the basis of summary scores; 3 overviews did not report any results of the AMSTAR assessment and 3 only reported details in tables; 5 performed statistical investigations of associations between AMSTAR items and various other features. A large proportion (60/80) of overviews assigned summary quality scores. The majority assigned 1 point for each AMSTAR item and summed scores to generate a summary score. Some stratified studies as high, medium or low

quality based on summary scores.

Chapter 6. Generating a list of items

Key points

- We propose 5 domains for ROBIS: (1) Review question and eligibility criteria; (2) searching for studies; (3) review process; (4) synthesis; (5) summarising the findings and reaching conclusions
- Based on the reviews conducted in phase 1 and 2 we have proposed a list of possible signalling questions for consideration for each domain
- Many existing tools and previous reviews using AMSTAR have used a summary quality score. We do not want to incorporate a summary score into ROBIS but will consider whether there are other ways of producing a summary assessment of study quality based on the ROBIS assessment
- The only modification made to AMSTAR that may have an impact on ROBIS is the decision by one of the tools authors' to remove the item on conflict of interest.
- Most of the tools included in Phase 2 of this project were generic in focus or did not state a specific focus.. Systematic reviews included in the overviews evaluated as part of Phase 3 included a variety of study designs (RCTs, observational studies, case-studies/series, n of 1 studies and descriptive studies) but all either investigated epidemiological associations or interventions; none included diagnostic or prognostic studies.

6.1 Domains

The classification of items for Phase 1 and 2 based on our 5 proposed domains worked well. We therefore propose retaining these domains and naming them as follows:

Domain 1: Review question and eligibility criteria

Domain 2: Searching for studies

Domain 3: Review Process

Domain 4: Synthesis

Domain 5: Summarising the findings and reaching conclusions

6.2 Possible signalling questions

Based on the reviews conducted as part of Phase 1 and 2 we have developed the following list of items for consideration for inclusion in ROBIS, grouped according to domain. These tables summarise the evidence from Phase and 1 and 2 of the project showing whether they were identified as possible questions by the MECIR project or if they were included in previous tools, how many previous tools they were included in:

Domain 1: Review question and eligibility criteria

Possible signalling question	**MECIR or number of tools**
Were review objectives clearly specified?	MECIR
Did the review ask a well-defined focused question?	24
Was there a narrow focus of the question?	2
Were inclusion criteria clearly defined?	MECIR
Were inclusion and exclusion criteria defined/explicit?	21
Were the inclusion criteria appropriate?	7
Were criteria for handling studies that include only a subset of relevant participants specified?	MECIR
Were inclusion criteria defined in terms of population/clinical context?	3
Were inclusion criteria appropriate in terms of population?	1
Was ambiguity in inclusion criteria for interventions and comparators avoided?	MECIR
Were inclusion criteria defined in terms of intervention/index test?	3
Were inclusion criteria appropriate in terms of intervention?	1
Was it clear whether outcomes were specified as inclusion criteria?	MECIR
Were inclusion criteria defined in terms of outcome/reference standard?	3
Was ambiguity in inclusion criteria for study design avoided?	MECIR, 2
Were studies eligible for inclusion irrespective of publication status?	MECIR
Were studies excluded from the review post hoc for reasons not specified as inclusion criteria?	MECIR
Were outcomes pre-defined?	MECIR
Were criteria for selection of outcome measures specified?	MECIR
Was timing of outcome measurement pre-specified?	MECIR
Were inclusion criteria appropriate in terms of outcome?	2
Were inclusion criteria appropriate in terms of study design?	1

Domain 2: Searching for studies

Possible signalling question	**MECIR or number of tools**
Was a comprehensive literature search performed?	23
Are descriptions provided to ensure representativeness of the sample?	1
Was a two phase search strategy described (identification of search terms and search for studies)?	1
Did the review search an appropriate range of databases?	MECIR

Possible signalling question	MECIR or number of tools
Was the internet searched?	1
Were multiple bibliographical databases searched?	8
Search Strategy (At least one electronic database was searched and the names of the databases are provided).	1
Were details of the search procedures provided?	12
Was the full search strategy presented?	6
If the review focused on specific types of data, e.g. economic or qualitative questions, were specific searches carried out for these data?	MECIR
Was the search structured appropriately?	MECIR, 2
Were search terms appropriate?	MECIR
Were filters used appropriately?	MECIR
Were trial registers searched?	MECIR
Were grey literature sources searched?	MECIR, 3
Were in house collections searched?	1
Were previous reviews on the same topic screened?	MECIR
Were reference lists (of included studies and relevant) reviews screened?	MECIR, 8
Were hand-searches conducted?	5
Were experts and/or relevant organisations contacted for additional studies?	MECIR, 4
Was industry contacted for additional studies?	2
Were any restrictions on date, publication format, or language appropriate?	MECIR
Were language restrictions avoided?	8
Were searches carried out for unpublished studies?	6
Were age restrictions avoided?	1
Were quality restrictions avoided?	1
Has a search for multiple publications of the same trial or patient data been undertaken?	1

Domain 3: Review Process

Possible signalling question	MECIR or number of tools
Did inclusion assessment involve at least two reviewers?	MECIR, 10
Was inclusion assessment blinded to study results?	2
Was the process for resolving disagreements specified?	MECIR
Were studies included irrespective of how outcome data were reported?	MECIR
Was each study rather than report included as the unit of interest?	MECIR
Were inclusion criteria applied in an unbiased way?	1
Is the selection of trials objective and independent of the results (ideally blinded selection)?	1
Did data extraction involve at least two reviewers?	MECIR, 13
Was agreement between reviewers reported for data extraction?	4
Were methods to discuss disagreements in data extraction reported?	1

Possible signalling question	MECIR or number of tools
Was a recognised and agreed upon data extraction tool used?	1
Was data extraction done using a standardised form (or were data categories extracted listed)?	1
Were data extraction forms pilot tested?	1
Was there a detailed explicit coding book for data extraction?	1
Was data extraction blinded to treatment groups?	2
Were additional sources used to identify data not included in published reports?	MECIR, 7
Does a theoretical framework serve as the basis for coding, hypothesis testing and interpretation of results?	1
Was the risk of bias (quality) of the included studies formally assessed?	MECIR, 33
Were criteria used to assess quality appropriate?	3
Did risk of bias (quality) assessment involve at least two reviewers?	MECIR, 7
Were all of the trials RCTs?	1
Was agreement between reviewers reported (and acceptable for the quality assessment)?	3
Is the method used to assess primary studies reproducible?	1
Was a description of the methodology used included?	1
Are decision rules made explicit at each step of the process?	1
Excluded trials listed (and reasons reported)	5
Details of included studies reported/tabulated	16

Domain 4: Synthesis

Possible signalling question	MECIR or number of tools
Were the statistical methods described?	4
What was the overall effect?	3
How precise were the results?	3
Has the review question been answered?	1
Were major findings of the review summarised?	1
Were results reported in sufficient detail to enable replication of results by the reviewer?	1
Was there a forest plot/graphical display of study specific results?	2
How were the results of the primary studies combined?	1
Were differences between studies (heterogeneity) assessed?	19
Were the results consistent across studies?	5
Were studies sufficiently similar to be pooled?	9
Were reasons for variation discussed?	1
Was heterogeneity taken into account when interpreting the results?	MECIR
Was a sensible strategy used to address statistical heterogeneity in meta-analyses?	1
Was a narrative synthesis presented?	3
Was a quantitative (pooled) analysis presented?	7
Were meta-analysis methods reported?	2
If a meta-analysis was conducted were appropriate methods used?	MECIR, 10
If pooling was not performed, were reasons for this reported?	1
Was the power of trials with negative findings discussed?	1

Possible signalling question	MECIR or number of tools
Was reporting (publication) bias assessed?	MECIR, 15
Were sensitivity analyses used to assess the robustness of results?	MECIR, 5
Did the researchers use more than one method of statistical pooling to provide multiple indicators for interpreting the results?	1
How sensitive were the results to the way the review has been conducted?	2
Were RCTs discussed separately from other study designs or were only RCTs pooled?	2
Was the robustness of the results discussed	1
Were subgroup analyses performed?	3
Were outcomes related to study characteristics?	1
Are analytic methods used which differentiate whether characteristics affect diagnostic accuracy or test threshold?	1
Were subgroup analyses compared using formal statistical tests?	MECIR, 1
Were subgroup analyses pre-specified?	MECIR, 2
Was the rationale for the choice of subgroups given?	1
Was risk of bias (quality) considered in the synthesis of results?	MECIR, 7
Were data analysed on an ITT basis?	1
Were missing outcome data considered?	MECIR, 1
Was the impact of non-standard design features on the analysis considered?	MECIR
Were cross-over trials mentioned?	1
Were cluster randomised trials mentioned?	1
Were other study designs mentioned?	1
Were comparisons sensible within each meta-analysis?	1
Were outcomes sensible within each meta-analysis?	1
Was double counting of individuals avoided?	2
Was the choice of effect size appropriate (e.g. MD vs. SMD)?	1
Was the possibility and implications of skewed data considered for continuous outcomes?	MECIR, 1
Were methods appropriate to rare events/sparse data?	1
Were cut-points to dichotomize continuous/ordinal outcomes justified?	1
Were time-to-event data appropriately dealt with?	1
Were ordinal data appropriately dealt with?	1
Were indirect comparisons performed appropriately?	2
Was the unit of analysis consistent across studies?	1
Were multi-arm studies analysed appropriately?	MECIR
Were combined tests of significance accompanied with estimates of effect size?	1
Did the researcher examine multiple independent and dependent variables separately through blocking, mediating effects?	1
Were nonparametric measures of effect size used when appropriate, such as with ordinal or dichotomous data?	1
Does the method of pooling sensitivity and specificity take account of their interdependence?	1
Are the estimation of the treatment effect and its CI, and the results of the association tests given?	2
When multiple test categories are available, are they used in the summary?	1
Were all important outcomes considered?	2
Was the dataset available?	1
Was the statistical programme reported?	1

Possible signalling question	MECIR or number of tools
Were the pooled data appropriate for testing the hypothesis?	1
Was the comparability of the cases and controls assessed?	1
Were missing outcome data considered?	MECIR

Domain 5: Summarising the findings and reaching conclusions

Possible signalling question	MECIR or number of reviews
Were assessments of the quality of the body of evidence justified?	MECIR
Were the review conclusions supported by the results of the review?	MECIR, 13
Were results appropriately interpreted in the light of risk of bias in included studies?	4
Were the limitations of the meta-analysis identified?	3[19, 30, 41]
Did the researcher consider alternative explanations for the results obtained?	2[41, 46]
Were subgroup analyses interpreted cautiously?	3
Were all important outcomes considered?	1
Were results appropriately interpreted in the light of risk of reporting bias?	1
Were recommendations linked to the strength of the evidence?	1
Were results appropriately interpreted in the light of any multiplicity?	1
If there was "no evidence of effect" was caution taken not to interpret this as "evidence ofno effect"?	1
Were sources of support/conflicts of interest specified?	6

6.3 Implications of reviews for ROBIS structure

General structure

The majority of existing tools use a simple checklist approach. Only one used a more domain based approach,[14] although this was the most recently published tool. This may reflect advances in the area of quality assessment where more recently developed tools in other areas have adopted a domain based approach (e.g. Cochrane risk of bias tool and QUADAS-2).

Item rating

The majority of tools included a simple rating of yes/no with some also including a not clear/not reported option. AMSTAR also used this rating system – items were rated as “Yes”, “no”, “can't answer”, “not applicable”. The tool developed by Higgins et al, the only one to proposes a domain based approach, used a “Yes”, “Probably Yes”, “Unsure”, “Probably No” and “No” rating

for the domain questions.[14] This is similar to the rating system that we are considering for the signalling questions in ROBIS. Other tools used more complex scoring systems but these were each used by a small number of tools. Some tools were descriptive/open ended in their answers; these would be very difficult to apply objectively in practice and would be more useful for general critical appraisal.

Some of the tools identified as part of Phase 2 incorporate a summary quality score into their rating system. Most rated items equally but one applied different weights to different items according to their perceived importance. A large proportion of the reviews included in Phase 3 (60/80) also assigned summary quality scores. The majority assigned 1 point for each AMSTAR item and summed scores to generate a summary score. Some stratified studies as high, medium or low quality based on summary scores. This shows that reviewers would like some form of overall summary of study quality based on the results of the quality assessment. However, there are a number of problems associated with the use of quality scores which means that their use is not generally recommended and is something that we may want to avoid with ROBIS.[115, 116] We may therefore want to consider whether there are other ways of producing a summary assessment of study quality based on the ROBIS assessment; this is something that can be considered once the tool has been developed.

Modification to AMSTAR

Of the 80 reviews that used AMSTAR very few made modifications, of which most were minor. The only modification that may have an impact on ROBIS is the decision by one of the tools authors' to remove the item on conflict of interest. There is general debate regarding whether conflict of interest/source of funding is something that should be considered as part of the quality assessment as a potential source of bias.

Incorporation of the results of the quality assessment

How the results of the ROBIS assessment will be used and incorporated into the overview is something to consider when deciding on its structure. Based on the review of how AMSTAR was used, most overviews only included a narrative discussion of the AMSTAR assessment. However, a small number of reviews used a formal incorporation of results restricting inclusion based on the

results of the AMSTAR assessment or performing statistical investigations of associations between AMSTAR items and various other features

Study designs to be targeted

A key decision regarding ROBIS is whether it should be generic in focus i.e. targeting all systematic reviews of any study design whether of RCTs, diagnostic accuracy studies, observational studies or prognostic studies etc. or to initially be more focused in design. Most of the tools included in Phase 2 of this project were generic in focus or did not state a specific focus although some of the tools were targeted to specific study designs most commonly RCTs (5 tools) but there were also tools aimed at reviews of observational studies and diagnostic accuracy studies. Systematic reviews included in the overviews evaluated as part of Phase 3 included a variety of study designs (RCTs, observational studies, case-studies/series, n of 1 studies and descriptive studies) but all either investigated epidemiological associations or interventions; none included diagnostic or prognostic studies.

Chapter 7. Items for discussion at face-to-face meeting

Key points

- Agree preliminary conceptual decision made by the steering group
    - Definition of risk of bias
    - General application
    - Structure
    - Rating of domains
    - Comprehensive nature of tool
- Tool properties to be discussed:
    - Generic or RCTs only?
    - Number of signalling questions
    - Rating of signalling questions
    - Overall rating of study quality without using summary quality scores?
    - Should conflict of interest/source of funding be included as a risk of bias item?
- Tool content
    - Small group discussions on signalling questions for each domain
    - Risk of bias questions for each domain

7.1 Agree preliminary conceptual decision made by the steering group (Chapter 2)

- **Definition of risk of bias in systematic reviews used for the ROBIS project**

  "a systematic error or deviation from the truth, in the summary estimates and/or review conclusions" and is therefore related only to the internal validity of the review. ROBIS will not consider applicability.

- **ROBIS will have the following general applications:**
    - Allow those conducting overviews of systematic reviews to assess the risk of bias in included studies
    - Allow consistent and reliable assessment of risk of bias by reviewers with different backgrounds
    - Distinguish between reviews at high and low risk of bias

- **Structure**
    - Relatively short and straightforward to complete.
        - Domain based structure similar to those used in Cochrane Risk of Bias tools and QUADAS-2.
        - Domains: Selection, Searching, Review Process, Synthesis, Conclusions
        - Signalling questions will be included to help judge the risk of bias; these questions flag aspects of study design related to the potential for bias and aim to help reviewers judge risk of bias.
        - ROBIS should not incorporate a summary quality score.

- **Rating:** three phased approach to scoring risk of bias
    - (1) Information used to support the judgment of risk of bias, (2) signalling questions, and (3) judgment of risk of bias.
    - Use of "low," "high," or "unclear" as domain-level judgements.

- **Comprehensive nature of the tool:** When developing ROBIS we need to aim to develop a set of independent criteria that work together, i.e. to ensure that there is no overlap between items.

7.2 Tool properties to be discussed:

- Should ROBIS aim to assess the risk of bias of a systematic review in generic terms (relevant to all types of systematic review e.g. reviews of RCTs, observational studies, DTA studies, prognostic studies etc.) or should it initially focus only on reviews of RCTs?
- Scoring of items. We have agreed that domains should be scored as "high", "low", or "unclear" risk of bias. However, the scoring of signalling questions needs further discussion. There are two possibilities: either score items as "yes", "no" or "unclear" or move to a rating of "yes", "probably yes", "probably no", "no", "no information".
- Can we provide guidance on produce some overall rating of a study's quality without using summary quality scores?
- How many signalling questions should we be aiming for per domain?
- Should conflict of interest/source of funding be included as a risk of bias item?

7.3 Tool content

- Small group discussions to discuss signalling questions for each of the five domains
- Propose risk of bias questions for each domain

8. References

[1] Bai AS, V.K.; Bak, G.; Wells, G. *Quality Assessment Tools Project Report*. Ottawa: Canadian Agency for Drugs and Technologies in Health, 2012

[2] Whiting P, Sterne J. Systematic Reviews and Meta-anlysis. In: Ben-Shlomo Y, Brookes S, Hickman M, editors. *Lecture notes. Epidemiology, evidence-based medicine, and public health*. 6th ed. Chichester, West Sussex: John Wiley & Sons, 2013: 102-111.

[3] Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Med* 2009;6(7):e1000100.

[4] Chandler J, Churchill R, Higgins J, Lasserson T, Tovey D. Methodological Expectations of Cochrane Intervention Reviews (MECIR): Methodological standard for the conduct of new Cochrane Intervention Reviews. The Cochrane Collaboration, 2012. Version 2.2, 17 December 2012. Available from: http://www.editorial-unit.cochrane.org/mecir

[5] Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Med. Res. Methodol.* 2007;7:10.

[6] Shea BJ, Hamel C, Wells GA, Bouter LM, Kristjansson E, Grimshaw J, et al. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *J. Clin. Epidemiol.* 2009;62(10):1013-1020.

[7] Higgins JPT, Altman DG, Gøtzsche PC, Moher D, Oxman AD, Savovic J, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomized trials. *British Medical Journal* 2011.

[8] Sterne JAC, Savovic J, Reeves BC, Moher D, Turner L. Extending the Cochrane Risk of Bias tool to assess risk of bias in randomised trials with non-parallel-group designs, and non-randomised studies. Cochrane Collaboration, 2012.

[9] Whiting P, Rutjes AW, Dinnes J, Reitsma J, Bossuyt PM, Kleijnen J. Development and validation of methods for assessing the quality of diagnostic accuracy studies. *Health Technol Assess* 2004;8(25):iii, 1-iii234.

[10] 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.

[11] Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;155(8):529-536.

[12] Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. *PLoS Med* 2010;7(2):e1000217.

[13] Oxman AD, Guyatt GH. Validation of an index of the quality of review articles. *J. Clin. Epidemiol.* 1991;44(11):1271-8.

[14] Higgins J, Lane PW, Anagnostelis B, Anzures-Cabrera J, Baker NF, Cappelleri JC, et al. A tool to assess the quality of a meta-analysis. *Royal Society of Medicine* In Press.

[15] Ho RC, Ong HS, Kudva KG, Cheung MW, Mak A. How to critically appraise and apply meta-analyses in clinical practice. *International Journal of Rheumatic Diseases* 2010;13(4):294-299.

[16] Santaguida P, Oremus M, Walker K, Wishart LR, Siegel KL, Raina P. Systematic reviews identify important methodological flaws in stroke rehabilitation therapy primary studies: Review of reviews. *J. Clin. Epidemiol.* 2012;65(4):358-367.

[17] Auperin A, Pignon JP, Poynard T. Review article: critical review of meta-analyses of randomized clinical trials in hepatogastroenterology. *Alimentary Pharmacology & Therapeutics* 1997;11(2):215-25.

[18] Network SIG. Methodology checklist 1: systematic reviews and meta-analyses. In: *SIGN 50: a guideline developer's handbook*. Edinburgh: SIGN, 2009. Available from: http://www.sign.ac.uk/guidelines/fulltext/50/checklist1.html

[19] Li TJ, Vedula SS, Scherer R, Dickersin K. What Comparative Effectiveness Research Is Needed? A Framework for Using Guidelines and Systematic Reviews to Identify Evidence Gaps and Research Priorities. *Ann. Intern. Med.* 2012;156(5):367-U199.

[20] Mailis A, Taenzer P. Evidence-based guideline for neuropathic pain interventional treatments: Spinal cord stimulation, intravenous infusions, epidural injections and nerve blocks. *Pain Res. Manag.* 2012;17(3):150-158.

[21] Shamliyan T, Kane RL, Jansen S. Quality of Systematic Reviews of Observational Nontherapeutic Studies. *Prev. Chronic Dis.* 2010;7(6).

[22] Knox EM, Thangaratinam S, Kilby MD, Khan KS. A review of the methodological features of systematic reviews in fetal medicine. *European Journal of Obstetrics Gynecology and Reproductive Biology* 2009;146(2):121-128.

[23] Smith AF. An analysis of review articles published in four anaesthesia journals. *Canadian Journal of Anaesthesia* 1997;44(4):405-9.

[24] Assendelft WJ, Koes BW, Knipschild PG, Bouter LM. The relationship between methodological quality and conclusions in reviews of spinal manipulation. *JAMA* 1995;274(24):1942-8.

[25] Wilson A, Henry DA. Meta-analysis. Part 2: Assessing the quality of published meta-analyses. *Medical Journal of Australia* 1992;156(3):173-4, 177-80, 184-7.

[26] Smith MC, Stullenbarger E. Meta-analysis: an overview. *Nursing Science Quarterly* 1989;2(3):114-5.

[27] Mulrow CD. The medical review article: state of the science. *Ann. Intern. Med.* 1987;106(3):485-8.

[28] (NMHRC) NHaMRC. *How to review the evidence: assessment and application of scientific evidence*. Canberra, Australia, 2000.

[29] Zambon M, Biondi-Zoccai G, Bignami E, Ruggeri L, Zangrillo A, Landoni G. A comprehensive appraisal of meta-analyses focusing on nonsurgical treatments aimed at decreasing perioperative mortality or major cardiac complications. [Review]. *Journal of Anesthesia* 2012;26(4):509-515.

[30] Geller NL, Proschan M. Meta-analysis of clinical trials: a consumer's guide. *Journal of Biopharmaceutical Statistics* 1996;6(4):377-94.

[31] Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. *New England Journal of Medicine* 1987;316(8):450-5.

[32] Institute JB. RAPid: Rapid Appraisal protocol internet database. . Adelaide: The Joanna Briggs Institute, 2006.

[33] Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, et al. Guidelines for meta-analyses evaluating diagnostic tests. *Ann. Intern. Med.* 1994;120(8):667-76.

[34] Minelli C, Thompson JR, Abrams KR, Thakkinstian A, Attia J. The Quality of Meta-Analyses of Genetic Association Studies: A Review With Recommendations. *Am. J. Epidemiol.* 2009;170(11):1333-1343.

[35] Mokkink LB, Terwee CB, Stratford PW, Alonso J, Patrick DL, Riphagen I, et al. Evaluation of the methodological quality of systematic reviews of health status measurement instruments. *Qual. Life Res.* 2009;18(3):313-333.

[36] Light RJ, Pillemer DB. *The science of reviewing research*. Cambridge, MA: Harvard University Press, 1984.

[37] Philibert A, Loyce C, Makowski D. Assessment of the quality of meta-analysis in agronomy. *Agriculture Ecosystems & Environment* 2012;148:72-82.

[38] Crombie IK. *The pocket guide to critical appraisal: A handbook for health care professionals*. London: BMJ Publishing group, 1996.

[39] Unit PHR. Critical Appraisal Skills Programme (CASP). 10 questions to help you make sense of reviews. Oxford: Public Health Resource Unit, 2006.

[40] Glenny AM, Esposito M, Coulthard P, Worthington HV. The assessment of systematic reviews in dentistry. *European Journal of Oral Sciences* 2003;111(2):85-92.

[41] Beck CT. Use of meta-analysis as a teaching strategy in nursing research courses. *Journal of Nursing Education* 1997;36(2):87-90.

[42] Sheikh L, Johnston S, Thangaratinam S, Kilby MD, Khan KS. A review of the methodological features of systematic reviews in maternal medicine. *BMC Med.* 2007;5 , 2007. Article Number.

[43] Greenhalgh T. Papers that summarise other papers (systematic reviews and meta-analyses). *BMJ* 1997;315(7109):672-5.

[44] Lundh A, Sismondo S, Lexchin J, Busuioc OA, Bero L. Industry sponsorship and research outcome. *Cochrane Database of Systematic Reviews* 2012(12).

[45] FOCUS. FOCUS critical appraisal tool. London: The Royal College of Psychiatrists, 2001.

[46] Thacker SB, Peterson HB, Stroup DF. Metaanalysis for the obstetrician-gynecologist. *American Journal of Obstetrics & Gynecology* 1996;174(5):1403-7.

[47] Oxman AD, Cook DJ, Guyatt GH. Users' guides to the medical literature. VI. How to use an overview. Evidence-Based Medicine Working Group. *JAMA* 1994;272(17):1367-71.

[48] Smith V, Devane D, Begley CM, Clarke M, Higgins S. A systematic review and quality assessment of systematic reviews of fetal fibronectin and transvaginal length for predicting preterm birth. *European Journal of Obstetrics Gynecology and Reproductive Biology* 2007;133(2):134-142.

[49] Nony P, Cucherat M, Haugh MC, Boissel JP. Critical reading of the meta-analysis of clinical trials. *Therapie* 1995;50(4):339-51.

[50] Oxman AD. Checklists for review articles. *BMJ* 1994;309(6955):648-51.

[51] Oxman AD, Guyatt GH. Guidelines for reading literature reviews. *CMAJ Canadian Medical Association Journal* 1988;138(8):697-703.

[52] Shea BJ, Bouter LM, Peterson J, Boers M, Andersson N, Ortiz Z, et al. External Validation of a Measurement Tool to Assess Systematic Reviews (AMSTAR). *Plos One* 2007;2(12).

[53] Sacks HS, Reitman D, Pagano D, Kupelnick B. Meta-analysis: an update. *Mount Sinai Journal of Medicine* 1996;63(3-4):216-24.

[54] Oxman AD, Guyatt GH, Singer J, Goldsmith CH, Hutchison BG, Milner RA, et al. Agreement among reviewers of review articles. *J. Clin. Epidemiol.* 1991;44(1):91-8.

[55] Anttila H, Samuelsson K, Salminen AL, Brandt S. Quality of evidence of assistive technology interventions for people with disability: An overview of systematic reviews. *Technology and Disability* 2012;24(1):9-48.

[56] Saokaew S, Oderda GM. Quality assessment of the methods used in published opioid conversion reviews. *Journal of Pain and Palliative Care Pharmacotherapy* 2012;26(4):341-347.

[57] Berkhof M, van Rijssen HJ, Schellart AJM, Anema JR, van der Beek AJ. Effective training strategies for teaching communication skills to physicians: An overview of systematic reviews. *Patient Educ. Couns.* 2011;84(2):152-162.

[58] List T, Axelsson S. Management of TMD: evidence from systematic reviews and meta-analyses. *J. Oral Rehabil.* 2010;37(6):430-451.

[59] Melchiors AC, Correr CJ, Venson R, Pontarolo R. An analysis of quality of systematic reviews on pharmacist health interventions. *Int. J. Clin. Phar,.* 2012;34(1):32-42.

[60] Minozzi S, Amato L, Davoli M. Development of dependence following treatment with opioid analgesics for pain relief: a systematic review. *Addiction* 2013;108(4):688-698.

[61] Wells C, Kolt GS, Marshall P, Hill B, Bialocerkowski A. Effectiveness of Pilates exercise in treating people with chronic low back pain: a systematic review of systematic reviews. *BMC Med. Res. Methodol.* 2013;13.

[62] Zwicker JG, Mayson TA. Effectiveness of Treadmill Training in Children With Motor Impairments: An Overview of Systematic Reviews. *Pediatr. Phys. Ther.* 2010;22(4):361-377.

[63] Oestergaard S, Moldrup C. Improving outcomes for patients with depression by enhancing antidepressant therapy with non-pharmacological interventions: A systematic review of reviews. *Public Health* 2011;125(6):357-367.

[64] Chan CL, Wang CW, Ho RT, Ng SM, Chan JS, Ziea ET, et al. A systematic review of the effectiveness of qigong exercise in supportive cancer care. [Review]. *Supportive Care in Cancer* 2012;20(6):1121-1133.

[65] Bouchard K, Dubuisson W, Simard J, Dorval M. Systematic mixed-methods reviews are not ready to be assessed with the available tools. *J. Clin. Epidemiol.* 2011;64(8):926-928.

[66] Cates Christopher J, Oleszczuk M, Stovold E, Wieland LS. Safety of regular formoterol or salmeterol in children with asthma: an overview of Cochrane reviews. *Cochrane Database of Systematic Reviews* 2012(10): John Wiley & Sons, Ltd. Available from: http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD010005.pub2/abstract http://onlinelibrary.wiley.com/store/10.1002/14651858.CD010005.pub2/asset/CD010005.pdf?v=1&t=hi49s2ql&s=adc394e4a58be3c737da3849ee2fb9675bc5d2e4

[67] Moore RA, Derry S, McQuay Henry J, Wiffen Philip J. Single dose oral analgesics for acute postoperative pain in adults. *Cochrane Database of Systematic Reviews* 2011(9): John Wiley & Sons, Ltd. Available from: http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD008659.pub2/abstract http://onlinelibrary.wiley.com/store/10.1002/14651858.CD008659.pub2/asset/CD008659.pdf?v=1&t=hi49v5c1&s=91dc0346ae734767382d48124c31621a05fdf1c8

[68] Jacobs WCH, Kruyt MC, Verbout AJ, Oner FC. Effect of methodological quality measures in spinal surgery research: a metaepidemiological study. *Spine Journal* 2012;12(4):339-348.

[69] Singh JA, Christensen R, Wells GA, Suarez-Almazor ME, Buchbinder R, Lopez-Olivo MA, et al. Biologics for rheumatoid arthritis: an overview of Cochrane reviews. *Cochrane Database of Systematic Reviews* 2009(4).

[70] Winters BD, Weaver SJ, Pfoh ER, Yang T, Pham JC, Dy SM. Rapid-Response Systems as a Patient Safety Strategy A Systematic Review. *Ann. Intern. Med.* 2013;158(5):417-+.

[71] Kang DY, Wu YX, Hu D, Hong Q, Wang JL, Zhang X. Reliability and External Validity of AMSTAR in Assessing Quality of TCM Systematic Reviews. *Evid.-based Complement Altern. Med.* 2012.

[72] McGee RG, Craig JC, Rogerson TE, Webster AC. Systematic reviews of surgical procedures in children: Quantity, coverage and quality. *Journal of Paediatrics and Child Health* 2013;49(4):319-324.

[73] Ma B, Guo JW, Qi GQ, Li HM, Peng JY, Zhang YL, et al. Epidemiology, Quality and Reporting Characteristics of Systematic Reviews of Traditional Chinese Medicine Interventions Published in Chinese Journals. *Plos One* 2011;6(5).

[74] Braga LH, Pemberton J, DeMaria J, Lorenzo AJ. Methodological Concerns and Quality Appraisal of Contemporary Systematic Reviews and Meta-Analyses in Pediatric Urology. *J. Urol.* 2011;186(1):266-271.

[75] Ma B, Qi GQ, Lin XT, Wang T, Chen ZM, Yang KH. Epidemiology, Quality, and Reporting Characteristics of Systematic Reviews of Acupuncture Interventions Published in Chinese Journals. *J. Altern. Complement Med.* 2012;18(9):813-817.

[76] MacDonald SL, Canfield SE, Fesperman SF, Dahm P. Assessment of the Methodological Quality of Systematic Reviews Published in the Urological Literature From 1998 to 2008. *J. Urol.* 2010;184(2):648-653.

[77] Panesar SS, Rao C, Vecht JA, Mirza SB, Netuveli G, Morris R, et al. Development of the Veritas plot and its application in cardiac surgery: an evidence-synthesis graphic tool for the clinician to assess multiple meta-analyses reporting on a common outcome. *Can. J. Surg.* 2009;52(5):E137-E145.

[78] Papageorgiou SN, Papadopoulos MA, Athanasiou AE. Evaluation of methodology and quality characteristics of systematic reviews in orthodontics. *Orthod. Craniofac. Res.* 2011;14(3):116-137.

[79] Popovich I, Windsor B, Jordan V, Showell M, Shea B, Farquhar CM. Methodological Quality of Systematic Reviews in Subfertility: A Comparison of Two Different Approaches. *Plos One* 2012;7(12).

[80] Seo HJ, Kim KU. Quality assessment of systematic reviews or meta-analyses of nursing interventions conducted by Korean reviewers. *BMC Med. Res. Methodol.* 2012;12.

[81] Sequeira-Byron P, Fedorowicz Z, Jagannath VA, Sharif MO. An AMSTAR assessment of the methodological quality of systematic reviews of oral healthcare interventions published in the Journal of Applied Oral Science (JAOS). *J. Appl. Oral Sci.* 2011;19(5):440-447.

[82] Torloni MR, Riera R. Design and level of evidence of studies published in two Brazilian medical journals recently indexed in the ISI Web of Science database. *Sao Paulo Med. J.* 2010;128(4):202-205.

[83] Weed DL, Althuis MD, Mink PJ. Quality of reviews on sugar-sweetened beverages and health outcomes: a systematic review. *Am. J. Clin. Nutr.* 2011;94(5):1340-1347.

[84] Jagannath V, Mathew JL, Asokan GV, Fedorowicz Z. Quality assessment of systematic reviews of health care interventions using AMSTAR. *Indian Pediatrics* 2011;48(5):383-385.

[85] Jin YH, Ma ET, Hua W, Dou HY. Reporting and methodological quality of systematic reviews and meta-analyses in nursing field in China. *Chinese Journal of Evidence-Based Medicine* 2012;12(9):1148-1155.

[86] Aziz T, Compton S, Nassar U, Matthews D, Ansari K, Flores-Mir C. Methodological quality and descriptive characteristics of prosthodontic-related systematic reviews. *J. Oral Rehabil.* 2013;40(4):263-278.

[87] Burda BU, Norris SL, Holmer HK, Ogden LA, Smith MEB. Quality varies across clinical practice guidelines for mammography screening in women aged 40-49 years as assessed by AGREE and AMSTAR instruments. *J. Clin. Epidemiol.* 2011;64(9):968-976.

[88] Chafen JJS, Newberry SJ, Riedl MA, Bravata DM, Maglione M, Suttorp MJ, et al. Diagnosing and Managing Common Food Allergies A Systematic Review. *JAMA-J. Am. Med. Assoc.* 2010;303(18):1848-1856.

[89] de Bot CMA, Moed H, Berger MY, Roder E, van Wijk RG, van der Wouden JC. Sublingual immunotherapy in children with allergic rhinitis: quality of systematic reviews. *Pediatr. Allergy Immunol.* 2011;22(6):548-558.

[90] Elangovan S, Avila-Ortiz G, Johnson GK, Karimbux NY, Allareddy V. Quality Assessment of Systematic Reviews on Periodontal Regeneration in Humans. *J. Periodont.* 2013;84(2):176-185.

[91] Faggion CM, Listl S, Giannakopoulos NN. The methodological quality of systematic reviews of animal studies in dentistry. *Vet. J.* 2012;192(2):140-147.

[92] Fleming PS, Seehra J, Polychronopoulou A, Fedorowicz Z, Pandis N. A PRISMA assessment of the reporting quality of systematic reviews in orthodontics. *Angle Orthod.* 2013;83(1):158-163.

[93] Littell JH, Shlonsky A. Making Sense of Meta-Analysis: A Critique of "Effectiveness of Long-Term Psychodynamic Psychotherapy". *Clin. Soc. Work J.* 2011;39(4):340-346.

[94] Suebnukarn S, Ngamboonsirisingh S, Rattanabanlang A. A Systematic Evaluation of the Quality of Meta-analyses in Endodontics. *J. Endod.* 2010;36(4):602-608.

[95] Andersen JH, Fallentin N, Thomsen JF, Mikkelsen S. Risk Factors for Neck and Upper Extremity Disorders among Computers Users and the Effect of Interventions: An Overview of Systematic Reviews. *Plos One* 2011;6(5).

[96] Chipps J, Brysiewicz P, Mars M. Effectiveness and feasibility of telepsychiatry in resource constrained environments? A systematic review of the evidence. *Afr. J. Psychiatry* 2012;15(4):235-243.

[97] Hagen KB, Dagfinrud H, Moe RH, Osteras N, Kjeken I, Grotle M, et al. Exercise therapy for bone and muscle health: an overview of systematic reviews. *BMC Med.* 2012;10.

[98] Li L, Ying XJ, Sun TT, Yi K, Tian HL, Sun R, et al. Overview of Methodological Quality of Systematic Reviews about Gastric Cancer Risk and Protective Factors. *Asian Pac. J. Cancer Prev.* 2012;13(5):2069-2079.

[99] Moe RH, Kjeken I, Uhlig T, Hagen KB. There Is Inadequate Evidence to Determine the Effectiveness of Nonpharmacological and Nonsurgical Interventions for Hand Osteoarthritis: An Overview of High-Quality Systematic Reviews. *Phys. Ther.* 2009;89(12):1363-1370.

[100] Vidal L, Gafter-Gvili A, Shpilberg O. Immunotherapy for Patients with Follicular Lymphoma: The Contribution of Systematic Reviews. *Acta Haematol.* 2011;125(1-2):23-31.

[101] Kung J, Chiappelli F, Cajulis OO, Avezova R, Kossan G, Chew L, 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. *The open dentistry journal* 2010;4:84-91.

[102] Coker TR, Chan LS, Newberry SJ, Limbos MA, Suttorp MJ, Shekelle PG, et al. Diagnosis, Microbial Epidemiology, and Antibiotic Treatment of Acute Otitis Media in Children A Systematic Review. *JAMA-J. Am. Med. Assoc.* 2010;304(19):2161-2169.

[103] Jaspers MWM, Smeulers M, Vermeulen H, Peute LW. Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. *J. Am. Med. Inf. Assoc.* 2011;18(3):327-334.

[104] Ryan R, Santesso N, Hill S, Lowe D, Kaufman C, Grimshaw J. Consumer-oriented interventions for evidence-based prescribing and medicines use: an overview of systematic reviews. *Cochrane Database of Systematic Reviews* 2011(5).

[105] Wiysonge CS, Ngcobo NJ, Jeena PM, Madhi SA, Schoub BD, Hawkridge A, et al. Advances in childhood immunisation in South Africa: where to now? Programme managers' views and evidence from systematic reviews. *BMC Public Health* 2012;12.

[106] Lougheed MD, Lemiere C, Ducharme FM, Licskai C, Dell SD, Rowe BH, et al. Canadian Thoracic Society 2012 guideline update: Diagnosis and management of asthma in preschoolers, children and adults. *Can. Respir. J.* 2012;19(2):127-164.

[107] Dent R, Blackmore A, Peterson J, Habib R, Kay GP, Gervais A, et al. Changes in Body Weight and Psychotropic Drugs: A Systematic Synthesis of the Literature. *Plos One* 2012;7(6).

[108] Matheson SL, Shepherd AM, Laurens KR, Carr VJ. A systematic meta-review grading the evidence for non-genetic risk factors and putative antecedents of schizophrenia. *Schizophr. Res.* 2011;133(1-3):133-142.

[109] Sakzewski L, Ziviani J, Boyd R. Systematic Review and Meta-analysis of Therapeutic Management of Upper-Limb Dysfunction in Children With Congenital Hemiplegia. *Pediatrics* 2009;123(6):E1111-E1122.

[110] Matjasko JL, Vivolo-Kantor AM, Massetti GM, Holland KM, Holt MK, Dela Cruz J. A systematic meta-review of evaluations of youth violence prevention programs: Common and divergent findings from 25 years of meta-analyses and systematic reviews. *Aggress. Violent Behav.* 2012;17(6):540-552.

[111] Monasta L, Batty GD, Cattaneo A, Lutje V, Ronfani L, van Lenthe FJ, et al. Early-life determinants of overweight and obesity: a review of systematic reviews. *Obes. Rev.* 2010;11(10):695-708.

[112] Windsor B, Popovich I, Jordan V, Showell M, Shea B, Farquhar C. Methodological quality of systematic reviews in subfertility: a comparison of Cochrane and non-Cochrane systematic reviews in assisted reproductive technologies. *Hum. Reprod.* 2012;27(12):3460-3466.

[113] Shepherd AM, Laurens KR, Matheson SL, Carr VJ, Green MJ. Systematic meta-review and quality assessment of the structural brain alterations in schizophrenia. *Neurosci. Biobehav. Rev.* 2012;36(4):1342-1356.

[114] Faggion CM, Schmitter M. Using the Best Available Evidence to Support Clinical Decisions in Implant Dentistry. *Int. J. Oral Maxillofac. Implants* 2010;25(5):960-969.

[115] Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* 1999;282(11):1054-60.

[116] 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.

[117] Amato L, Minozzi S, Davoli M. Efficacy and safety of pharmacological interventions for the treatment of the Alcohol Withdrawal Syndrome. *Cochrane Database of Systematic Reviews* 2011(6).

[118] Brouwers MC, Garcia K, Makarski J, Daraz L, Evidence Expert P, Res KTCCCP. The landscape of knowledge translation interventions in cancer control: What do we know and where to next? A review of systematic reviews. *Implement. Sci.* 2011;6.

[119] Chan RJ, Larsen E, Chan P. Re-examining the Evidence in Radiation Dermatitis Management Literature: An Overview and a Critical Appraisal of Systematic Reviews. *Int. J. Radiat. Oncol. Biol. Phys.* 2012;84(3):E357-E362.

[120] Flodgren G, Eccles Martin P, Shepperd S, Scott A, Parmelli E, Beyer Fiona R. An overview of reviews evaluating the effectiveness of financial incentives in changing healthcare professional behaviours and patient outcomes. *Cochrane Database of Systematic Reviews* 2011(7): John Wiley & Sons, Ltd. Available from: http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD009255/abstract http://onlinelibrary.wiley.com/store/10.1002/14651858.CD009255/asset/CD009255.pdf?v=1&t=hi49shar&s=a3723700a68fc3f408b821fc88500efa74e6e215

[121] Friedman AJ, Cosby R, Boyko S, Hatton-Bauer J, Turnbull G. Effective Teaching Strategies and Methods of Delivery for Patient Education: A Systematic Review and Practice Guideline Recommendations. *J. Cancer Educ.* 2011;26(1):12-21.

[122] Hopton A, MacPherson H. Acupuncture for Chronic Pain: Is Acupuncture More than an Effective Placebo? A Systematic Review of Pooled Data from Meta-analyses. *Pain Pract.* 2010;10(2):94-102.

[123] Johnson R, Bryant S, Huntley AL. Green tea and green tea catechin extracts: An overview of the clinical evidence. *Maturitas* 2012;73(4):280-287.

[124] Jones L, Othman M, Dowswell T, Alfirevic Z, Gates S, Newburn M, et al. Pain management for women in labour: an overview of systematic reviews. *Cochrane Database of Systematic Reviews* 2012(3): John Wiley & Sons, Ltd. Available from: http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD009234.pub2/abstract

[125] Kamioka H, Tsutani K, Okuizumi H, Mutoh Y, Ohta M, Handa S, et al. Effectiveness of Aquatic Exercise and Balneotherapy: A Summary of Systematic Reviews Based on Randomized Controlled Trials of Water Immersion Therapies. *J. Epidemiol.* 2010;20(1):2-12.

[126] Kim TH, Lim HJ, Kim MS, Lee MS. Dietary supplements for benign prostatic hyperplasia: An overview of systematic reviews. *Maturitas* 2012;73(3):180-185.

[127] Kumar A, Galeb S, Djulbegovic B. Treatment of Patients with Multiple Myeloma: An Overview of Systematic Reviews. *Acta Haematol.* 2011;125(1-2):8-22.

[128] Leucht S, Hierl S, Kissling W, Dold M, Davis JM. Putting the efficacy of psychiatric and general medicine medication into perspective: review of meta-analyses. *Br. J. Psychiatry* 2012;200(2):97-106.

[129] Li YK, White C, Ignacy TA, Thoma A. Comparison of Trapeziectomy and Trapeziectomy with Ligament Reconstruction and Tendon Interposition: A Systematic Literature Review. *Plast. Reconstr. Surg.* 2011;128(1):199-207.

[130] Mahtani KR, Protheroe J, Slight SP, Demarzo MMP, Blakeman T, Barton CA, et al. Can the London 2012 Olympics 'inspire a generation' to do more physical or sporting activities? An overview of systematic reviews. *BMJ Open* 2013;3(1).

[131] Michiels B, Govaerts F, Remmen R, Vermeire E, Coenen S. A systematic review of the evidence on the effectiveness and risks of inactivated influenza vaccines in different target groups. *Vaccine* 2011;29(49):9159-9170.

[132] Mickenautsch S, Yengopal V. Extent and quality of systematic review evidence related to minimum intervention in dentistry: essential oils, powered toothbrushes, triclosan, xylitol. *Int. Dent. J.* 2011;61(4):179-192.

[133] Mikton C, Butchart A. Child maltreatment prevention: a systematic review of reviews. *Bull. World Health Organ.* 2009;87(5):353-361.

[134] Parker JA, Barroso F, Stanworth SJ, Spiby H, Hopewell S, Doree CJ, et al. Gaps in the evidence for prevention and treatment of maternal anaemia: a review of systematic reviews. *BMC Pregnancy Childbirth* 2012;12.

[135] Payne C, Wiffen Philip J, Martin S. Interventions for fatigue and weight loss in adults with advanced progressive illness. *Cochrane Database of Systematic Reviews* 2012(1): John Wiley & Sons, Ltd. Available from: http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD008427.pub2/abstract http://onlinelibrary.wiley.com/store/10.1002/14651858.CD008427.pub2/asset/CD008427.pdf?v=1&t=hi49vohs&s=651f27b71c519dbf7e97d46b9e9b69acbb27af77

[136] Prior M, Guerin M, Grimmer-Somers K. The effectiveness of clinical guideline implementation strategies - a synthesis of systematic review findings. *J. Eval. Clin. Pract.* 2008;14(5):888-897.

[137] Rookmoneea M, Dennis L, Brealey S, Rangan A, White B, McDaid C, et al. The effectiveness of interventions in the management of patients with primary frozen shoulder. *J. Bone Joint Surg.-Br. Vol.* 2010;92B(9):1267-1272.

[138] Savard LA, Thompson DR, Clark AM. A meta-review of evidence on heart failure disease management programs: the challenges of describing and synthesizing evidence on complex interventions. *Trials* 2011;12.

[139] Singh JA, Christensen R, Wells GA, Suarez-Almazor ME, Buchbinder R, Lopez-Olivo MA, et al. A network meta-analysis of randomized controlled trials of biologics for rheumatoid arthritis: a Cochrane overview. *Can. Med. Assoc. J.* 2009;181(11):787-796.

[140] Spearing NM, Connelly LB. Is compensation "bad for health"? A systematic meta-review. *Injury-Int. J. Care Inj.* 2011;42(1):15-24.

[141] van der Linde RM, Stephan BCM, Savva GM, Dening T, Brayne C. Systematic reviews on behavioural and psychological symptoms in the older or demented population. *Alzheimers Res. Ther.* 2012;4(4).

Appendix 1: Steering Group Members

Name	Organisation
Deborah Caldwell	CMIMG
Rachel Churchill	University of Bristol
Philippa Davies	University of Bristol
Julian Higgins	Several Cochrane entities
Jos Kleijnen	Kleijnen Systematic Reviews
Toby Lasserson	Cochrane Library
Jelena Savović	Cochrane BMG
Beverley Shea	AMSTAR development
David Tovey	Cochrane Library
George Wells	AMSTAR development
Penny Whiting	Kleijnen Systematic Reviews/University of Bristol

Appendix 2: Face-to-Face meeting attendees

Name	Organisation
Patrick Bossuyt	Cochrane Diagnostic Methods Group
Kay Dickersin	US Cochrane Centre
Kerry Dwan	COMET
Julie Glanville	YHEC/Searching Methods
Julia Kreis	IQWIG
Silvia Minozzi	Overview Review Author
Carl Moons	Prognostic Methods Group
Matthew Page	Australasian Cochrane Centre
Barney Reeves	Cochrane Non-randomised Studies Group
Nancy Santesso	GRADE
Lesley Stewart	CRD
Peter Tugwell	Musculoskeletal disorders/J Clin Epi
Meera Viswanathan	RTI

Other contributors to the development of ROBIS

Phil Alderson	NICE
Zarko Alfirevic	Pregnancy and Childbirth Group/Review Group Cochrane Co-editor
Doug Altman	Prognostic Methods Group
Laura Amato	Overview Review Author
Lars Beckmann	IQWiG
Jon Deeks	Diagnostic Methods Group
Brian Hutton	CADTH
Fergus Macbeth	NICE/Cochrane
Alan Pearson	Joanna Briggs Institute
Therese Pigott	Campbell Collaboration
Stefan Sauerland	IQWiG
Chris Schmid	Institute of Medicine Standards
Jonathan Sterne	Bias Methods Group
Lesley Stewart	CRD

Appendix 3: Search strategy Review 2

The following databases were searched: MEDLINE (September 2007 to May 2013), EMBASE (2007 to May 2013), The Cochrane Methodology Register (2007 to 2013), and BIOSIS (2007 to 2013). Date of search 8th May 2013. The search strategies are below.

Database: Medline 1950 to present
Search Strategy:
---

1 "Review Literature as Topic"/ (4621)
2 meta-analysis/ (39407)
3 meta-analysis as topic/ (12660)
4 systematic review$.tw. (36763)
5 (meta-analys$ or metaanalys$).tw. (45463)
6 or/1-5 (89852)
7 Checklist/ (1571)
8 Quality Control/ (38779)
9 Guidelines as Topic/ (29619)
10 Total Quality Management/ (11672)
11 Reference Standards/ (31676)
12 or/7-11 (108883)
13 ((tool or tools or instrument$ or checklist$ or check list$ or scale or scales) and (quality or methodolog$ or method or methods)).ti. (5379)
14 (quality adj10 (score or scores or scoring or rating or rate) adj5 (methodolog$ or method or methods)).tw. (863)
15 (guideline$ and (quality or methodolog$ or method or methods)).ti. (1649)
16 ((assess$ or apprais$ or critical$) adj3 (systematic review$ or meta-analys$ or metaanalys$)).ti. (326)
17 ((score or scores or scoring or rating or rate) and (quality or methodolog$ or method or methods)).ti. (4218)
18 ((quality or methodology) adj3 (review or meta-analys$ or metaanalys$) adj3 (assess$ or method$)).tw. (1463)
19 (quality adj3 article$).tw. (1093)
20 (critical$ adj2 (apprais$ or evaluat$)).tw. (13506)
21 ((apprais$ or evaluat$) adj3 (systematic review$ or meta-analys$ or metaanalys$)).tw. (2094)
22 (guideline$ adj3 (systematic review$ or meta-analys$ or metaanalys$)).tw. (628)
23 or/13-22 (30168)
24 12 or 23 (137307)
25 Publication Bias/ (2158)
26 exp "bias (epidemiology)"/ (48129)
27 "Reproducibility of Results"/ (250918)
28 "Review Literature as Topic"/ (4621)
29 meta-analysis as topic/ (12660)
30 (bias adj3 (systematic review$ or meta-analys$ or metaanalys$)).tw. (268)
31 ((quality or bias or methodolog$) adj3 (systematic review$ or meta-analys$ or metaanalys$)).tw. (1543)
32 or/25-31 (299954)
33 6 and 24 and 32 (2171)
34 (200709$ 20071$ or 2008$ or 2009$ or 201$).ed. (4049446)
35 33 and 34 (1165)

Database: Embase <1980 to 2013 Week 18>
Search Strategy:

---
1 "meta analysis (topic)"/ (7056)
2 meta analysis/ (70619)
3 "systematic review (topic)"/ (3730)
4 systematic review/ (59703)
5 systematic review$.tw. (52119)
6 (meta-analys$ or metaanalys$).tw. (65855)
7 or/1-6 (149246)
8 "Review Literature as Topic"/ (44922)
9 (bias adj3 (systematic review$ or meta-analys$ or metaanalys$)).tw. (389)
10 ((quality or bias or methodolog$) adj3 (systematic review$ or meta-analys$ or metaanalys$)).tw. (2313)
11 "internal validity"/ (1268)
12 publishing/ (29591)
13 reproducibility/ (137306)
14 "systematic review (topic)"/ (3730)
15 "meta analysis (topic)"/ (7056)
16 or/8-15 (223210)
17 ((tool or tools or instrument$ or checklist$ or check list$ or scale or scales) and (quality or methodolog$ or method or methods)).ti. (7261)
18 (quality adj10 (score or scores or scoring or rating or rate) adj5 (methodolog$ or method or methods)).tw. (1387)
19 (guideline$ and (quality or methodolog$ or method or methods)).ti. (2176)
20 ((assess$ or apprais$ or critical$) adj3 (systematic review$ or meta-analys$ or metaanalys$)).ti. (451)
21 ((score or scores or scoring or rating or rate) and (quality or methodolog$ or method or methods)).ti. (5390)
22 ((quality or methodology) adj3 (review or meta-analys$ or metaanalys$) adj3 (assess$ or method$)).tw. (2108)
23 (quality adj3 article$).tw. (1473)
24 (critical$ adj2 (apprais$ or evaluat$)).tw. (17103)
25 ((apprais$ or evaluat$) adj3 (systematic review$ or meta-analys$ or metaanalys$)).tw. (3067)
26 (guideline$ adj3 (systematic review$ or meta-analys$ or metaanalys$)).tw. (917)
27 or/17-26 (39931)
28 checklist/ (5501)
29 quality control/ (108799)
30 total quality management/ (19016)
31 standard/ (334076)
32 or/28-31 (443558)
33 27 or 32 (477942)
34 7 and 16 and 33 (2913)
36 or/8-15 (223210)
37 35 and 36 (2913)
38 limit 37 to embase (1984)
39 (2007$ or 2008$ or 2009$ or 201$).em. (6895566)
40 38 and 39 (1609)
41 limit 40 to conference abstract (412) [note: downloaded separately]
42 40 not 41 (1197)

The Cochrane Methodology Register on The Cochrane Library
#1 ((tool or tools or instrument* or checklist* or check list* or scale or scales) and (quality or methodolog* or method or methods)):ti (Word variations have been searched)
#2 (guideline* and (quality or methodolog* or method or methods)):ti

#3 ((assess* or apprais* or critical*) near/3 (systematic review* or meta-analys* or metaanalys*)):ti
#4 ((score or scores or scoring or rating or rate) and (quality or methodolog* or method or methods)):ti
#5 (quality near/10 (score or scores or scoring or rating or rate) near/10 (methodolog* or method or methods))
#6 ((quality or methodology) near/3 (review or meta-analys* or metaanalys*) near/3 (assess* or method*))
#7 (quality near/3 article*)
#8 (critical* near/2 (apprais* or evaluat*))
#9 (apprais* or evaluat*) near/3 (systematic review* or meta-analys* or metaanalys*)
#10 (guideline* near/3 (systematic review* or meta-analys* or metaanalys*))
#11 #1 or #2 or #3 or #4 or #5 or #6 or #7 or #8 or #9 or #10
#12 ((quality or bias or methodolog*) near/3 (systematic review* or meta-analys* or metaanalys*))
#13 (bias near/3 (systematic review* or meta-analys* or metaanalys*))
#14 "publication bias"
#15 reproducibility
#16 "publishing"
#17 #12 or #13 or #14 or #15 or #16
#18 #11 and #17
#19 "systematic review*" or meta-analys* or metaanalys* from 2007 to 2013, in Methods Studies (Word variations have been searched)
#20 #18 and #19

BIOSIS on ISI Web of Knowledge
15 1,357 #14 AND #13 AND #1
14 TS=(bias or reproducibility or quality or methodology)
13 #12 or #11 or #10 OR #9 OR #8 OR #7 OR #6 OR #5 OR #4 OR #3 OR #2
12 TS=((tool NEAR/5 method*) or (tools NEAR/5 method*) or (instrument* NEAR/5 method*) or (checklist* NEAR/5 method*) or (“check list*” NEAR/5 method*) or (scale NEAR/5 method*) or (scales NEAR/5 method*))
11 TS=((tool NEAR/5 methodology) or (tools NEAR/5 methodology) or (instrument* NEAR/5 methodology) or (checklist* NEAR/5 methodology) or (“check list*” NEAR/5 methodology) or (scale NEAR/5 methodology) or (scales NEAR/5 methodology))
10 TS=((tool NEAR/5 quality) or (tools NEAR/5 quality) or (instrument* NEAR/5 quality) or (checklist* NEAR/5 quality) or ("check list*" NEAR/5 quality) or (scale NEAR/5 quality) or (scales NEAR/5 quality))
9 TS=(methodology NEAR/3 “systematic review*” or methodology NEAR/3 meta-analys* or methodology NEAR/3 metaanalys*)
8 TS=(quality NEAR/3 “systematic review*” or quality NEAR/3 meta-analys* or quality NEAR/3 metaanalys*)
7 TS=(quality NEAR/3 article*)
6 TS=(critical* NEAR/2 apprais* or critical* NEAR/2 evaluat*)
5 TS=(evaluat* NEAR/3 “systematic review*” or evaluat* NEAR/3 meta-analys* or evaluat* NEAR/3 metaanalys*)
4 TS=((apprais* NEAR/3 “systematic review*” or apprais* NEAR/3 meta-analys* or apprais* NEAR/3 metaanalys*))
3 TS=((guideline* NEAR/3 "systematic review*") or (guideline* NEAR/3 meta-analys*) or (guideline* NEAR/3 metaanalys*)).
2 TI=(guideline* and (quality or methodolog* or method or methods))
1 TS=("systematic review*" or meta-analys* or metaanalys*)

Appendix 4: MECIR coding

Item No.	Status	Item name	Standard	Classification
Setting the research question (s) to inform the scope of the review				
C1	Mandatory	Formulating review questions	Ensure that the review question and particularly the outcomes of interest, address issues that are important to stakeholders such as consumers, health professionals and policy makers.	Bias
C2	Mandatory	Pre-defining objectives	Define in advance the objectives of the review, including participants, interventions, comparators and outcomes.	Bias
C3	Mandatory	Considering potential adverse effects	Consider any important potential adverse effects of the intervention(s) and ensure that they are addressed.	Process
C4	Highly desirable	Considering equity and specific populations	Consider in advance whether issues of equity and relevance of evidence to specific populations are important to the review, and plan for appropriate methods to address them if they are. Attention should be paid to the relevance of the review question to populations such as low socioeconomic groups, low or middle income regions, women, children and older people.	Applicability
Setting eligibility criteria for including studies in the review				
C5	Mandatory	Pre-defining unambiguous criteria for participants	Define in advance the eligibility criteria for participants in the studies.	Bias
C6	Highly desirable	Pre-defining a strategy for studies with a subset of eligible participants	Define in advance how studies that include only a subset of relevant participants will be handled.	Bias
C7	Mandatory	Pre-defining unambiguous criteria for interventions and comparators	Define in advance the eligible interventions and the interventions against which these can be compared in the included studies.	Bias
C8	Mandatory	Clarifying role of outcomes	Clarify in advance whether outcomes listed under 'Criteria for considering studies for this review' are used as criteria for including studies (rather than as a list of the outcomes of interest within whichever studies are included).	Bias
C9	Mandatory	Pre-defining study designs	Define in advance the eligibility criteria for study designs in a clear and unambiguous way, with a focus on features of a study's design rather than design labels.	Bias

Item No.	Status	Item name	Standard	Classification
C10	Mandatory	Including randomized trials	Include randomized trials as eligible for inclusion in the review, if they are feasible for the interventions and outcomes of interest.	NR
C11	Mandatory	Justifying choice of study designs	Justify the choice of eligible study designs.	Process
C12	Mandatory	Excluding studies based on publication status	Include studies irrespective of their publication status, unless explicitly justified.	Bias
C13	Mandatory	Changing eligibility criteria	Justify any changes to eligibility criteria or outcomes studied. In particular, *post hoc* decisions about inclusion or exclusion of studies should keep faith with the objectives of the review rather than with arbitrary rules.	Bias
Selecting outcomes to be addressed for studies included in the review				
C14	Mandatory	Pre-defining outcomes	Define in advance which outcomes are primary outcomes and which are secondary outcomes.	Bias
C15	Highly desirable	Choosing outcomes	Keep the total number of outcomes selected for inclusion in the review as small as possible. Choose outcomes that are relevant to stakeholders such as consumers, health professionals and policy makers. Avoid trivial outcomes and biochemical, interim and process outcomes, but consider the importance of resource-use outcomes.	Process
C16	Highly desirable	Pre-defining outcome details	Define in advance details of what are acceptable outcome measures (e.g. diagnostic criteria, scales, composite outcomes).	Process
C17	Highly desirable	Pre-defining choices from multiple outcome measures	Define in advance how outcome measures will be selected when there are several possible measures (e.g. multiple definitions, assessors or scales).	Bias
C18	Highly desirable	Pre-defining time points of interest	Define in advance the timing of outcome measurement.	Bias
Planning the review methods at protocol stage				
C19	Mandatory	Planning the search	Plan in advance the methods to be used for identifying studies. Design searches to capture as many studies as possible meeting the eligibility criteria, ensuring that relevant time periods and sources are covered and not restricting by language or publication status.	Process

Item No.	Status	Item name	Standard	Classification
C20	Mandatory	Planning the assessment of risk of bias in included studies	Plan in advance the methods to be used for assessing risk of bias in included studies, including the tool(s) to be used, how the tool(s) will be implemented, and the criteria used to assign studies, for example, to judgements of low risk, high risk and unclear risk of bias.	Process
C21	Mandatory	Planning the synthesis of results	Plan in advance the methods to be used to synthesize the results of the included studies, including whether a quantitative synthesis is planned, how heterogeneity will be assessed, choice of effect measure (e.g. odds ratio, risk ratio, risk difference or dichotomous outcomes), and methods for meta-analysis (e.g. inverse variance or Mantel Haenszel, fixed-effect or random-effects model).	Process
C22	Mandatory	Planning subgroup analyses	Pre-define potential effect modifiers (e.g. for subgroup analyses) at the protocol stage; restrict these in number; and provide rationale for each.	Process
C23	Highly desirable	Planning a 'Summary of findings' table	Plan in advance the methods to be used for summarizing the findings of the review, including the assessment of the quality of the body of evidence. If a formal 'Summary of findings' table is anticipated, specify which outcomes will be included, and which comparisons and subgroups will be covered(if appropriate).	Process
Searching for studies				
C24	Mandatory	Searching key databases	Search the Cochrane Review Group's Specialized Register (internally, e.g. via the Cochrane Register of Studies, or externally via CENTRAL). Ensure that CENTRAL and MEDLINE (e.g. via PubMed) have been searched (either for the review or for the Review Group's Specialized Register).	Bias
C25	Highly desirable	Searching specialist bibliographic databases	Search appropriate national, regional and subject specific bibliographic databases.	
C26	Mandatory	Searching for different types of evidence	If the review has specific eligibility criteria around study design to address adverse effects, economic issues or qualitative research questions, undertake searches to address them.	Bias
C27	Mandatory	Searching trials registers	Search trials registers and repositories of results, where relevant to the topic through ClinicalTrials.gov, the WHO International Clinical Trials Registry Platform (ICTRP)	Bias

Item No.	Status	Item name	Standard	Classification
			portal and other sources as appropriate.	
C28	Highly desirable	Searching for grey literature	Search relevant grey literature sources such as reports/dissertations/theses databases and databases of conference abstracts.	Bias
C29	Highly desirable	Searching within other reviews	Search within previous reviews on the same topic.	Bias
C30	Mandatory	Searching reference lists	Check reference lists in included studies and any relevant systematic reviews identified.	Bias
C31	Highly desirable	Searching by contacting relevant individuals and organisations	Contact relevant individuals and organisations for information about unpublished or ongoing studies.	Bias
C32	Mandatory	Structuring search strategies for bibliographic databases	Inform the structure of search strategies in bibliographic databases around the main concepts of the review, using appropriate elements from PICO and study design. In structuring the search, maximize sensitivity whilst striving for reasonable precision. Ensure correct use of the AND and OR operators.	Bias
C33	Mandatory	Developing search strategies for bibliographic databases	Identify appropriate controlled vocabulary (e.g. MeSH, Emtree, including 'exploded' terms) and free-text terms (considering, for example, spelling variants, synonyms, acronyms, truncation and proximity operators).	Bias
C34	Highly desirable	Using search filters	Use specially designed and tested search filters where appropriate including the Cochrane Highly Sensitive Search Strategies for identifying randomized trials in MEDLINE, but do not use filters in pre-filtered databases e.g. do not use a randomized trial filter in CENTRAL or a systematic review filter in DARE.	Bias
C35	Mandatory	Restricting database searches	Justify the use of any restrictions in the search strategy on publication date, publication format or language.	Bias
C36	Mandatory	Documenting the search process	Document the search process in enough detail to ensure that it can be reported correctly in the review.	Process
C37	Mandatory	Rerunning searches	Rerun or update searches for all relevant databases within 12 months before publication of the review or review update, and screen the results for potentially	Applicability

Item No.	Status	Item name	Standard	Classification
			eligible studies.	
C38	Highly desirable	Incorporating findings from rerun searches	Incorporate fully any studies identified in the rerun or update of the search within 12 months before publication of the review or review update.	Applicability
Selecting studies into the review				
C39	Mandatory	Making inclusion decisions	Use (at least) two people working independently to determine whether each study meets the eligibility criteria, and define in advance the process for resolving disagreements.	Bias
C40	Mandatory	Excluding studies without useable data	Include studies in the review irrespective of whether measured outcome data are reported in a 'usable' way.	Bias
C41	Mandatory	Documenting decisions about records identified	Document the selection process in sufficient detail to complete a PRISMA flow chart and a table of 'Characteristics of excluded studies'.	Process
C42	Mandatory	Collating multiple reports	Collate multiple reports of the same study, so that each study rather than each report is the unit of interest in the review.	Bias
Collecting data from included studies				
C43	Mandatory	Using data collection forms	Use a data collection form, which has been piloted.	Process
C44	Mandatory	Describing studies	Collect characteristics of the included studies in sufficient detail to populate a table of 'Characteristics of included studies'.	Process
C45	Highly desirable	Extracting study characteristics in duplicate	Use (at least) two people working independently to extract study characteristics from reports of each study, and define in advance the process for resolving disagreements.	Bias
C46	Mandatory	Extracting outcome data in duplicate	Use (at least) two people working independently to extract outcome data from reports of each study, and define in advance the process for resolving disagreements.	
C47	Mandatory	Making maximal use of data	Collect and utilize the most detailed numerical data that might facilitate similar analyses of included studies. Where 2×2 tables or means and standard deviations are not available, this might include effect estimates (e.g. odds ratios, regression coefficients), confidence intervals, test statistics (e.g. t, F, Z, chi-squared) or P values, or even data for individual participants.	Process

Item No.	Status	Item name	Standard	Classification
C48	Highly desirable	Examining errata	Examine any relevant retraction statements and errata for information.	Process
C49	Highly desirable	Obtaining unpublished data	Seek key unpublished information that is missing from reports of included studies.	Bias
C50	Mandatory	Choosing intervention groups in multi- arm studies.	*If a study is included with more than two intervention arms*, include in the review only intervention and control groups that meet the eligibility criteria.	Process
C51	Mandatory	Checking accuracy of numeric data in the review.	Compare magnitude and direction of effects reported by studies with how they are presented in the review, taking account of legitimate differences.	Process
Assessing risk of bias in included studies				
C52	Mandatory	Assessing risk of bias	Assess the risk of bias for each included study. For randomized trials, the Cochrane 'Risk of bias' tool should be used, involving judgements and supports for those judgements across a series of domains of bias, as described in Chapter 8 of the Cochrane Handbook (version 5 or later).	Bias
C53	Mandatory	Assessing risk of bias in duplicate	Use (at least) two people working independently to apply the risk of bias tool to each included study, and define in advance the process for resolving disagreements.	Bias
C54	Mandatory	Supporting judgements of risk of bias	Justify judgements of risk of bias (high, low and unclear) and provide this information in the 'Risk of bias' tables (as 'Support for judgement').	Process
C55	Highly desirable	Providing sources of information for risk of bias assessments	Collect the source of information for each risk of bias judgement (e.g. quotation, summary of information from a trial report, correspondence with investigator etc.).Where judgements are based on assumptions made on the basis of information provided outside publicly available documents, this should be stated.	Process
C56	Highly desirable	Differentiating between performance bias and detection bias.	Consider separately the risks of bias due to lack of blinding for (i) participants and study personnel (performance bias), and (ii) outcome assessment (detection bias).	Process
C57	Highly desirable	Assessing risk of bias due to lack of blinding for different outcomes	Consider blinding separately for different key outcomes.	Process
C58	Highly desirable	Assessing completeness of data for different outcomes	Consider the impact of missing data separately for different key outcomes to which an included study contributes data.	Process

Item No.	Status	Item name	Standard	Classification
C59	Highly desirable	Summarizing risk of bias assessments	Summarize the risk of bias for each key outcome for each study.	Process
C60	Highly desirable	Addressing risk of bias in the synthesis	Address risk of bias in the synthesis (whether qualitative or quantitative). For example, present analyses stratified according to summary risk of bias, or restricted to studies at low risk of bias.	Bias
C61	Mandatory	Incorporating assessments of risk of bias	*If randomized trials have been assessed using one or more tools in addition to the Cochrane 'Risk of bias' tool*, use the Cochrane tool as the primary assessment of bias for interpreting results, choosing the primary analysis, and drawing conclusions.	Process
Synthesizing the results of included studies				
C62	Mandatory	Combining different scales	If studies are combined with different scales, ensure that higher scores for continuous outcomes all have the same meaning for any particular outcome; explain the direction of interpretation; and report when directions were reversed.	Process
C63	Mandatory	Ensuring meta- analyses are meaningful	Undertake (or display) a meta-analysis only if participants, interventions, comparisons and outcomes are judged to be sufficiently similar to ensure an answer that is clinically meaningful.	Bias
C64	Mandatory	Assessing statistical heterogeneity	Assess the presence and extent of between- study variation when undertaking a meta- analysis.	Bias
C65	Highly desirable	Addressing missing outcome data	Consider the implications of missing outcome data from individual participants (due to losses to follow up or exclusions from analysis).	Bias
C66	Highly desirable	Addressing skewed data	Consider the possibility and implications of skewed data when analysing continuous outcomes.	Bias
C67	Mandatory	Addressing studies with more than two groups	*If multi-arm studies are included*, analyse multiple intervention groups in an appropriate way that avoids arbitrary omission of relevant groups and double-counting of participants.	Bias
C68	Mandatory	Comparing subgroups	*If subgroup analyses are to be compared*, and there are judged to be sufficient studies to do this meaningfully, use a formal statistical test to compare them.	Bias
C69	Mandatory	Interpreting subgroup analyses	*If subgroup analyses are conducted*, follow the subgroup analysis plan specified in the	Bias

Item No.	Status	Item name	Standard	Classification
			protocol without undue emphasis on particular findings.	
C70	Mandatory	Considering statistical heterogeneity when interpreting the results	Take into account any statistical heterogeneity when interpreting the results, particularly when there is variation in the direction of effect.	Bias
C71	Mandatory	Addressing non- standard designs	Consider the impact on the analysis of clustering, matching or other non-standard design features of the included studies.	Bias
C72	Highly desirable	Sensitivity analysis	Use sensitivity analyses to assess the robustness of results, such as the impact of notable assumptions, imputed data, borderline decisions and studies at high risk of bias.	Bias
C73	Mandatory	Interpreting results	Interpret a statistically non-significant P value (e.g. larger than 0.05) as a finding of uncertainty unless confidence intervals are sufficiently narrow to rule out an important magnitude of effect.	Process
C74	Highly desirable	Investigating reporting biases	Consider the potential impact of reporting biases on the results of the review or the meta-analyses it contains.	Bias
Summarizing the findings				
C75	Highly desirable	Including a 'Summary of Findings' table	Include a 'Summary of Findings' table according to recommendations described in Chapter 10 of the Cochrane Handbook (version 5 or later). Specifically: • include results for one population group (with few exceptions); • indicate the intervention and the comparison intervention; • include seven or fewer patient-important outcomes; • describe the outcomes (e.g. scale, scores, follow-up); • indicate the number of participants and studies for each outcome; • present at least one baseline risk for each dichotomous outcome (e.g. study • population or median/medium risk) and • baseline scores for continuous outcomes (if appropriate);	Process

Item No.	Status	Item name	Standard	Classification
			• summarize the intervention effect (if appropriate); and • include a measure of the quality of the body of evidence.	
C76	Mandatory	Assessing the quality of the body of evidence	Use the five GRADE considerations (study limitations, consistency of effect, imprecision, indirectness and publication bias) to assess the quality of the body of evidence for each outcome, and to draw conclusions about the quality of evidence within the text of the review.	Process
C77	Mandatory	Justifying assessments of the quality of the body of evidence	Justify and document all assessments of the quality of the body of evidence (for example downgrading or upgrading if using the GRADE tool).	Bias
Reaching conclusions				
C78	Mandatory	Formulating implications for practice	Base conclusions only on findings from the synthesis (quantitative or narrative) of studies included in the review.	Bias
C79	Mandatory	Avoiding recommendations	Avoid providing recommendations for practice.	NR
C80	Highly desirable	Formulating implications for research	Structure the implications for research to address the nature of evidence required, including population intervention comparison, outcome, and type of study.	NR

Items highlighted red are those coded as relating to bias.

Appendix 5: Data extraction tables Review 2

Review of existing quality assessment tools

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Assendelft(1995)[24] **Tool Name:** No name **Study designs targeted:** RCT of spinal manipulation **Item rating:** Numerical scoring system - maximum score included after each item **Tool development:** Used a list of standardized criteria based on previous work of Oxman and Guyatt, Light and Pillemer and Mulrow. **IRR:** Reviewers agreed on 95% of all items in the tool.	*A. Description of inclusion and exclusion criteria:* Description of study setting(s) included (i.e., industry, general practice, hospital), intervention type(s) included (manipulation, mobilization, or both; specific techniques or professions [chiropractors, osteopaths, Cyriax technique]), outcome type(s) included (pain, global assessment, mobility, functional status, time until recovery, medical consumption),years covered, language(s) covered (10)	*B. Search strategy:* Established bibliographical database included (e.g., Index Medicus, EMBASE), additional efforts to locate non-indexed RCTs (e.g., citation tracking, correspondence with experts, manual search of non-indexed journals) (5 points).	*D. Assessment of the validity RCTs:* Assessment per RCT included that is explicit (reproducible by readers of the review) regarding the similarity of treatment groups, similarity of treatment characteristics, adequacy in treatment of missing values (dropouts, loss to follow-up), success regarding blinding of outcome assessment, relevance of outcome measures, (12) *E. Number of reviewers:* At least two independent reviewers [methodological quality assessment] (4) *F. Blinding of reviewers*: Reviewer(s) blinded for at least the outcomes of the RCTs [methodological quality assessment] (2) G. *Agreement of reviewer(s):* Agreement between reviewers reported (quantitative) and acceptable [methodological quality assessment] (2) *H. Description of manipulative intervention(s)* (8) *I. Description of control intervention(s)* (7)	*C. Emphasis on randomized clinical trials (RCTs):* Randomized clinical trials only, or results of RCTs discussed separately from other study designs (10 points). (10) *J. Outcome presentation* (14) *K. Statistical pooling:* Statistical pooling of the most important outcome(s) or discussion of the reason why pooling is not indicated or warranted or pooling of the subset considered to be valid and similar enough (3) *L. Discussion power of negative RCTs:* Elaboration on the power of negative RCTs: calculation of the power of each negative RCT or narrative elaboration on the power of each negative RCT or overall narrative elaboration on the power of the negative RCTs (3)	*M. Overall conclusion:* Overall conclusion on the aggregated level of available RCTs on the effectiveness of manipulation presented (5) *N. Discussion of heterogeneity of RCTs and outcomes:* Identification of relevant subgroups (eg, study setting, disease classification) with explicit motivation (4 points). Discussion of the variety of treatment modalities in the intervention groups (eg, mobilization, manipulation; chiropractic, ostéopathie, Cyriax technique) (2 points). Discussion of the variety of treatment modalities in the control groups (placebo, existing modality) (2 points). Discussion of the relationship between the methodological quality of the RCTs and outcome (2 points). (10)

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Auperin(1997)[17] **Tool Name:** No Name **Study designs targeted:** Not specified **Item rating:** "For each item, 3 rating levels were available: `adequate' (2), ` partial (1) and `none or unknown' (0). Established a quality score as the sum of the rating of the 27 items, ranging from 0 to 54." **Tool development:** Modified version of Sacks [31] "...slightly modified version of the scoring method established by Sacks et al. We kept the 23 items of Sacks et al. and we added **four items** (contact with the investigators of the primary trials, intention-to-treat analysis design, carrying out of indirect analyses and discussion of the end-point quality)." **IRR:** "The intraclass correlation coefficient between the two scores given by each rater was 0.84. The intraclass correlation coefficients between the score given by each rater and the common score ranged from 0.89 - 0.96."	*1. Protocol*: presence of a protocol planned before the start of the study.	*2. Literature search*: several procedures completely described (not only computer searches).	*3. List of trials analysed*: list of analysed trials published. *4. Log of rejected trials:* reasons for exclusion and list of rejected trials published or available on request. *5. Selection method:* selection according to the methods used to perform the trials and blinded to the results. *7. Description of patients, treatments and diagnoses* : in each trial, mainly for treatment modalities and prognosis factors. *10. Trial quality* assessment: reported for each trial. *12. Data-extraction method and inter-observer agreement:* data extracted by more than one observer, blinded to the treatment groups and measure of the inter-observer agreement. ***13. Contact with trial investigators*** **: contact for all the trials.**	*6. Control of publication bias*: unpublished trials included or calculation of number of negative trials required to refute the meta-analysis result. *8. Clinical combinability criteria*: discussion of criteria used to decide whether trials were similar enough to be pooled. *9. Only randomized trials pooled*: main analysis performed with only randomized trials or with and without pseudo-randomized trials. ***11. Intention-to-treat analysis*****: analysis on all patients randomized (no withdrawal) for all the trials.** *14. Statistical methods*: referenced pooling method stratified for trials. ***15. End-point quality*****: relevant, objective and homogeneous.** *16. Sensitivity analysis*: analysis with varying end-points and statistical methods or with exclusion of some trials. *17. Subgroup analyses*: performed on the data of all trials. ***18. Indirect analyses:*** **test of interaction between predefined categories of trials and treatment effect.**	19. Specification of source of support: clear acknowledgement of source of support.

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Beck(1997)[41] **Tool Name:** Meta-analysis Appraisal Checklist **Study designs targeted:** Not specified **Item rating:** Yes, No and Comments **Tool development:** Author's own. **IRR:** Not reported	1. Were research questions identified? 2. Were specific hypotheses tested? 3. Did the researcher define criteria for the inclusion and exclusion of studies in the meta-analysis?	7. Were details of the search procedures provided? 8. Did the meta-analysts search for unpublished studies in order to test for a type 1 error publication bias? 9. Did the researcher avoid selecting studies based on criteria of methodological rigour age of study, or publication status?	4. Did the researcher enumerate the relevant studies which were excluded from the meta-analysis and the reasons for exclusion? 5. Were the study characteristics reported so that the nature and limits of the domain actually analysed can be understood? 6. Did the researcher publish or make available the final list of studies included in the meta-analysis? 10. Did the researcher develop and pilot test coding forms before coding characteristics for the meta-analysis?' 11. Did the meta-analyst develop a detailed, explicit codebook that was keyed to the coding forms? 12. Did the researcher measure and report inter-coder reliability as part of the meta-analysis?	13. Was a fail-safe N computed to decrease the likelihood of a type I publication bias error in finding more positive results than is really the case?' 14. What were the criteria which were used to decide that the studies were similar enough that they could be pooled? 15. Was weighting of studies by sample size or quality of study performed? 16. Were tests of homogeneity used to help identify those which represent outliers? 17. When a single study provided multiple results were separate meta-analyses for each type of dependent variable performed or were the different types of outcome measures combined in a single analysis? 18. Did the researcher examine multiple independent and dependent variables separately through blocking, mediating effects? 19. Were nonparametric measures of effect size used when appropriate, such as with ordinal or dichotomous data? 20. Did the researcher use more than one method of statistical pooling to provide multiple indicators for interpreting the results? 21. Were combined tests of significance accompanied with estimates of effect size? 22. Did the researcher provide an estimate with confidence intervals of the difference between the success rates of the interventions being compared rather than only the results of the significance tests?	23. Did the researcher consider alternative explanations for the results obtained? 25. Did the researcher limit generalizations of the findings to the domain specified by the meta-analysis? 26. Were the limitations of the meta-analysis identified? 27. Did the meta-analyst provide guidelines for future research concerning the relationship reviewed? 28. Was the complete study reported in enough detail to permit direct replication?

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Crombie(1996)[38] **Tool Name:** No name **Study designs targeted:** Not specified **Item rating:** Descriptive **Tool development:** Not reported **IRR:** Not reported **Comments:** 3 **essential questions** (a, b & c above) and 11 specific questions	1. Is the topic well defined?	**a. How were the papers identified? (essential question)**	3. Were the detailed study designs reviewed? 4. Was missing information sought? 5. Were the basic data adequately described?	**b. How was the quality of the papers assessed?** 2. Are the statistical methods described? 6. Was publication bias taken into account? 7. Was heterogeneity of effect investigated?	**c. How were the results summarised?** 8. What do the main findings mean? 9. Are there other findings which merit attention? 10. Are the conclusions justified? 11. How do the findings compare with previous reports?

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
FOCUS(2001)[45] **Tool Name:** FOCUS? **Study designs targeted:** Generic **Item rating:** Comments fields; descriptive **Tool development:** 'Adapted from material produced by the Centre for Evidence-Based Mental Health'. **IRR:** No information	*1. Is the question clearly focused?* (What is being reviewed? What is the population? What is the intervention/exposure? What are the outcomes?) *2. Is the search thorough? Did the authors look for the appropriate sort of papers?* (What sort of bibliographic databases were used? What years were searched? What languages were searched? Was any hand-searching conducted or references in relevant articles obtained? Are the inclusion criteria appropriate? Is the inclusion process discussed?)	*2. Is the search thorough? Did the authors look for the appropriate sort of papers?* (What sort of bibliographic databases were used? What years were searched? What languages were searched? Was any hand-searching conducted or references in relevant articles obtained? Are the inclusion criteria appropriate? Is the inclusion process discussed?	*3. Is the validity of included studies adequately assessed?* (reproducible, blind assessment? Method of random selection, is the analysis on an ITT basis? Is missing information obtained from investigators? Is publication bias an issue? Has quality been assessed?) *4. How many individual studies were included in the systematic review/meta-analysis?* (What type of studies were included? e.g. randomised controlled trials, cohort studies, case-control studies, etc., What are the sample sizes for each study group?, Were the patient characteristics, interventions, outcome measures and the efficacious and adverse results discussed/presented for each study? What were they? *5. In what countries were the treatment studies conducted?* *6. If medication was used, what were the dosages of medication used for each study?* *7. What was the duration of treatment (give the range)?* *8. Are the studies focused on boys or girls or both?* *9. Were the children receiving concomitant medication/treatment?*	*10. How big is the overall effect?* (On what scale is the effect measured? (odds ratio, number needed to treat?)) *11. Are the results consistent from study to study?* (How sensitive are the results to changes in the way the review was done?) *12. If the results of the review have been combined, was it reasonable to do so?* (Were the results similar from study to study? Are the results of the included studies clearly displayed? Are the results of the different studies similar? Are the reasons for any variations in results discussed?) *13. How precise are the results?* (Does the lower confidence limit include clinically relevant effects? Does the upper confidence limit exclude clinically relevant effects?)	*14. Do conclusions flow from evidence that is reviewed?* *15. Are subgroup analyses interpreted cautiously?* *17. Were all important outcomes considered?* *18. Are the benefits worth the harms and the costs?*

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Geller(1996)[30] **Tool Name:** No name **Study designs targeted:** RCTs **Item rating:** Not specified **Tool development:** No information. Reads like an educational article based on author opinion/ experience. **IRR:** No information.	*General considerations:* 1. Is the objective of the meta-analysis clearly stated? 2. Are the inclusion/exclusion criteria explicit?	*General considerations:* 3. Is the search mechanism for determination of suitable studies adequate?	*General considerations:* 4. Is the quality of the trials assessed? 5. Are all of the trials randomized?	*Statistical considerations:* 1. Is the analysis technically correct? 2. Is there adequate discussion concerning the combinability of trials (homogeneity)? 3. Is evidence presented that subgroup analyses were defined a priori? 4. Are there any graphics? Only tables? 5. Is some sensitivity analysis shown?	*General considerations:* 6. Does the discussion include mention of limitations? Put the results in context? 7. Are the conclusions justified by the data?

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Glenny(2003)[40] **Tool Name:** No name **Study designs targeted:** Dentistry intervention studies **Item rating:** Yes, no, can't tell and for some items not applicable **Tool development:** A quality assessment checklist was devised and piloted by all four reviewers on a sample of 10 reviews. The form was piloted on 10 reviews by all four reviewers. Areas of ambiguity were discussed and the assessment form revised. **IRR:** There were four raters, two clinicians and two methodologists. At least one clinician and one methodologist assessed each review. When studies were assessed by more than one either clinician or methodologist, the assessment used for analysis was selected at random. The per cent agreement was generally high, ranging across all assessments from 55% to 88%, with a median of 72%. Overall kappa values ranged from 0.06 to 0.81, with a median value of 0.46	1. Did review address a focused question? 2. Did authors look for appropriate papers?	3. Do you think authors attempted to identify all relevant studies? 4. Search for published and unpublished literature 5. Were all languages considered? 6. Was any hand-searching carried out?	7. Was it stated that the inclusion criteria were carried out by at least two reviewers? 8. Did reviewers attempt to assess the quality of the included studies? 10. Was it stated that the quality assessment was carried out by at least two reviewers?	9. If so did they include this in the analysis? (refers to quality assessment) 11. Are the results given in a narrative or pooled statistical analysis? 12. If the results have been combined was it reasonable to do so? 13. Are the results clearly displayed? 14. Was an assessment of heterogeneity made and reasons for variation discussed?	15. Were results of review interpreted appropriately?

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Greenhalgh(1997)[43] **Tool Name:** How to read a paper **Study designs targeted:** Not specified **Item rating:** Yes/No **Tool development:** Not described. Educational article by a sole author, so experience/opinion based. **IRR:** Not reported	Q 1: Can you find an important clinical question which the review addressed?	Q2: Was a thorough search done of the appropriate databases and were other potentially important sources explored?	Q3: Was methodological quality assessed and the trials weighted accordingly?	Q3: Was methodological quality assessed and the trials weighted accordingly? Q4: How sensitive are the results to the way the review has been done?	Q5: Have the numerical results been interpreted with common sense and due regard to the broader aspects of the problem?

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Higgins(In Press)[14] **Tool Name:** No Name **Study designs targeted:** RCTs **Item rating:** Summary questions: 'Yes', 'Probably Yes', 'Unsure', 'Probably No' and 'No'. Supportive questions as shown in brackets after item. **Tool development:** Based on AMSTAR, Cochrane Handbook for Systematic Reviews of Interventions and contributions from members of expert groups. Suggestions collated alongside AMSTAR items and list reviewed by all members in a series of iterations that led to the first draft of the tool. All proposed changes were discussed and agreed during	*(A) Data sources (Were the review methods adequate such that biases in location and assessment of studies were minimized or able to be identified?)* 1. Eligibility criteria were stated and suitably specific for (check all that apply)… (participants, intervention, comparator, outcomes, study designs)	*(A) Data sources (Were the review methods adequate such that biases in location and assessment of studies were minimized or able to be identified?)* 2. Were any further restrictions placed on eligibility of studies or reports? (Yes / No / Unclear) 3. Data for meta-analysis were sought from (check all that apply)… (published literature, online repositories, correspondence with trialists, in-house IPD, others' IPD) 4. Were data disclosed by industry sought specifically? (Yes / No / Unclear / Not relevant) 5.The search for trials included (check all that apply)… (bibliographic databases, grey literature, the web, in-house collections, reference lists, hand searching, correspondence with industry, other correspondence, other sources) 6. Which bibliographic databases are mentioned? (PubMed/MEDLINE, EMBASE,	*(A) Data sources (Were the review methods adequate such that biases in location and assessment of studies were minimized or able to be identified?)* 9. Study selection was done... (By one person / By one person, checked by another / By two or more people independently / Unstated or unclear / Not relevant (e.g. in-house data)) 10. Data extraction from published reports was done... (By one person / By one person, checked by another / By two or more people independently / Unstated or unclear / Not relevant (e.g. in-house data)) 11. Was risk of bias (or quality) assessed for each included study? (Yes / No / Unclear) 12. Risk of bias (or quality) was assessed using (check all that apply)… (scale, checklist, item-by-item assessment, only informally) 13. Risk of bias (quality assessment) or eligibility criteria included (check all	*(B) Analysis of individual studies by the meta-analyst (Were the individual studies analysed appropriately and without avoidable bias?)* Missing outcome data 15. Are adequate methods used to address missing outcome data? (Yes / No / Unclear / Not relevant) 16 . Cross-over trials were (Not found or not mentioned / Included appropriately / Included inappropriately / Explicitly excluded / Unclear) 17. Cluster-randomized trials were (Not found or not mentioned / Included appropriately / Included inappropriately / Explicitly excluded / Unclear) 18. Other study designs were (Not found or not mentioned / Included appropriately / Included inappropriately / Explicitly excluded / Unclear) *(C) General meta-analysis (Were the basic meta-analysis methods appropriate?)* 20. Were comparisons sensible within each meta-analysis? (Yes / No / Unclear)	*(D) Reporting and interpretation (Are the conclusions justified and the interpretation sound?)* 39. Were results appropriately interpreted in the light of risk of bias in included studies? (Yes / No / Unclear) 40. Were results appropriately interpreted in the light of risk of reporting bias? (Yes / No / Unclear) 41. Were results appropriately interpreted in the light of any multiplicity? (Yes / No / Unclear Comment) 43. Source of funding:

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
teleconferences of the full expert group. Resulting tool was piloted by pairs of assessors using 2 industry supported meta-analysis and 2 non-industry supported meta-analyses from 2005 or 2006. Further amendments were made in light of any difficulties encountered. Integral part of the tool was a guidance document. **IRR:** Unweighted raw agreement ranged from 35% (summary question C) to 42% (summary question A), and weighted raw agreement from 71% (summary question A) to 79% (summary question D). Weighted kappa measures ranged from 0.30 (summary question B) to 0.45 (summary question D). According to classifications of Landis and Koch (1977), these correspond to 'fair' or 'moderate' agreement.		CENTRAL/Cochrane Library, Science Citation Database/Web of Science, Others: how many) 7. The search strategy for bibliographic databases was (Not presented / Partially presented / Presented and comprehensive / Presented and not comprehensive) 8. Was the search for evidence reasonably comprehensive? Yes / No / Unclear	that apply)… (generation of allocation sequence, concealment of allocation sequence, blinding, attrition/drop-out/ITT, other) (D) Are the conclusions justified and the interpretation sound? 38. Were results of risk of bias (methodological quality) assessments reported? Yes in a table / Yes in the text / Unclear / No	21. Were outcomes sensible within each meta-analysis? (Yes / No / Unclear) 22. Do the authors avoid double-counting of individuals? Yes / No / Unclear 23. Presence of statistical heterogeneity was assessed by (check all that apply)… (visualisation, statistical test, I^2, other, not done) 24. The synthesis methods used in the paper included (check all that apply)… (pooling, fixed-effect meta-analysis, random effects meta-analysis, fixed effect meta-regression, random-effects meta-regression) 25. Synthesis methods were mainly (Classical - basic / Classical - advanced / Bayesian) 26. Was a sensible strategy used to address statistical heterogeneity in meta-analyses? (Yes / Unclear / No / No heterogeneity observed) 27. Were subgroups compared appropriately? (Yes / Unclear / No / Not applicable) 28. Were any subgroup	

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
				analyses apparently over-interpreted (e.g. because they were post hoc, or due to large number of subgroup analyses)? (Yes / Unclear / No / Not applicable) 29. Potential for reporting bias or small study effects was assessed using (check all that apply)… (funnel plots, Egger test, Begg-Mazumdar rank correlation test, other funnel plot asymmetry test, trim and fill, other) 30. Was the choice of effect size appropriate (e.g. MD vs SMD)? (Yes / Unclear / No / Not applicable) 31. Was skew of data a potential problem, not appropriately addressed? (Yes / Unclear / No / Not applicable) 32. Were methods appropriate to rare events/sparse data? (Yes / Unclear / No / Not applicable) 33. Were cut-points to dichotomize continuous/ordinal outcomes justified? (Yes / Unclear / No / Not applicable) 34. Were time-to-event data appropriately dealt with? (Yes	

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
				/ Unclear / No / Not applicable) 35. Were ordinal data appropriately dealt with? (Yes / Unclear / No / Not applicable) 36. Were indirect comparisons performed appropriately? (Yes / Unclear / No / Not applicable)	

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Ho(2010)[15] **Tool Name:** No name **Study designs targeted:** Not specified **Item rating:** General discussion rather than rating **Tool development:** Adapted from Oxman and Guyatt.[47] Authors opinions on how readers should interpret papers. Items have had very minor rewording and one item dropped. **IRR:** None	1. Did the review explicitly address a sensible and clearly focused clinical question? 2. Were the criteria used to select articles for inclusion appropriate?	3. Is it likely that relevant studies were missed?	4. Were the included studies evaluated for quality? 5. Is the method used to assess primary studies reproducible?	6. What are the overall results of the meta-analysis? 7. Were the results similar from study to study? 8. How precise were the results? 9. Were all clinically important outcomes considered?	

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Irwig(1994)[33] **Tool Name:** No Name **Study designs targeted:** DTA **Item rating:** Not specified **Tool development:** Not reported **IRR:** Not reported	*1. Determine the objective and scope of the meta-analysis* Is there a clear statement about: • The test of interest? • The disease of interest and the reference standard by which it is measured? • The clinical question and context? • Is the objective to evaluate a single test or to compare the accuracy of different tests? *2. Retrieve the relevant literature* • Are inclusion and exclusion criteria stated?	*2. Retrieve the relevant literature* • Is the literature retrieval procedure described with search and link terms given?	*3. Extract and display the data* • Are studies assessed by two or more readers? • Do the authors explain how disagreements between readers were resolved? • Is a full listing of diagnostic accuracy and study characteristics given for each primary study? *5. Assess the effect of variation in study validity on estimates of diagnostic accuracy*	*4. Estimate diagnostic accuracy* • Does the method of pooling sensitivity and specificity take account of their interdependence? • When multiple test categories are available, are they used in the summary? *5. Assess the effect of variation in study validity on estimates of diagnostic accuracy* Is the relation examined between estimates of diagnostic accuracy and study validity of the primary studies for each of the following design characteristics? • Appropriate reference standard • Independent assessment of the test or tests and reference standard • Avoidance of verification bias • In comparative studies, were either all of the tests of interest applied to each patient or were patients randomly allocated to the tests? • Are analytic methods used that estimated whether study design flaws affect diagnostic accuracy rather than just test threshold? *6. Assess the effect of variation in the characteristics of patients and test on estimates of diagnostic accuracy (generalizability*) • Is the relation examined between estimates of diagnostic accuracy and characteristics of the patients and test? • Are analytic methods used which differentiate whether characteristics affect diagnostic accuracy or test threshold?	

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Joanna Briggs(2006)[32] **Tool Name:** RAPiD tool (2006 version) **Study designs targeted:** Intervention studies **Item rating:** Yes, no or unclear **Tool development:** Not reported **IRR:** Not reported	*1. Well formated question* (Indicators: the question gives a clear understanding of what the review was trying to achieve). *2. Question clearly documented in the report* (Indicators: the question is clearly visible within the review report. This guides the review, and must be evident) *10. Describes who the target population were* (Indicators: this includes relevant demographic, disease/condition and intervention characteristics). *11. Describes the intervention/s in detail* (Indicators: it is clear what intervention/s were being reviewed, and that they are appropriate for the patient group). *12. Describes the outcomes in detail* (Indicators: all outcomes are stated and relevant to the interventions reviewed). *13. States the study design/s* (Indicators: it is clearly stated what study designs were to be included and what level of study was considered as	*3. Two phase search strategy is described* (Indicators:a two-phase search strategy is the minimum requirement, and includes an initial search to establish appropriate search terms and a second search of all relevant databases) *4. Phase 1 search terms appropriate* (Indicators: the search terms are specific to the topic in review.) *5. Phase 2 was relevant and exhaustive* (Indicators: the search strategy covers all aspects of the topic in review.) *6. Accessed a broad number of databases* (Indicators: the databases match the area of practice that is of concern, including all the major (broad) and minor (topic specific) databases) *7. References and bibliographies were searched* (Indicators: this is clearly documented). *8. Unpublished literature was sourced* (Indicators: the strategy for accessing unpublished literature is clearly defined).	*14. Critical Appraisal Method was appropriate* (Indicators: the criteria used are specific to the design of the included studies). *15. Critical Appraisal Method clearly reported* (Indicators: the methods used in the report are stated). *16. Any checklists or tools used are reported* (Indicators: all tools used are documented; this is often in the appendices). *17. Critical Appraisal determined by two independent reviewers* (Indicators: two independent reviewers have performed the critical appraisal process in an attempt to maintain consistency and eliminate bias). *18. Data Extraction methods were used to minimise errors* (Indicators: a recognised and agreed upon data extraction tool is used in an attempt to minimise errors, often in the appendices). *19. Data extraction methods were clearly reported* (Indicators: all tools used are documented; this is often in	*21. Results from individual studies are reported in a narrative, tabular or statistical summary* (Indicators: if studies could not be pooled statistically, the results are clearly reported). *22. The review question has been answered* (Indicators: the results of studies provide evidence that directly informs the review question). *23. If meta-analysis was undertaken, it was appropriate to combine studies* (Indicators: if studies were combined the study samples, interventions and outcomes are very similar (clinical homogeneity).) *24. The meta-analysis methods were appropriate* (Indicators: there are a variety of statistical methods available; their use is logical and appropriate). *25. The meta-analysis methods were reported* (Indicators: the above methods used in meta-analysis are reported). *26. Studies were tested for heterogeneity* (Indicators: this determines that the studies combined were sufficiently	*28. All issues of importance are addressed* (Indicators: the report includes all issues that are seen to be of importance). *29. Limitations are acknowledged* (Indicators: the review includes a list of what are seen to be the limitations of the study. This then acknowledges what is lacking in the study, and makes the reader aware of these areas.) *30. All issues arising from review addressed* (Indicators: everything that was discovered from the review is included. If there are particular issues, it could disclose the need for further work). *31. No major areas omitted* (Indicators: no area of the findings was left out of the recommendations for practice) *32. Recommendations for research are clear and unambiguous* (Indicators: any suggestion for further research as a result of the review is clearly and explicitly stated) *33. No major areas omitted* (Indicators: no area of the review was left out of the implications for further

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
	unacceptable). *14. Describes the exclusion in detail* (Indicators: the decisions about the inclusion criteria are justified in terms of the objectives of the review and any exclusions are also justified).	*9. There is reference to languages searched* (Indicators: if studies using languages other than the reviewer's first language are sourced, then they are documented. If not, then this is justified).	the appendices). *20. Double data entry by two independent reviewers* (Indicators: two independent reviewers have performed the data extraction process in an attempt to maintain consistency and eliminate bias).	statistically similar). *27. Summarises major findings of review* (Indicators: any major findings of the review are summarised and included in the report).	research)

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Knox (2009)[22] **Tool Name:** No name **Study designs targeted:** Generic **Item rating:** Yes, no, unclear, not reported **Tool development:** Based on QUORUM, Jadad et al, OQAQ and Users' guide to review articles. **IRR:** None	1. Question specified 2. Narrow focus of question 3. Explicit testable hypothesis	4. Search description 5. Use of multiple databases 6. Use of reference list 7. Search without language restriction	7. Study quality assessment 8. Tabulation of findings	9. Assessment of risk of missing studies 10. Assessment of risk heterogeneity 11. Meta-analysis	

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Li (2012)[19] **Tool Name:** No name **Study designs targeted:** RCT & NRS **Item rating:** Yes / No / Not reported / Not applicable Overall risk of bias rating for the review: Low / High / Unclear. Four key deficiencies from the 13 quality items were selected to classify the findings from a review as at low, high, or unclear risk of bias. Findings from a systematic review were classified as at high risk of bias if it contained a non-comprehensive literature search, did not assess the methodological quality of included studies, used inappropriate statistical methods for meta-analysis, or presented conclusions inconsistent with the review findings. **Tool development:** Adapted items from 3 instruments: AMSTAR, PRISMA, GRADE. No other details on development. **IRR:** Not reported.	1. Asked a focused question 2. Had pre-specified eligibility criteria	4. Performed comprehensive literature search	3. Assessed eligibility independently 5. Assessed methodological quality of included trials 6. Assessed methodological quality independently 7. Reported characteristics of included studies 8. Abstracted data independently	9. Synthesized evidence qualitatively 10. Used appropriate methods for meta-analysis	11. Discussed limitations at study level 12. Discussed limitations at review level 13. Conclusions consistent with review findings

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Light(1984)[36] **Tool Name:** No name **Study designs targeted:** "scientific and policy research" **Item rating:** **Tool development:** Authors' own. **IRR:** Not reported **Comments:** Wording slightly different in text; have extracted wording of items as specified as the tool.	1. What is the precise purpose of the review? 2. How were studies selected?			3. Is there publication bias? 4. Are treatments similar enough to combine? 5. Are control groups similar enough to combine? 6. What is the distribution of study outcomes? 7. Are outcomes related to research design? 8. Are outcomes related to characteristics of programs, participants, and settings? 9. Is the unit of analysis similar across studies?	10. What are guidelines for future research?

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Lundh(2012)[44] **Tool Name:** No name **Study designs targeted:** Generic **Item rating:** High, low, unclear **Tool development:** Not reported **IRR:** Not reported	1. Whether explicit and well defined criteria that could be replicated by others were used to select studies for inclusion/exclusion	2. Whether the search for studies was comprehensive	3. Whether there was an adequate study inclusion method, with two or more assessors selecting studies 4. Whether methodological differences and other characteristics that could introduce bias were controlled for or explored	4. Whether methodological differences and other characteristics that could introduce bias were controlled for or explored	

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Mailis(2012)[20] **Tool Name:** No Name **Study designs targeted:** Not specified **Item rating:** Reported, partially reported, not reported; QS items as yes, unclear, no. Guidance for reported or yes in brackets. Quality was rated numerically with respect to six quality subsections as follows: Good – six criteria met, or five criteria met and one criterion 'unclear'. Average – one criterion not met, or one criterion not met and one criterion 'unclear', or two criteria 'unclear'. Poor – at least two criteria not met. **Tool development:** In-house checklist adapted from a number of sources (1) Aggressive Research Intelligence Facility (ARIF) Critical Appraisal Checklist. University of Birmingham 2008. Available from: http://www.arif.bham.ac.uk/critical-appraisalchecklist. shtml (accessed September 10, 2008). 2) Fishbain D, Cutler RB, Rosomoff HL, Rosomoff RS. What is the quality of the implemented meta-analytic procedures in chronic pain treatment meta-analyses? Clinical Journal of Pain 2000;16(1):73-85. 3) Greenhalgh T. How to read a paper: Papers that summarise other papers (systematic	1. Study Question (The objectives of the review are clearly stated in the abstract, introduction, or methods). 2. Inclusion/Exclusion Criteria (All four elements (participants, interventions, outcome measures, types of studies) are reported in the abstract, introduction, or methods section of the review).	3. Search Strategy (At least one electronic database was searched and the names of the databases are provided). QS1: At least MEDLINE and EMBASE 4. Other sources (At least one additional resource or method, other than searching electronic databases, was used to identify relevant literature (e.g. pearling or review of reference lists in retrieved articles, hand searching of journals).	5. Data extraction method (The data extraction process is described.) QS 2: Standardized method (The data categories extracted are listed or the use of a standardized data extraction form is mentioned). QS 3: Independent data extraction by at least two reviewers 6. Criteria used to assess the validity of included studies (A quality assessment tool or checklist was used and details are provided (e.g. name or source) QS 4: Independent quality assessment by at least two reviewers (The quality of the included studies was assessed independently by at least two reviewers) 7. Inter-rater agreement (The review provides a statement of the degree of difference/equivalence between the reviewers or a statistical measure of inter-rater agreement)	Checklist divides synthesis into three different types - qualitative, semi-quantitative (statistical analysis of individual study without pooling results), Meta-analysis - with separate checklist criteria for each. *Qualitative review:* QS 5a: Study quality used in analysis or discussion of study results (Results of the included studies are discussed or analyzed in terms of their quality) *Semi-quantitative review*: QS 5b: Confidence interval/measures of dispersion reported (Confidence intervals or measures of dispersion (range, standard deviation, standard error of the mean) are reported for all relevant analyses) *Meta-analysis:* QS 5c: Precision of results reported (Confidence intervals are reported for all pooled effect estimates) QS 5d: Test of study heterogeneity conducted (A statistical analysis of study heterogeneity is reported for all pooled studies) 8. Test for publication bias (Publication bias was analysed or a reason provided for why it was	9. Potential methodological limitations (methodological limitations or advantages are described in a separate section or paragraph) 10. Incorporation of methodological quality (The methodological quality of the included studies is mentioned in the concluding section or discussion or statement of the review) QS: Conclusions supported by results (The conclusions drawn by the authors of the review are supported by the evidence presented in the results section) 11. Conflict of interest (A statement of conflict of interest (if any) is provided) 12. Sources of funding (Funding sources are mentioned; or the review was developed without external funding (e.g. authors employed by a university or volunteered time to produce a Cochrane Review).

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
reviews and meta-analyses). British Medical Journal 1997;315(7109):672-5). **IRR:** Not reported				not.)	

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Minelli(2009)[34] **Tool Name:** None **Study designs targeted:** Genetic association studies **Item rating:** Not clear; summary score based on general quality indicators calculated. **Tool development:** Not reported **IRR:** No information **Comments:** Also contained items on genetic quality indicators.	1. Inclusion/exclusion criteria not reported	2. Completely reproducible search strategy 3. Search methods not described	4. Duplicate eligibility checking and/or data extraction 5. Authors contacted for extra data 6. Quality assessment of individual studies 9. Designs of primary studies unclear 10. No details on study characteristics 11. No details on study-specific results	7. Statistical methods section in the paper 8. Forest plot of study-specific results 12. Formal tests for any interactions 13. Measure of size of heterogeneity (e.g., I 2) 14. No assessment of heterogeneity 15. P values without effect size estimate 16. Reason given for choice of fixed/random effects 17. Unclear whether fixed- or random-effects models were used 18. No assessment of publication bias 19. Study influence assessment	

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Mokkink(2009)[35] **Tool Name:** No Name **Study designs targeted:** Health status measurement instruments **Item rating:** Varied - yes/no, yes/no/unclear or specific answers **Tool development:** Not reported **IRR:** Not reported	1. Are the in- and exclusion criteria for articles described? (yes/no)	2. Is the search strategy used and described? (yes/no) 3. Number of databases searched (1, 2, 3, 4, >4) 4. Which databases are searched? (Pubmed, PsycINFO, CINAHL, EMBASE, Cochrane)	5. Is the selection of articles performed by at least two reviewers? (yes/no/unclear) 6. Is the data extraction performed by at least two reviewers? (yes/no/unclear)		

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Mulrow(1987)[27] **Tool Name:** No Name **Study designs targeted:** Not specified **Item rating:** Rated as 'specified', 'unclear' or 'not specified' **Tool development:** adapted from published guidelines for information synthesis (not clear which ones). **IRR:** Not reported	1. Specified purpose 3. Data selection	2. Data identification	4. Validity assessment	5. Qualitative synthesis 6. Quantitative synthesis	7. Summary 8. Future directives

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
NMHRC(2000)[28] **Tool Name:** No name **Study designs targeted:** Not specified **Item rating:** Not specified **Tool development:** Based on articles by Greenhalgh (1997) and Hunt and McKibbon (1997) **IRR:** Not reported	2. Were the inclusion criteria appropriate and applied in an unbiased way?	1. Was an adequate search strategy used?	2. Were the inclusion criteria appropriate and applied in an unbiased way? 3. Was a quality assessment of included studies undertaken? 4. Were the characteristics and results of the individual studies appropriately summarised?'	5. Were the methods for pooling the data appropriate? 6. Were sources of heterogeneity explored?	

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Nony(1995)[49] **Tool Name:** No name **Study designs targeted:** Not specified **Item rating:** Not stated **Tool development:** No information about tool development **IRR:** Not reported	1. Is the objective clearly stated? 2. Are the sources and rationale for the hypothesis tested indicated? 3. Are the proposed endpoints suitable? 4. Are the proposed end points reliable? 10. Are the inclusion and exclusion criteria clearly stated?	5. Are the sources exhaustive (computerized bibliographic databases and others)? 6. Is the search strategy fully described (computerised and others)? 7. Have unpublished trials been searched for (contact with investigators and for pharmaceutical companies)? 8. Has a search for multiple publications of the same trial or patient data been undertaken?	9. Is the selection of trials objective and independent of the results (ideally blinded selection)? 11. Is the quality assessment of the trial methods described? 12. Are excluded trials described (with reasons for exclusion)? 13. Full details of treatment studies: Are these coherent with the objectives of the meta-analysis? Are these homogenous in terms of the coherence? Are these compatible with the hypothesis and/or current medical practice? 14. Description of the patients included? Are these compatible with the objectives of the meta-analysis? Can the target population be described from the details given? 15. Are the extracted data summarised in a table? Can the calculation be checked and redone?	16. Description of statistical methods: rationale, software or methods used - Is the effect model suitable a priori, have several methods been used or are the reasons given for the choice of one method? 17. Are the estimation of the treatment effect and its CI, and the results of the association and homogeneity tests given? 18. Has the heterogeneity been analysed and if this was not possible has a practical interpretation been given and the sources of the heterogeneity been identified? 19. Were the subgroups defined a priori? 20. Is the rationale for the choice of subgroups given? 21. Is the robustness of the results discussed?	22. Are the conclusions consistent with the original goals and objectives of the meta-analysis? 23. Have the internal and external coherence been analysed and the implications of the results discussed?

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Oxman(1994)[47] **Tool Name:** Users' Guide **Study designs targeted:** Not specified **Item rating:** **Tool development:** No information about tool development. **IRR:** Not reported **Comments:** Tool structured in the three sections, "are the results of the study valid?", "what are the results?" and "will the results help me in caring for my patients?". We only extracted data for the first of these sections.	1. Did the overview address a focused clinical question? 2. Were the criteria used to select articles for inclusion appropriate?	3. Is it likely that important, relevant studies were missed?	4. Was the validity of included studies appraised? 5. Were assessment of studies reproducible?	6. Were the results similar from study to study?	

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Oxman(1994)[50] **Tool Name:** Not stated **Study designs targeted:** Not specified **Item rating:** Not stated **Tool development:** Not reported/unclear **IRR:** None	1. Problem formulation: Is the question clearly focused? 3. Study selection: Are the inclusion criteria appropriate?	2. Study identification: Is the search for relevant studies thorough?	4. Appraisal of studies: Is the validity of included studies adequately assessed? 5. Data collection: Is missing information obtained from investigators?	6. Data synthesis: How sensitive are the results to changes in the way the review is done?	Interpretation of results 7. Do the conclusions flow from the evidence that is reviewed? 8. Are recommendations linked to the strength of the evidence? 9. Are judgments about preferences (values) explicit? 10. If there is "no evidence of effect" is caution taken not to interpret this as "evidence of no effect"? 11. Are subgroup analyses interpreted cautiously?

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Oxman(1988)[51] **Tool Name:** No Name **Study designs targeted:** Generic **Item rating:** Not stated **Tool development:** Not reported **IRR:** Not reported	1. Were the questions and methods clearly stated? 3. Were explicit methods used to determine which articles to include in the review?	2. Were comprehensive search methods used to locate relevant studies?	4. Was the validity of the primary studies assessed? 5. Was the assessment of the primary studies reproducible and free from bias?	6. Was variation in the findings of the relevant studies analysed? 7. Were the findings of the primary studies combined appropriately?	8. Were the reviewers' conclusions supported by the data cited?

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Oxman(1991)[13, 54] **Tool Name:** OQAC (Overview Quality Assessment Checklist) **Study designs targeted:** Generic **Item rating:** 7-point scale response - 7 highest quality, 1 lowest quality **Tool development:** Preliminary set of criteria based on a review of the literature. The inclusion criteria used to select items were that they should measure "scientific quality" and should be applicable to overviews of practical questions in the health sciences i.e. questions regarding causation, prognosis, diagnosis, therapy, prevention or policy. Items were excluded if they were redundant, irrelevant to scientific quality or were not generalizable to both quantitative and qualitative overviews (meta-analyses and traditional narrative overviews) of clinically relevant topics. Items were initially selected based on the subjective assessment of one of the authors and were subsequently refined through an iterative process of discussions, pretesting and revision. In addition, much helpful advice was received from numerous investigators who had published relevant material. A mailed survey of editors and additional methodological expert known to be engaged in meta-analytic research did not generate any additional items or general concepts. In a pilot study nine overviews were each evaluated by nine judges. In addition to identifying any remaining ambiguities in the evaluation instrument and providing a basis for further revisions of the form, the pilot test was an important component of the training that the judges received. Twenty-five items were included in the instrument that was used in the consistency study. They were subsequently reduced by eliminating items that did not discriminate between overviews of high and low scientific quality. **IRR:** ICCs (and 95% CIs). Experts in research methodology: 0.77 (0.65 - 0.97). MDs with research training 0.74 (0.51 - 0.79). Research assistants 0.62 (0.38 - 0.78). All judges: 0.71 (0.59 - 0.81)	3. Were the inclusion criteria reported? 4. Was selection bias avoided?	1. Were the search methods reported? 2. Was the search comprehensive ?	5. Were the validity criteria reported? 6. Was validity assessed appropriately?	7. Were the methods used to combine studies reported? 8. Were the findings combined appropriately?	9. Were the conclusions supported by the reported data? 10. What was the overall scientific quality of the overview?

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Philibert (2012)[37] **Tool Name:** No Name **Study designs targeted:** Agronomy **Item rating:** NR **Tool development:** Based on the findings of previous studies (Borenstein et al., 2009; Roberts et al., 2006; Gates, 2002) **IRR:** None		(1) Correct description of the bibliographic search procedures used by the authors to select the individual studies (i.e. papers) and the repeatability of these procedures.	(2) Listing of the references of the selected individual studies used in the meta-analysis.	(3) Analysis of the variability of the results of individual studies, including checking to see whether the results vary between the selected individual studies and, when relevant, investigation of the sources of between-study variability (e.g. using random-effects model). Evaluation of the between-study variability of the response variable and of differences in the accuracy of individual estimates is an important step in a meta-analysis and several statistical methods have been proposed for the estimation of between- and within-study variances (Borenstein et al., 2009). (4) Analysis of the sensitivity of the conclusions to any change in the dataset and/or in the statistical method used to analyze the data. Sensitivity analyses should be carried out to identify influential data and to assess the robustness of the main conclusions of a meta-analysis to the assumptions made in the statistical analysis. (5) Assessment of the publication bias, which occurs when only studies with highly significant results are published. In this case, a meta-analysis can lead to a biased conclusion and an overestimation of the effect of a given factor. Publication bias is a predominant issue in meta-analysis and several methods such as funnel plots (e.g., Borenstein et al., 2009; Light and Pillemer, 1984) have been developed to detect the presence of such bias in datasets including published results. (6) Data weighting. When the results reported in the individual studies differ in their levels of accuracy, weighting of the data according to their levels of precision is recommended, based, for example, on the inverse of the variance of the measurements, as suggested by Hedges and Olkin (1985). (7) Availability of the dataset. (8) Availability of the program used for statistical analysis. These last two criteria are used to determine whether the meta-analysis could easily be re-run.	

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
PHRU(2006)[39] **Tool Name:** CASP **Study designs targeted:** Not specified **Item rating:** Yes, can't tell, no; 6 & 7 descriptive **Tool development:** No information about the development on the checklist (based on Oxman and Guyatt's Users' guide to the medical literature). **IRR:** Not reported	*1. Did the review ask a clearly-focused question?* (HINT: Consider if the question is 'focused' in terms of: the population studied, the intervention given or exposure, the outcomes considered) *2. Did the review include the right type of study?* (HINT: Consider if the included studies: address the review's question, have an appropriate study design)	*3. Did the reviewers try to identify all relevant studies?* (HINT: Consider: which bibliographic databases were used, if there was follow-up from reference lists, if there was personal contact with experts, if the reviewers searched for unpublished studies, if the reviewers searched for non-English-language studies)	*4. Did the reviewers assess the quality of the included studies?* (HINT: Consider: if a clear, pre-determined strategy was used to determine which studies were included. Look for: a scoring system, more than one assessor)	*5. If the results of the studies have been combined, was it reasonable to do so?* (HINT: Consider whether: the results of each study are clearly displayed, the results were similar from study to study (look for tests of heterogeneity) the reasons for any variations in results are discussed) *6. How are the results presented and what is the main result?* (HINT: Consider: how the results are expressed (e.g. odds ratio, relative risk, etc.), how large this size of result is and how meaningful it is, how you would sum up the bottom-line result of the review in one sentence) *7. How precise are these results?* (HINT: Consider: if a confidence interval were reported. Would your decision about whether or not to use this intervention be the same at the upper confidence limit as at the lower confidence limit? if a p-value is reported where confidence intervals are unavailable) *9. Were all important outcomes considered?* (HINT: Consider outcomes from the point of view of the: individual, policy makers and professionals, family/carers, wider community)	*10. Should policy or practice change as a result of the evidence contained in this review?* (HINT: Consider: whether any benefit reported outweighs any harm and/or cost. If this information is not reported can it be filled in from elsewhere?)

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Sacks(1987)[31, 53] **Tool Name:** No name **Study designs targeted:** Controlled clinical trials **Item rating:** Adequate, partial, none or unknown **Tool development:** Described as 'a scoring sheet listing what we considered to be the important elements of a meta-analysis'. No details on how items were selected. Tool contained 23 items, which were divided into six main areas: study design, combinability, control of bias, statistical analysis, sensitivity analysis, and problems of applicability. **IRR:** Not reported	1. Protocol	2. Literature search	3. List of trials analysed 4. Log of rejected trials 5. Treatment assignment 6. Ranges of patients 7. Ranges of treatment 8. Ranges of diagnoses 11. Selection bias 12. Data extraction bias 13. Inter-observer agreement 14. Sources of support (for primary studies)	9. Combinability criteria 10. Combinability measurement 11. Statistical methods (refers to acceptable methods of pooling studies) 12. Statistical errors 13.Confidence intervals 14. Subgroup analyses 15. Sensitivity analysis: quality assessment 16. Sensitivity analysis: varying methods 17. Sensitivity analysis: publication bias	

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Santaguida(2012)[16] **Tool Name:** No Name **Study designs targeted:** Not specified **Item rating:** Yes (2 points), partially (1 point), and no (0 points). Summed to create summary score. Scores greater than 14 were considered high overall quality, less than 13 to 11 as moderate overall quality, and less than or equal to 10 as low overall quality. **Tool development:** Used a previously modified tool - AHRQ modified OQAQ. URL: http://www.cms.gov/Medicare/Coverage/DeterminationProcess/downloads/id59TA.pdf **IRR:** None.	3. Were the criteria used for deciding which studies to include in the review reported?	1. Were the search methods used to find evidence (primary studies) on the primary question(s) stated? 2. Was the search for evidence reasonably comprehensive?	4. Was bias in the selection of articles avoided? 5. Were the criteria used for assessing the validity of the studies that were reviewed reported? 6. Was the validity for each study cited assessed using appropriate criteria (either in selecting studies for inclusion or in analyzing the studies that are cited)?	6. Was the validity for each study cited assessed using appropriate criteria (either in selecting studies for inclusion or in analyzing the studies that are cited)? 7. Were the methods used to combine the findings for the relevant studies (to reach a conclusion) reported? 8. Were findings of relevant studies combined appropriately relative to the primary question the review addresses?	9. Were the conclusions made by the author(s) supported by the data or analysis reported in the review?

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Shamliyan(2010)[21] **Tool Name:** No Name **Study designs targeted:** Observational **Item rating:** Varied according to item (see item details) **Tool development:** Based on guidelines for determining the reporting and methodological quality of systematic reviews: MOOSE, STROBE, Tooth et al. 2005, AMSTAR, AHRQ methods guide for comparative effectiveness reviews. **IRR:** Not reported		*1. Literature search*: no information, documented partially, complete documentation *2. Articles published in language other than English*: not addressed, included or justified exclusion of non-English publications *3. Grey literature*: not assessed, reported method of handling abstracts and unpublished studies	*4. Contact with authors of included studies*: no information, authors contacted *5. Formal internal quality evaluation of included study*: formal evaluation, some mention, no internal quality evaluation, reliability of internal quality evaluation reported, internal quality evaluation masked	*8. Pooled model obtained in the review*: pooling not obtained, fixed effect model obtained, random effects model obtained *9. Heterogeneity across included studies*: not reported, not significant, significant	*6. Conflict of interest from included studies*: not extracted *7. Sponsorship of included studies*: not analysed, analysed

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Shea (2009)[6, 52] **Tool Name:** AMSTAR **Study designs targeted:** RCT **Item rating:** Yes, no, can't answer, not applicable **Tool development:** A 37-item assessment tool was formed by combining the enhanced Overview Quality Assessment Questionnaire (OQAQ), a checklist created by Sacks, and three additional items recently judged to be of methodological importance. This tool was applied to 151 systematic reviews. Exploratory factor analysis was used to identify underlying components. The results were considered by methodological experts using a nominal group technique aimed at item reduction and design of an assessment tool with face and content validity **IRR:** The inter-rater agreement of the individual items of AMSTAR had a mean kappa of 0.70 (95% confidence interval [CI]: 0.57, 0.83) (range: 0.38 - 1.0) .	1. Was an *a priori* design provided?	3. Was a comprehensive literature search performed? 4. Was the status of publication (i.e. grey literature) used as an inclusion criterion?	2. Was there duplicate study selection and data extraction? 5. Was a list of studies (included and excluded) provided? 6. Were the characteristics of the included studies provided? 7. Was the scientific quality of the included studies assessed and documented?	9. Were the methods used to combine the findings of studies appropriate? 10. Was the likelihood of publication bias assessed?	8. Was the scientific quality of the included studies used appropriately in formulating conclusions? 11. Were potential conflicts of interest included?

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Sheikh(2007)[42] **Tool Name:** No Name **Study designs targeted:** Therapeutic / prognostic / diagnostic **Item rating:** Yes/No **Tool development:** Unclear **IRR:** Not reported	1. Question specified 2. Question relevant 3. Narrow focus of question 4. Explicit testable hypothesis	5. Adequate search description (incl. names of databases and search terms) 6. Use of reference list 7. Search without language restriction 8. Inclusion of unpublished data	Quality assessment of included studies based on the following: 9. Potential sources of bias (i.e. Randomisation) 10. Data collection (prospective/retrospective) 11. Follow-up 12. Blinding of assessors* 13. Description of intervention* (*applicable only to interventional reviews)	14. Assessment for risk of missing studies	

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
SIGN(2009)[18] **Tool Name:** SIGN **Study designs targeted:** Not specified **Item rating:** Well covered, adequately addressed, poorly addressed, not addressed, not reported, not applicable **Tool development:** Authors searched for existing checklists and selected those of the New South Wales Department of Health as they had undergone rigorous development and validation procedures. The checklists were further evaluated and adapted by the group in order to meet SIGN's requirements for a balance between methodological rigour and practicality of use. **IRR:** Not reported	1. The study addresses an appropriate and clearly focused question.	3. The literature search is sufficiently rigorous to identify all studies.	2. A description of the methodology used is included. 4. Study quality is assessed and taken into account.	5. There are enough similarities between the studies selected to make combining them reasonable.	6. How well was the study done to minimise bias? Code ++, +, or -

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Smith(1989)[26] **Tool Name:** No Name **Study designs targeted:** Not specified **Item rating:** Not stated **Tool development:** Derived from checklists by Bullock and Svyantek (1985), Light and Pillemer (1984) and Smither (1988). **IRR:** Not reported	1. Are the purpose and problem questions specified?	3. Are descriptions provided to ensure representativeness of the sample?	2. Does a theoretical framework serve as the basis for coding, hypothesis testing and interpretation of results? 4. Are decision rules made explicit at each step of the process? 7. Are checks for reliability and bias described at each step of the process?	5. Is there sufficient similarity among constructs, treatments and control groups for study comparisons? 6. Is the unit of analysis consistent across studies? 8. Are outcomes related to study characteristics?	9. Are alternative explanations in the form of rival hypotheses provided? 11. Is the report presented in sufficient detail for replication? 12. Are recommendations for the future specified?

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Smith(1997)[23] **Tool Name:** None **Study designs targeted:** Not specified **Item rating:** Yes/No and descriptive **Tool development:** Based on 'published guidelines and previous work'. References are to Oxman 1994 (ID 4253) and Mulrow (ID 4264). **IRR:** Not reported	1. Was the purpose of the review specified? 3. Were explicit criteria used to decide which articles to include in the review?	2. Were the search methods used to locate relevant studies comprehensive?	4. Was the methodological quality of the primary studies assessed?	5. How were the results of the primary studies combined?	6. Were suggestions made for future research?

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Smith(2007)[48] **Tool Name**: No Name **Study designs targeted:** Not specified **Item rating:** Not specified **Tool development:** Unclear **IRR:** none	1. Description of study selection and inclusion criteria	2. The extent of searching undertaken	1. Description of study selection and inclusion criteria 3. Description of methods used to assess the quality of included studies	4. Comparability of included studies 5. Assessment of publication bias 6. Assessment of heterogeneity	

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Thacker(1996)[46] **Tool Name:** No Name **Study designs targeted:** Not specified **Item rating:** Descriptive **Tool development:** Unclear. 'We propose an approach using the following series of 15 questions to be used by the reader to evaluate a published meta-analysis'. **IRR:** Not reported	1. Is the purpose of the study (i.e., the hypothesis) clearly identified? 3. Were explicit inclusion and exclusion criteria used to specify studies eligible for the meta-analysis?	2. Was an active, comprehensive effort made to include all available studies in the analysis?	7. Were multiple raters used to assess coding? If so, were they blinded and were measures of inter-rater reliability provided? 8. Were the selection and coding of data based on sound clinical principles or convenience? 9. Was documentation provided that explained how the data were coded and analyzed?	4. Was there an assessment of publication bias (i.e., bias resulting from reporting only those results that are statistically significant, which tends to overestimate the effect under study)? 6. Were the pooled data appropriate for testing the hypothesis? 10. Was the comparability of the cases and controls assessed? 11. Was heterogeneity testing conducted and reported appropriately? 12. Were results reported in sufficient detail to enable replication of results by the reviewer?	13. Were alternative explanations for observed results considered in the discussion? 15. Were guidelines provided for future research?

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Wilson(1992)[25] **Tool Name:** No name **Study designs targeted:** Generic **Item rating:** Descriptive **Tool development:** 'In compiling … we have drawn on the work of Light and Pillemer, Sacks et al and Oxman and Guyatt'. **IRR:** No information	1. Did the authors work to a written protocol? 2. Have the authors defined the research questions clearly? 3. Have the authors described their search strategy and how studies were chosen for inclusion?	3. Have the authors described their search strategy and how studies were chose for inclusion?	4. How have the study authors assessed the quality of individual studies? 5. How have the authors abstracted the information from individual studies? 6. Have the authors provided adequate details of the subjects in the studies being analysed?	7. Have the authors plotted their results? 8. Have the authors inspected the data for heterogeneity of outcome? 9. How have the authors calculated a summary estimate of the effect of the intervention? 10. Have the authors inspected the data for evidence of publication bias?	

Tool Details	Domain 1: Research question and eligibility criteria	Domain 2: Searching for studies	Domain 3: Review Process	Domain 4: Synthesis	Domain 5: Summarising the findings and reaching conclusions
Zambon(2012)[29] **Tool Name:** No name **Study designs targeted:** RCT **Item rating:** Yes/no or yes/no/not reported Reviews meeting items 1, 3, 4, 5, & 8 were judged at low risk of bias, those meeting two to four of these criteria were judged at moderate risk of bias, and the others at high risk of bias or at unclear risk of bias, depending on thoroughness of reporting. **Tool development:** Not reported **IRR:** No information	1. Explicit methods for study selection, abstraction and pooling 2. Only RCT included 3. Only double-blind RCT included	4. Study search explicit and extensive	1. Explicit methods for study selection, abstraction and pooling	5. Statistical heterogeneity/inconsistency	6. Competing conflicts of interest 7. Funding for review 8. Discrepancy between quantitative results and authors' recommendations

Appendix 6: Data extraction tables Review 3

Review of overviews that have used AMSTAR

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Amato (2011)[117]	**Review topic** Efficacy and safety of pharmacological interventions for the treatment of the Alcohol Withdrawal Syndrome **Study designs included** RCTs **Number of reviews**:5 **Type of synthesis**: Narrative	No	'we identified and discussed differences in quality between reviews, and used the quality assessment to interpret the results.'	None
Andersen(2011)[95]	**Review topic** 1) the evidence for causal relationships between computer work and the occurrence of carpal tunnel syndrome (CTS) or upper extremity musculoskeletal disorders (UEMSDs), and 2) intervention studies among computer users/or office workers. **Study designs included** Not reported. Cochrane reviews likely to have been RCT-based. Cohort studies reported as included in several reviews. Cross-sectional and retrospective studies also referred to within the text. **Number of reviews**:17 **Type of synthesis**: Narrative	**Items removed:** Conflict of interest item removed	Description of quality within results and commentary in discussion	Summed the number of items that were scored positively, maximum obtainable score of 9. 0-4 considered as low quality, and 5 or more as moderate to high quality

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Anttila(2012)[55]	**Review topic** Assistive technology interventions for people with disability **Study designs included** RCTs, CCT Observational studies (cross-sectional, 'post-test', 'pre/post' , Before-and-After, retrospective studies) Qualitative studies Case studies, 'ABA single subject', **Number of reviews**:44 **Type of synthesis**: Narrative	**Items added:** Two items relating to external validity added concerning reporting of participants' functional limitations and the study settings.	Described within results	"minor limitations" at least eight criteria met; "moderate" at least five; "major" fewer than five.
Aziz (2013)[86]	**Review topic** prosthodontics **Study designs included** RCTs Observational studies **Number of reviews**:106 **Type of synthesis**: Narrative summary of quality	**Scoring response modified:** Items were scored as Yes No Partially Cannot answer or N/A	Primary aim of paper was to assess quality of SRs in prosthodontics therefore the validity assessments represent the results of the overview.	Each component that was completely addressed in the SR was given a score of 2, partially addressed was given a score of 1, and not addressed or cannot answer was given a score of 0. Maximum achievable score of 22

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Berkhof(2011)[57]	**Review topic** Training strategies for teaching communication skills to physicians **Study designs included** RCTs Observational studies: CBA, ITS, observational study, pre-post-test; post-test only; case study; open effect study; descriptive studies **Number of reviews**:12 **Type of synthesis**: Narrative	**Items added:** Whether or not the outcome measures in the reviews were clearly described and integrated in the results	Description of quality within results and commentary in discussion	Summed the number of items that were scored positively. Classified the reviews as: high methodological quality (9-12 times a score of 'yes'), medium methodological quality (5-8 times a score of 'yes'), or low methodological quality (0-4 times a score of 'yes')
Bouchard(2011)[65]	**Review topic** Comparative critical appraisal of mixed methods reviews vs. quantitative reviews **Study designs included** Qualitative and unspecified quantitative studies **Number of reviews**:22 **Type of synthesis**: Narrative summary of quality	No	Examined suitability of AMSTAR for appraisal of mixed methods reviews	One point for each question answered "yes", 0 for any other answer. Scores summed.

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Braga(2011)[74]	**Review topic** Urology **Study designs included** RCTs Observational studies **Number of reviews**:57 **Type of synthesis**: Narrative summary of quality	No	To provide a general descriptive assessment of the methodological quality of the SRs	Assigned a score of 1 when a criterion was met, and 0 when not met. For each SR a summary AMSTAR score was calculated using a score of 0 to 11 with higher values reflecting better methodological quality. Mean AMSTAR score was 4.8 +/-2.0 (range 1 to 8). Mean kappa as a measure of interobserver agreement was 0.73 (range 0.44 to 0.93).
Brouwers (2011)[118]	**Review topic** Knowledge translation interventions in cancer control **Study designs included** Not specified (there were no restrictions to study types) **Number of reviews**:34 **Type of synthesis**: Narrative	No	Description of quality within results and commentary in discussion	"AMSTAR ratings can range between 1 and 11, with 11 denoting highest quality." and "The overall quality of the systematic reviews targeting consumer interventions was variable, ranging from poor to high. The average AMSTAR score was 7, with scores ranging from 3 to 10."
Burda(2011)[87]	**Review topic** Mammography screening in asymptomatic, average-risk women 40-49 years of age **Study designs included** Not reported **Number of reviews**:9 **Type of synthesis**: Narrative summary of quality	No	Used to comment on the quality of the evidence reviews contained within clinical practice guidelines.	Each item was given a score of 1 if the specific criterion was met or a score of 0 if the information was not reported, was unclear or the criterion was not applicable. For each review, scores for each item were averaged across the five assessors. The total AMSTAR score for each review was calculated by adding the average scores for all 11 items.

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Cates(2012)[66]	**Review topic** Safety of formoterol or salmeterol in children with asthma **Study designs included** RCTs **Number of reviews**:6 **Type of synthesis**: Network meta-analysis	No	Description of quality within results and commentary in discussion	"the AMSTAR ratings were high (all achieved a score of at least 9 out of a possible 11)"
Chafen(2010)[88]	**Review topic** Diagnosis and management of common food allergies **Study designs included** Not reported (prevalence studies therefore likely to be cross-sectional) **Number of reviews**:1 **Type of synthesis**: Narrative summary of quality	No	No results of AMSTAR assessment reported	None
Chan(2012)[119]	**Review topic** Prevention/management of radiation dermatitis. **Study designs included** RCTs Observational studies Qualitative studies **Number of reviews**:6 **Type of synthesis**: Narrative	No	Description of quality within results and commentary in discussion	Summed the number of items that were scored positively. Reviews achieving scores of 8-11 were deemed to have high methodological quality, those with scores of 4-7 had medium methodological quality, and those with scores 0-3 had low methodological quality

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Chipps(2012)[96]	**Review topic** Effectiveness and feasibility of videoconference-based telepsychiatry services for resource constrained environments **Study designs included** RCTs Observational studies **Number of reviews**:10 **Type of synthesis**: Narrative	'Revised assessment of multiple reviews' (R-AMSTAR) was used	**Criterion for inclusion in review:** Reviews with a QS of >=22 were classified as eligible for full review and assessment of quality	'revised assessment of multiple reviews' (R-AMSTAR) was used to assess systematic reviews. A total quality score (QS) out of 44 was computed by counting ratings per item.
de Bot(2011)[89]	**Review topic** Sublingual immunotherapy for allergic rhinitis in children **Study designs included** RCTs **Number of reviews**:10 **Type of synthesis**: Narrative summary of quality	No	Main aim of the overview was summarise quality therefore validity assessment formed the results of the overview.	Summed the number of items that were scored positively. Scores of 0-4 indicate that the review is of low quality, 5-8 of moderate quality, and 9-11 of high quality
Dent(2012)[107]	**Review topic** Changes in Body Weight and Psychotropic Drugs **Study designs included** Table 1 implies RCTs only but not entirely clear **Number of reviews**:20 **Type of synthesis**: Narrative	No	Only reported quality gradings in table, no further details of AMSTAR assessment	Reviews were graded as good (A), fair (B), and poor (C). Not clear how this was done.

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Elangovan(2013)[90]	**Review topic** Periodontal regeneration in humans **Study designs included** Unclear/Not reported **Number of reviews**:14 **Type of synthesis**: Narrative summary of quality	No	Constituted the results of the study	All of the yes scores (which were given the value of 1) summed to give an overall score (min total score 0, max 11). Score of <3 has been suggested to be of poor quality.
Faggion(2012)[91]	**Review topic** Animal studies in dentistry **Study designs included** Not described: "Animal research in dentistry" **Number of reviews**:54 **Type of synthesis**: Narrative summary of quality	**Scoring response modified:** 'No' and 'can't answer' collapsed into one category	Main purpose of the overview was to summarise quality of reviews therefore AMSTAR assessment constitutes the results of the paper.	The overall score was categorised into three levels: 8-11 = high quality; 4-7 = medium quality, and 0-3 = low quality
Faggion(2010)[114]	**Review topic** Implant dentistry **Study designs included** RCTs **Number of reviews**:2 **Type of synthesis**: Narrative	No	Description of quality within results and commentary in discussion	"The methodologic quality was determined from the percentage of Yes scores for each study."

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Fleming(2013)[92]	**Review topic** Orthodontics **Study designs included** Not stated **Number of reviews**:109 **Type of synthesis**: Narrative summary of quality	No	Association between PRISMA and AMSTAR scores assessed using linear regression. AMSTAR score used as the dependent variable in a regression analysis to identify characteristics of reviews that are associated with quality (reported in separate publication).	An overall percentage score was assigned to each review based on the sum of the applicable items
Flodgren(2011)[120]	**Review topic** Financial incentives for changing healthcare provider behaviour and health outcomes **Study designs included** RCTs, CCTs Observational studies (CBA, ITS) **Number of reviews**:4 **Type of synthesis**: Narrative	No	Description in results and comment in discussion. 'We had hoped to examine variation in review quality to see if it explained variations in the results of the reviews. However, because we had to use vote counting, this was not possible.'	Included reviews were categorised into bottom (score 0 to 3), middle (score 4 to 7), and upper (score 8 to 11) tertiles.
Friedman(2011)[121]	**Review topic** Teaching strategies and methods of delivery for patient education (PE) **Study designs included** Not reported **Number of reviews**:23 **Type of synthesis**: Narrative	No	Reported in the results	Scored 1 for each Yes and summed to give a total score.

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Hagen(2012)[97]	**Review topic** Exercise therapy for bone and muscle health **Study designs included** RCTs **Number of reviews**:9 **Type of synthesis**: Narrative	**Scoring response modified:** "The 11 [AMSTAR] criteria were rated as 'met,' 'unclear/partly met,' or 'not met'."	Described within the results	"The 11 [AMSTAR] criteria were rated as 'met,' 'unclear/partly met,' or 'not met'. A second reviewer independently verified the accuracy of the numeric results." ... "Four reviews were assessed to be of high methodological quality (all 11 criteria met), whereas in three reviews eight to nine criteria were met. Finally, in two reviews only three to four criteria were met."
Hopton(2010)[122]	**Review topic** Acupuncture for chronic pain **Study designs included** RCTs **Number of reviews**:8 **Type of synthesis**: Narrative	No	Described in results and commentary in discussion	None

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Jacobs(2012)[68]	**Review topic** Spinal surgery **Study designs included** RCTs Observational studies **Number of reviews**:7 **Type of synthesis**: Narrative summary of quality; meta-analysis of included reviews	**Items added:** Was the effect of methodological bias analyzed? **Scoring modified:** Expanded to give strict instructions for when to give a yes or a no response	As a variable in meta-regression: Unable to use in meta-regression due to small number of reviews matching inclusion criteria. AMSTAR assessments reported within results.	One point for each criterion met.
Jagannath(2011)[84]	**Review topic** Quality of systematic reviews published in five leading Indian medical journals **Study designs included** Not reported **Number of reviews**:22 **Type of synthesis**: Narrative summary of quality	No	Quality ratings constituted the results of the paper	A score of >4 out of 11 was deemed to be acceptable quality
Jaspers(2011)[103]	**Review topic** Effects of clinical decision-support systems on practitioner performance and patient outcomes **Study designs included** RCTs Observational studies **Number of reviews**:17 **Type of synthesis**: Narrative	No	**Criterion for inclusion in review:** Reviews with a score of 9 or greater were included in the review	Scores of 0-4 indicate that the review is of low quality; 5-8 that the review is of moderate quality; and 9-11 that the review is of high quality.

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Jin(2012)[85]	**Review topic** Reporting and methodological quality of systematic reviews or meta-analyses in the nursing field in China **Study designs included** RCTs, non-randomised controlled trials Observational studies **Number of reviews**:63 **Type of synthesis**: Narrative summary of quality	**Scoring response modified:** 'Can't answer' and not applicable' collapsed into one category within the results. Not clear if this was done prior to reviewing SRs or after.	Formed the results of the overview	A score of 0-4, 5-8, and 9-11 indicates a poor, moderate, and high quality for the review, respectively
Johnson(2012)[123]	**Review topic** Green tea and green tea catechin extracts **Study designs included** RCTs Observational studies **Number of reviews**:8 **Type of synthesis**: Narrative	No	Described within the results	None
Jones(2012)[124]	**Review topic** Pain management for women in labour **Study designs included** RCTs **Number of reviews**:18 **Type of synthesis**: Narrative	No	Described within the results and commentary within discussion	Each Yes is 1 point, summed up to give a score out of 11.

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Kamioka(2010)[125]	**Review topic** Aquatic exercise and balneotherapy **Study designs included** RCTs **Number of reviews**:7 **Type of synthesis**: Narrative	No	Described within results	No summary scores
Kang(2012)[71]	**Review topic** Chinese medicine - Reliability and validity of AMSTAR in Chinese studies **Study designs included** Not reported **Number of reviews**:41 **Type of synthesis**: Narrative summary of quality	Tool translated into Chinese language	Used to examine the reliability and external validity of AMSTAR in reviews of Chinese medicine.	"If an item was scored yes, it would be given one point, otherwise, 0 point. We added up these to calculate a total score."
Kim(2012)[126]	**Review topic** Dietary supplements for benign prostatic hyperplasia with lower urinary tract symptoms **Study designs included** RCTs **Number of reviews**:6 **Type of synthesis**: Narrative	No	Described in results and comment in discussion	The highest quality (if all 11 criteria were met), high quality (if 8-11 were met), medium quality (if 4-7 were met) or low quality (if 0-3 were met)

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Kumar(2011)[127]	**Review topic** Treatments for multiple myeloma **Study designs included** RCTs **Number of reviews**:11 **Type of synthesis**: Narrative	No	Description in results and comment in discussion	None
Leucht(2012)[128]	**Review topic** Pharmacological treatments for common medical and psychiatric disorders **Study designs included** RCTs **Number of reviews**:127 **Type of synthesis**: Narrative	No	Described within the results	"The quality of the included systematic reviews was evaluated with the AMSTAR score (range of possible values 0-11)"
Li(2012)[98]	**Review topic** Gastric cancer risk and protective factors **Study designs included** RCTs Observational studies **Number of reviews**:59 **Type of synthesis**: Narrative summary of quality	Used R-AMSTAR	Reported within results and commentary in discussion	R-AMSTAR – score out of 44

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Li(2011)[129]	**Review topic** Treatment of trapeziometacarpal osteoarthritis **Study designs included** Unclear. **Number of reviews**:2 **Type of synthesis**: Narrative	No	Described within results and comment in discussion	None
List(2010)[58]	**Review topic** Management of temporomandibular disorders **Study designs included** RCTs Observational studies Case series. 'uncontrolled studies'. **Number of reviews**:30 **Type of synthesis**: Narrative	No	Description in results and comment in discussion	One point for each of the 11 criteria met, total score between 0 and 11.
Littell(2011)[93]	**Review topic** Long term psychodynamic psychotherapy **Study designs included** RCTs Observational studies **Number of reviews**:1 **Type of synthesis**: Narrative summary of quality	No	Used to critique a published meta-analysis.	None

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Lougheed(2012)[106]	**Review topic** Diagnosis and management of asthma in pre-schoolers, children and adults **Study designs included** RCTs **Number of reviews**:16 **Type of synthesis**: Narrative	No	Results of AMSTAR assessment not reported	None
McGee(2013)[72]	**Review topic** Surgical procedures in children **Study designs included** RCTs **Number of reviews**:15 **Type of synthesis**: Narrative summary of quality	**Scoring response modified:** The quality in each AMSTAR domain was graded as high, low or unclear for each review according to the criteria in the AMSTAR tool	Validity assessment formed part of the results of the overview	None
Ma(2011)[73]	**Review topic** Traditional Chinese medicine interventions **Study designs included** Not reported **Number of reviews**:369 **Type of synthesis**: Narrative summary of quality	No	Reported as part of the results of the overview.	None

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Ma(2012)[75]	**Review topic** Acupuncture reviews in Chinese journals **Study designs included** Not stated **Number of reviews**:88 **Type of synthesis**: Narrative summary of quality	No	Formed part of the results of the overview	None
MacDonald(2010)[76]	**Review topic** Urology **Study designs included** RCTs Observational studies **Number of reviews**:57 **Type of synthesis**: Narrative summary of quality	No	Validity assessments constitute the results of the paper	Summary score was calculated. Details not given but score out of 11 so assume each item assigned score of 1 if fulfilled.
Mahtani(2013)[130]	**Review topic** Participation in physical or sporting activities after olympic or paralympic games **Study designs included** Not reported **Number of reviews**:2 **Type of synthesis**: Narrative	No	Reported within the results	Used numeric summary score out of 11.

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Matheson(2011)[108]	**Review topic** Non-genetic risk factors and putative antecedents of schizophrenia **Study designs included** Observational studies (population-level ecological design; mixed design studies) **Number of reviews**:24 **Type of synthesis**: Narrative	No	Summary quality rating reported in tables	AMSTAR ratings below 27% were considered of low quality; 27-73% of moderate quality; and 73-100% of high quality. The cut-offs were calculated using equal thirds of each rating scale.
Matjasko(2012)[110]	**Review topic** Youth violence prevention programs **Study designs included** RCTs **Number of reviews**:52 **Type of synthesis**: Narrative	No	Both the AMSTAR scale and a categorical variable based on the AMSTAR rating were used to determine the relationship between program effects and study quality using ANOVA.	An AMSTAR score was calculated by adding all of the 'yes' responses for each article reviewed; the maximum score is 11. Scores of 0 to 4 indicate that the review is of low quality; 5 to 8 indicate moderate quality; and 9 to 11 indicate high quality.

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Melchiors(2012)[59]	**Review topic** Pharmacist-delivered health interventions **Study designs included** RCTs, CCTs, quasi-experimental studies Observational studies (before and after studies, observational studies) Case studies **Number of reviews**:31 **Type of synthesis**: Narrative summary of quality	**Scoring modified:** Item 5 is scored as 'yes' when systematic reviews provide a list of studies that were included and excluded according to the instructions of the authors. However, reviews were also scored as 'yes' in item 5 if the revisions had a flowchart of excluded and included studies with their reasons for exclusion.	Formed the results of the overview	The total score using AMSTAR was obtained by summing one point for each 'yes' and no points for any other score ('no', 'can't answer' and 'not applicable'), ranging from 0 to 11. A score of 0-4 indicates a poor quality review, 5-8 indicate moderate quality and 9-11 indicate high quality.
Michiels(2011)[131]	**Review topic** Efficacy, effectiveness and risks of the use of inactivated influenza vaccines in children, healthy adults, elderly individuals and individuals with comorbidities **Study designs included** RCTs Observational studies **Number of reviews**:12 **Type of synthesis**: Narrative	No	Brief comment on quality of included reviews within the results section.	None

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Mickenautsch(2011)[1][32]	**Review topic** Minimum intervention in dentistry: powered toothbrushes, triclosan toothpaste, essential oil mouthwashes, xylitol chewing gum **Study designs included** RCTs and non-randomised studies **Number of reviews**:5 **Type of synthesis**: Narrative	No	Reported within the results and comment within discussion	Total score out of 11 given by summing yes responses
Mikton (2009)[133]	**Review topic** Universal and selective and selective child maltreatment prevention interventions **Study designs included** RCTs Observational studies Non-randomised controlled studies; no control group; 'other' **Number of reviews**:26 **Type of synthesis**: Narrative	No	Constituted one of the results of the study	Score out of eleven based on summing items receiving a yes response. Also classified as low (i.e. AMSTAR scores between 0-4), moderate (5-8), and two, or high quality (9-11)

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Minozzi(2013)[60]	**Review topic** Incidence or prevalence of opioid dependence syndrome in adults (with and without previous history of substance abuse) following treatment with opioid analgesics for pain relief **Study designs included** RCTs, case series, unclear and not reported. **Number of reviews**:3 **Type of synthesis**: Narrative	**Scoring response modified:** 'Not reported' and 'unclear' (as opposed to 'can't tell'?) appear to have been included as options (table 1).	Reported within the results	None
Moe(2009)[99]	**Review topic** Non-pharmacological and nonsurgical interventions for hand osteoarthritis **Study designs included** RCTs **Number of reviews**:4 **Type of synthesis**: Narrative	**Scoring response modified:** Criteria were rated as 'met', 'unclear/partly met', or 'not met'	**Criterion for inclusion in review:** One review was excluded as it met none of the AMSTAR criteria (reported in the results section not mentioned in the methods). AMSTAR scoring for all included reviews reported within the results section.	From methods: "with overall scores ranging from 0 to 10 (out of a maximum of 11 criteria)" However, the results of AMSTAR assessments were presented appropriately in Table 1, without the total scores.
Monasta(2010)[111]	**Review topic** Early-life determinants of overweight and obesity **Study designs included** Observational studies **Number of reviews**:22 **Type of synthesis**: Narrative	No	Reported within the results	"The maximum score is 11, scores of 0-4 indicating low quality, 5-8 moderate quality, and 9-11 high quality. If the total scores of the independent evaluation differed by one or two points, the average was calculated. If the differences were wider, a third author carried out an additional independent evaluation."

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Moore(2011)[67]	**Review topic** Single dose oral analgesics for acute postoperative pain in adults **Study designs included** RCTs **Number of reviews**:35 **Type of synthesis**: Network meta-analysis	No	Reported within the results and commentary within the discussion.	None
Oestergaard(2011)[63]	**Review topic** Non-pharmacological and pharmacological interventions versus pharmacological alone for depression **Study designs included** RCTs, non-random comparison of control and intervention groups; controlled clinical trials; Observational studies ('epidemiological descriptive studies', CBA; ITS) **Number of reviews**:19 **Type of synthesis**: Narrative	No	Reported in the results and comment in discussion	Sum of all items given a yes response.

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Panesar(2009)[77]	**Review topic** Cardiac surgery patients undergoing conventional coronary artery bypass (CCAB) vs. off-pump coronary artery bypass (OPCAB) surgery). **Study designs included** RCTs Observational studies **Number of reviews**:7 **Type of synthesis**: Narrative summary of quality	No	AMSTAR score was one item by which reviews were displayed on the veritas plot. Studies were ranked by AMSTAR score. The study with the best score received n points where n = the number of studies. The second best study received n -1 points, and so on. In the case of 2 studies performing equally well, the study with the next highest score would receive n -2 points.	Total score calculated based on number of yes responses and used to rank the reviews. "A yes gives a score of 1; any other response results in a score of 0. The overall score is out of 11."
Papageorgiou(2011)[78]	**Review topic** Orthodontics **Study designs included** 18 reviews limited included study designs to RCTs, but other designs included are not reported **Number of reviews**:110 **Type of synthesis**: Narrative summary of quality	**Scoring response modified:** Each item was assessed using a four point scale: Yes, Can't tell, No and Not applicable. A criterion was defined as can't tell if it was half met.	Study characteristics were used as predictors using the AMSTAR score as the dependent variable in linear regression. Variables found to be significant at the p <= 0.05 level were entered into multivariate linear regression models to assess for potential confounding factors. Risk ratios (RR) with 95% CI were used as summary statistics to compare quality and reporting between specific time points	Non-applicable items were excluded from the maximum scoring capability of each SR. Summary scores were extracted by giving one point for each Yes and half a point for each Can't tell in an attempt to maximize data output. Summary scores are reported as percentages

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Parker(2012)[134]	**Review topic** Prevention and treatment of maternal anaemia **Study designs included** RCTs **Number of reviews**:27 **Type of synthesis**: Narrative	No	Reported within the results and comment within discussion	Reviews were divided into the following categories - high quality: 9 or more positive answers; intermediate quality: 5-8 positive answers; low quality: 4 or less positive answers.
Payne(2012)[135]	**Review topic** Fatigue and weight loss in adults with advanced progressive illness **Study designs included** RCTs **Number of reviews**:27 **Type of synthesis**: Narrative	No	Reported within results and comment in discussion	'We deemed Cochrane Systematic Reviews achieving a score of 8 to 11 of high methodological quality, 4 to 7 of medium quality and 0 to 3 of low quality'
Popovich(2012)[79, 112]	**Review topic** Assisted reproduction for subfertility **Study designs included** RCTs **Number of reviews**:60 **Type of synthesis**: Narrative summary of quality	No	AMSTAR 'grades' were compared with those obtained using R-AMSTAR to compare the conclusions formed between the two assessment tools. Cochrane vs non-Cochrane reviews compared for individual AMSTAR items. From discussion: "The domains that need to be addressed by future CR authors are domains 2, 10 and 11."	Scores were converted to percentages, based on the maximum possible score (for the R-AMSTAR) and the number of domains with a 'yes' score (for the AMSTAR). Domains given a not-applicable ('NA') score were not used in the calculation. Based on the resulting percentage scores, grades were assigned to each review (A; 90%, B; 80%, C; 70%, D; 60%, E; 50%, F; ,50%.)

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Prior(2008)[136]	**Review topic** Clinical guideline implementation strategies **Study designs included** RCTs, CCT Observational studies, before-after, time series, cross-sectional **Number of reviews**:33 **Type of synthesis**: Narrative	No	Reported within the results.	Total count of Yes answers presented in Table 1 (in addition to individual answers to each item).
Rookmoneea(2010)[137]	**Review topic** Management of primary frozen shoulder **Study designs included** RCTs **Number of reviews**:11 **Type of synthesis**: Narrative	No	Described within the results and comment within the discussion	None

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Ryan(2011)[104]	**Review topic** Consumer-oriented interventions for evidence-based prescribing and medicines use **Study designs included** RCTs, quasi-randomised controlled trials (CCTs), Observational studies (controlled before-and-after studies (CBAs), interrupted time series (ITS) or before-and-after (BA) studies) **Number of reviews**:37 **Type of synthesis**: Narrative	No	**Criterion for inclusion in review:** Non-Cochrane reviews were excluded if rated as low quality or had serious methodological flaws according to the Centre for Reviews and Dissemination assessment of the review published as part of the DARE abstract; and as assessed by the reviewers using the AMSTAR assessment tool (rating of less than 4). AMSTAR also used to summarise quality of all included reviews within the results section and comment in the discussion.	Reviews classified as high (8-11), medium (4-7) or low (0-3) quality based on summed number of yes responses.
Sakzewski(2009)[109]	**Review topic** Therapeutic management of upper-limb dysfunction in children with congenital hemiplegia **Study designs included** RCTs, quasi-randomised controlled trials, 'non-RCTs', observational studies (pre-post studies, time series). **Number of reviews**:7 **Type of synthesis**: Narrative	No	Results of AMSTAR assessment only reported in table	A score of 1 was recorded for each criterion present, with a total possible score of 11.

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Saokaew.(2012)[56]	**Review topic** Opioid conversion **Study designs included** RCTs Observational studies: Crossover, Retrospective, Prospective (not randomized), Cross-sectional study Case series, n-of-1 crossover. **Number of reviews**:5 **Type of synthesis**: Narrative	No	Formed the results of the overview	Poor (0-4 of 'Yes'), Moderate (5-8 of 'Yes'), High (9-11 of 'Yes')
Savard, L.A.T., D. R. Clark, A. M.(2011)[138]	**Review topic** Heart failure disease management programs **Study designs included** RCTs **Number of reviews**:15 **Type of synthesis**: Narrative	No	Reported within the results	A score of 1 was recorded for each criterion present, with a total possible score of 11.

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Seo(2012)[80]	**Review topic** Quality of systematic reviews or meta-analyses for nursing interventions conducted by Korean researchers **Study designs included** RCTs True observational studies (studies with non-equivalent control group; quasi-experimental trials using a pre-test/post-test design) **Number of reviews**:22 **Type of synthesis**: Narrative summary of quality	No	Formed the results of the overview.	Total score calculated by summing one point for each yes and no point for others, including no, can't answer, and not applicable, resulting in summary scores from 0 to 11. Authors applied the following three categories: a score of 0-4 is classified as low quality, 5-8 indicates moderate quality, and 9-11 high quality
Sequeira-Byron(2011)[81]	**Review topic** Oral healthcare interventions published in the Journal of Applied Oral Science **Study designs included** Not reported **Number of reviews**:4 **Type of synthesis**: Narrative summary of quality	No	Constituted the results of the overview.	Summary scores calculated from total number of yes responses. Reviews graded as high, medium or low quality 'in concordance with the rating system used by the CADTH'.
Shepherd(2012)[113]	**Review topic** Structural alteration in both chronic and first-episode schizophrenia. **Study designs included** Not reported but most likely to be Case-control based on the review question **Number of reviews**:32 **Type of synthesis**: Narrative	No	Reported within the results and comment within the discussion	Ratings below 27% were considered low quality; 27-73% moderate quality; and 73-100% high quality. The overall quality rating of each review was a composite of both review methodological quality and the strength of the evidence

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Singh(2009)[69, 139]	**Review topic** Biologics for rheumatoid arthritis **Study designs included** RCTs **Number of reviews**:6 **Type of synthesis**: Network meta-analysis	No	Reported within the results	None
Spearing(2011)[140]	**Review topic** injury compensation and health outcomes **Study designs included** Observational studies **Number of reviews**:11 **Type of synthesis**: Narrative	No	Reported within the results and comment in the discussion	Score out of a total of 11.
Suebnukarn(2010)[94]	**Review topic** Endodontics **Study designs included** Observational studies **Number of reviews**:16 **Type of synthesis**: Narrative summary of quality	No	Constituted the results of the overview	Overall score categorized into three levels: 8 to 11 is high quality, 4 to 7 is medium quality, and 0 to 3 is low quality

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Torloni(2010)[82]	**Review topic** Design and level of evidence of articles published in 2007, in two recently indexed Brazilian journals (Clinics and Revista da Associação Médica Brasileira), and to evaluate the methodological quality of the SRs. **Study designs included** Not reported **Number of reviews**:4 **Type of synthesis**: Narrative summary of quality	No	Formed part of the results of the paper.	Summed items receiving a yes response and given a total score out of 11
van der Linde(2012)[141]	**Review topic** Behavioural and psychological symptoms in the older or demented population **Study designs included** Not reported **Number of reviews**:36 **Type of synthesis**: Narrative	No	Reported within the results and comment in discussion	Based on the results section, reviews appear to have been categorised as high, moderate (5-8 points) or low quality.

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Vidal(2011)[100]	**Review topic** Immunotherapy for follicular lymphoma **Study designs included** RCTs **Number of reviews**:11 (only 9 assessed with AMSTAR) **Type of synthesis**: Narrative	**Scoring response modified:** Little detail given in the methods, but in the results some of the questions have been answered as 'not reported' which isn't an option in the original AMSTAR tool.	Reported within the results In conclusions: "Criteria for assessment of systematic reviews, including the AMSTAR should be further explored and validated."	None
Weed(2011)[83]	**Review topic** Sugar-sweetened beverages and health outcomes **Study designs included** RCTs Observational studies **Number of reviews**:17 **Type of synthesis**: Narrative summary of quality	No	Constituted the results of the overview. Student's t test was used to compare mean AMSTAR scores for 1) reviews the investigators concluded as positively associated compared with all others and 2) reviews that were identified as systematic compared with all others	A score of one was assigned to a yes answer, and a score of zero was assigned to all other answers. Items summed to create an overall score out of 11.

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Wells(2013)[61]	**Review topic** Pilates exercise for chronic back pain **Study designs included** RCTs, pseudo-randomised controlled trial, case series. **Number of reviews**:5 **Type of synthesis**: Narrative summary of quality	Used R-AMSTAR criteria	Reported within the results and comment within the discussion.	R-AMSTAR summary score (out of 44)
Winters(2013)[70]	**Review topic** Hospital rapid response systems **Study designs included** Observational studies **Number of reviews**:43 **Type of synthesis**: Meta-analysis	No	**Criterion for inclusion in review:** Seven SRs were identified. The highest quality review (as determined by AMSTAR score) was selected and updated.	"The highest-quality systematic review (3) (assessment of multiple systematic reviews criteria score, 10 of 11)"
Wiysonge(2012)[105]	**Review topic** Interventions targeting barriers to effective immunisation programs **Study designs included** RCTs Observational studies (CBA, ITS) **Number of reviews**:10 **Type of synthesis**: Narrative	No	**Criterion for inclusion in review:** Authors only report data in the paper from reviews that they considered to be reliable (as determined by AMSTAR).	None

Review details	Details of included reviews	Modifications to AMSTAR	Incorporation of validity assessment	Summary quality rating
Zwicker(2010)[62]	**Review topic** Treadmill training in children with motor impairments **Study designs included** Unclear. Levels of Evidence from II to V included which implies small RCTs, ecological, cohort, case-control, case series, expert opinion, case series. **Number of reviews**:5 **Type of synthesis**: Narrative	No	Reported within the results and comment in the discussion	One point for each item scored as yes to give a total score out of 11.