Page, M. J., McKenzie, J. E., & Higgins, J. P. T. (2018). Tools for assessing risk of reporting biases in studies and syntheses of studies: A systematic review. *BMJ Open*, *8*(3), [e019703]. https://doi.org/10.1136/bmjopen-2017-019703

BMJ Open

Tools for assessing risk of reporting biases in studies and syntheses of studies: a systematic review

Matthew J Page,[1,2] Joanne E McKenzie,[1] Julian P T Higgins[2]

To cite: Page MJ, McKenzie JE, Higgins JPT. Tools for assessing risk of reporting biases in studies and syntheses of studies: a systematic review. *BMJ Open* 2018;**8**:e019703. doi:10.1136/bmjopen-2017-019703

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2017-019703).

Received 20 September 2017
Revised 10 January 2018
Accepted 22 January 2018

[1]School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia
[2]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

Correspondence to
Dr Matthew J Page;
matthew.page@monash.edu

ABSTRACT

Background Several scales, checklists and domain-based tools for assessing risk of reporting biases exist, but it is unclear how much they vary in content and guidance. We conducted a systematic review of the content and measurement properties of such tools.

Methods We searched for potentially relevant articles in Ovid MEDLINE, Ovid Embase, Ovid PsycINFO and Google Scholar from inception to February 2017. One author screened all titles, abstracts and full text articles, and collected data on tool characteristics.

Results We identified 18 tools that include an assessment of the risk of reporting bias. Tools varied in regard to the type of reporting bias assessed (eg, bias due to selective publication, bias due to selective non-reporting), and the level of assessment (eg, for the study as a whole, a particular result within a study or a particular synthesis of studies). Various criteria are used across tools to designate a synthesis as being at 'high' risk of bias due to selective publication (eg, evidence of funnel plot asymmetry, use of non-comprehensive searches). However, the relative weight assigned to each criterion in the overall judgement is unclear for most of these tools. Tools for assessing risk of bias due to selective non-reporting guide users to assess a study, or an outcome within a study, as 'high' risk of bias if no results are reported for an outcome. However, assessing the corresponding risk of bias in a synthesis that is missing the non-reported outcomes is outside the scope of most of these tools. Inter-rater agreement estimates were available for five tools.

Conclusion There are several limitations of existing tools for assessing risk of reporting biases, in terms of their scope, guidance for reaching risk of bias judgements and measurement properties. Development and evaluation of a new, comprehensive tool could help overcome present limitations.

Strengths and limitations of this study

- Tools for assessing risk of reporting biases, and studies evaluating their measurement properties, were identified by searching several relevant databases using a search string developed in conjunction with an information specialist.
- Detailed information on the content and measurement properties of existing tools was collected, providing readers with pertinent information to help decide which tools to use in evidence syntheses.
- Screening of articles and data collection were performed by one author only, so it is possible that some relevant articles were missed, or that errors in data collection were made.
- The search of grey literature was not comprehensive, so it is possible that there are other tools for assessing risk of reporting biases, and unpublished studies evaluating measurement properties, that were omitted from this review.

BACKGROUND

The credibility of evidence syntheses can be compromised by reporting biases, which arise when dissemination of research findings is influenced by the nature of the results.[1] For example, there may be bias due to selective publication, where a study is only published if the findings are considered interesting (also known as publication bias).[2] In addition, bias due to selective non-reporting may occur, where findings (eg, estimates of intervention efficacy or an association between exposure and outcome) that are statistically non-significant are not reported or are partially reported in a paper (eg, stating only that 'P>0.05').[3] Alternatively, there may be bias in selection of the reported result, where authors perform multiple analyses for a particular outcome/association, yet only report the result which yielded the most favourable effect estimate.[4] Evidence from cohorts of clinical trials followed from inception suggest that biased dissemination is common. Specifically, on average, half of all trials are not published,[1 5] trials with statistically significant results are twice as likely to be published[5] and a third of trials have outcomes that are omitted, added or modified between protocol and publication.[6]

Audits of systematic review conduct suggest that most systematic reviewers do not assess risk of reporting biases.[7–10] For example, in a cross-sectional study of 300 systematic reviews indexed in MEDLINE in February 2014,[7] the risk of bias due to selective publication was not

considered in 56% of reviews. A common reason for not doing so was that the small number of included studies, or inability to perform a meta-analysis, precluded the use of funnel plots. Only 19% of reviews included a search of a trial registry to identify completed but unpublished trials or prespecified but non-reported outcomes, and only 7% included a search of another source of data disseminated outside of journal articles. The risk of bias due to selective non-reporting in the included studies was assessed in only 24% of reviews.[7] Another study showed that authors of Cochrane reviews routinely record whether any outcomes that were measured were not reported in the included trials, yet rarely consider if such non-reporting could have biased the results of a synthesis.[11]

Previous researchers have summarised the characteristics of tools designed to assess various sources of bias in randomised trials,[12–14] non-randomised studies of interventions (NRSI),[14 15] diagnostic test accuracy studies[16] and systematic reviews.[14 17] Others have summarised the performance of statistical methods developed to detect or adjust for reporting biases.[18–20] However, no prior review has focused specifically on tools (ie, structured instruments such as scales, checklists or domain-based tools) for assessing the risk of reporting biases. A particular challenge when assessing risk of reporting biases is that existing tools vary in their level of assessment. For example, tools for assessing risk of bias due to selective publication direct assessments at the level of the synthesis, whereas tools for assessing risk of bias due to selective non-reporting within studies can direct assessments at the level of the individual study, at the level of the synthesis or at both levels. It is unclear how many tools are available to assess different types of reporting bias, and what level they direct assessments at. It is also unclear whether criteria for reaching risk of bias judgements are consistent across existing tools. Therefore, the aim of this research was to conduct a systematic review of the content and measurement properties of such tools.

METHODS

Protocol

Methods for this systematic review were prespecified in a protocol which was uploaded to the Open Science Framework in February 2017 (https://osf.io/9ea22/).

Eligibility criteria

Papers were included if the authors described a tool that was designed for use by individuals performing evidence syntheses to assess risk of reporting biases in the included studies or in their synthesis of studies. Tools could assess any type of reporting bias, including bias due to selective publication, bias due to selective non-reporting or bias in selection of the reported result. Tools could assess the risk of reporting biases in any type of study (eg, randomised trial of intervention, diagnostic test accuracy study, observational study estimating prevalence of an exposure) and in any type of result (eg, estimate of intervention efficacy or harm, estimate of diagnostic accuracy, association between exposure and outcome). Eligible tools could take any form, including scales, checklists and domain-based tools. To be considered a scale, each item had to have a numeric score attached to it, so that an overall summary score could be calculated.[12] To be considered a checklist, the tool had to include multiple questions, but the developers' intention was not to attach a numerical score to each response, or to calculate an overall score.[13] Domain-based tools were those that required users to judge risk of bias or quality within specific domains, and to record the information on which each judgement was based.[21]

Tools with a broad scope, for example, to assess multiple sources of bias or the overall quality of the body of evidence, were eligible if one of the items covered risk of reporting bias. Multidimensional tools with a statistical component were also eligible (eg, those that require users to respond to a set of questions about the comprehensiveness of the search, as well as to perform statistical tests for funnel plot asymmetry). In addition, any studies that evaluated the measurement properties of existing tools (eg, construct validity, inter-rater agreement, time taken to complete assessments) were eligible for inclusion. Papers were eligible regardless of the date or format of publication, but were limited to those written in English.

The following were ineligible:

- articles or book chapters providing guidance on how to address reporting biases, but which do not include a structured tool that can be applied by users (eg, the 2011 Cochrane Handbook chapter on reporting biases[22]);
- tools developed or modified for use in one particular systematic review;
- tools designed to appraise published systematic reviews, such as the Risk Of Bias In Systematic reviews (ROBIS) tool[23] or A MeaSurement Tool to Assess systematic Reviews (AMSTAR)[24];
- articles that focus on the development or evaluation of statistical methods to detect or adjust for reporting biases, as these have been reviewed elsewhere.[18–20]

Search methods

On 9 February 2017, one author (MJP) searched for potentially relevant records in Ovid MEDLINE (January 1946 to February 2017), Ovid Embase (January 1980 to February 2017) and Ovid PsycINFO (January 1806 to February 2017). The search strategies included terms relating to reporting bias which were combined with a search string used previously by Whiting *et al* to identify risk of bias/quality assessment tools[17] (see full Boolean search strategies in online supplementary table S1).

To capture any tools not published by formal academic publishers, we searched Google Scholar using the phrase 'reporting bias tool OR risk of bias'. One author (MJP) screened the titles of the first 300 records, as recommended by Haddaway *et al*.[25] To capture any papers that may have been missed by all searches, one author (MJP)

screened the references of included articles. In April 2017, the same author emailed the list of included tools to 15 individuals with expertise in reporting biases and risk of bias assessment, and asked if they were aware of any other tools we had not identified.

Study selection and data collection

One author (MJP) screened all titles and abstracts retrieved by the searches. The same author screened any full-text articles retrieved. One author (MJP) collected data from included papers using a standardised data-collection form. The following data on included tools were collected:

- type of tool (scale, checklist or domain-based tool);
- types of reporting bias addressed by the tool;
- level of assessment (ie, whether users direct assessments at the synthesis or at the individual studies included in the synthesis);
- whether the tool is designed for general use (generic) or targets specific study designs or topic areas (specific);
- items included in the tool;
- how items within the tool are rated;
- methods used to develop the tool (eg, Delphi study, expert consensus meeting);
- availability of guidance to assist with completion of the tool (eg, guidance manual).

The following data from studies evaluating measurement properties of an included tool were collected:

- tool evaluated
- measurement properties evaluated (eg, inter-rater agreement)
- number of syntheses/studies evaluated
- publication year of syntheses/studies evaluated
- areas of healthcare addressed by syntheses/studies evaluated
- number of assessors
- estimate (and precision) of psychometric statistics (eg, weighted kappa; κ).

Data analysis

We summarised the characteristics of included tools in tables. We calculated the median (IQR) number of items across all tools, and tabulated the frequency of different criteria used in tools to denote a judgement of 'high' risk of reporting bias. We summarised estimates of psychometric statistics, such as weighted κ to estimate inter-rater agreement,[26] by reporting the range of values across studies. For studies reporting weighted κ, we categorised agreement according to the system proposed by Landis and Koch,[27] as poor (0.00), slight (0.01–0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80) or almost perfect (0.81–1.00).

RESULTS

In total, 5554 records were identified from the searches, of which we retrieved 165 for full-text screening (figure 1). The inclusion criteria were met by 42 reports summarising 18 tools (table 1) and 17 studies evaluating the measurement properties of tools.[3 4 21 28–66] A list of excluded papers is presented in online supplementary table S2. No additional tools were identified by the 15 experts contacted.

General characteristics of included tools

Nearly all of the included tools (16/18; 89%) were domain-based, where users judge risk of bias or quality within specific domains (table 2; individual characteristics of each tool are presented in online supplementary table S3). All tools were designed for generic rather than specific use. Five tools focused solely on the risk of reporting biases[3 28 29 47 48]; the remainder addressed reporting biases and other sources of bias/methodological quality (eg, problems with randomisation, lack of blinding). Half of the tools (9/18; 50%) addressed only one type of reporting bias (eg, bias due to selective non-reporting only). Tools varied in regard to the study design that they assessed (ie, randomised trial, non-randomised study of an intervention, laboratory animal experiment). The publication year of the tools ranged from 1998 to 2016 (the earliest was the Downs-Black tool,[31] a 27-item tool assessing multiple sources of bias, one of which focuses on risk of bias in the selection of the reported result).

Assessments for half of the tools (9/18; 50%) are directed at an individual study (eg, tool is used to assess whether *any outcomes in a study* were not reported). In 5/18 (28%) tools, assessments are directed at a specific outcome or result within a study (eg, tool is used to assess whether *a particular outcome in a study*, such as pain, was not reported). In a few tools (4/18; 22%), assessments are directed at a specific synthesis (eg, tool is used to assess whether *a particular synthesis*, such as a meta-analysis of studies examining pain as an outcome, is missing unpublished studies).

The content of the included tools was informed by various sources of data. The most common included a literature review of items used in existing tools or a literature review of empirical evidence of bias (9/18; 50%), ideas generated at an expert consensus meeting (8/18; 44%) and pilot feedback on a preliminary version of the tool (7/18; 39%). The most common type of guidance available for the tools was a brief annotation per item/response option (9/18; 50%). A detailed guidance manual is available for four (22%) tools.

Tool content

Four tools include items for assessing risk of bias due to both selective publication and selective non-reporting.[29 33 45 49] One of these tools (the AHRQ tool for evaluating the risk of reporting bias[29]) directs users to assess a particular synthesis, where a single risk of bias judgement is made based on information about unpublished studies and under-reported outcomes. In the other three tools (the GRADE framework, and two others which are based on GRADE),[33 45 49] the different sources of

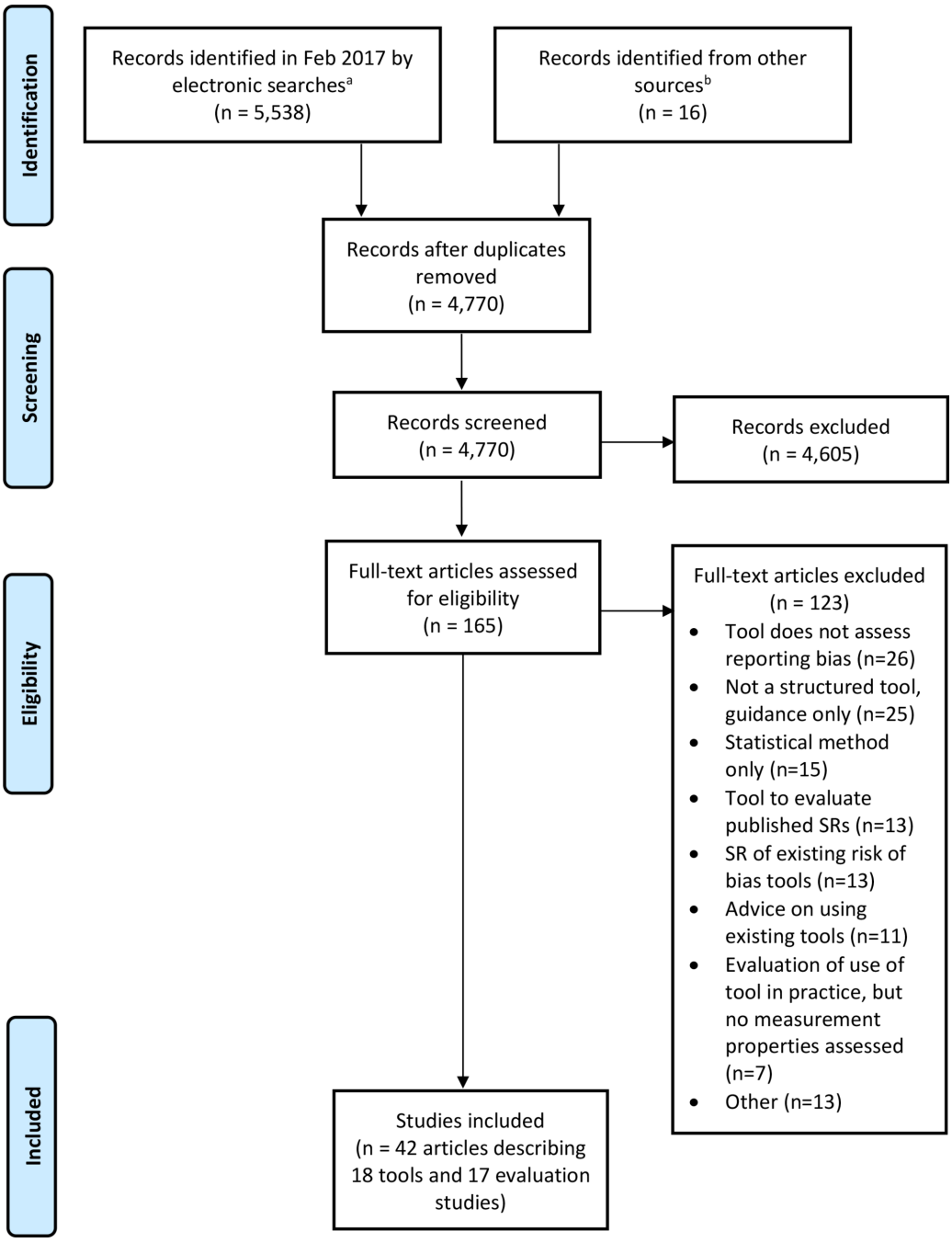

Figure 1 Flow diagram of identification, screening and inclusion of studies. [a]Records identified from Ovid MEDLINE, Ovid Embase, Ovid PsycINFO and Google Scholar. [b]Records identified from screening references of included articles. SR, systematic review.

reporting bias are assessed in separate domains (bias due to selective non-reporting is considered in a 'study limitations (risk of bias)' domain, while bias due to selective publication is considered in a 'publication bias' domain).

Five tools[21 28 43 44 47] guide users to assess risk of bias due to both selective non-reporting and selection of the reported result (ie, problems with outcomes/results that *are not* reported and those that *are* reported, respectively). Four of these tools, which include the Cochrane risk of bias tool for randomised trials[21] and three others which are based on the Cochrane tool,[43 44 47] direct assessments at the study level. That is, a whole study is rated at 'high' risk of reporting bias if *any* outcome/result in the study has been omitted, or fully reported, on the basis of the findings.

Some of the tools designed to assess the risk of bias due to selective non-reporting ask users to assess, for particular outcomes of interest, whether the outcome was not reported or only partially reported in the study on the basis of its results (eg, Outcome Reporting Bias In Trials (ORBIT) tools,[3 48] the AHRQ outcome reporting bias framework,[28] and GRADE.[34] This allows

Table 1 List of included tools

Article ID	Tool	Scope of tool	Types of reporting biases assessed			Level of assessment*
			Selective publication	Selective non-reporting	Selection of the reported result	
Balshem *et al*[28]	Agency for Healthcare Research and Quality (AHRQ) outcome and analysis reporting bias framework	Reporting bias only		✓	✓	Specific outcome/result in a study
Berkman *et al*[29]	AHRQ tool for evaluating the risk of reporting bias	Reporting bias only	✓	✓		Specific synthesis of studies
Downes *et al*[30]	Appraisal tool for Cross-Sectional Studies (AXIS) tool	Multiple sources of bias		✓		Study
Downs and Black[31]	Downs-Black tool	Multiple sources of bias			✓	Study
Guyatt *et al*[33–37]	Grading of Recommendations Assessment, Development and Evaluation (GRADE)	Multiple sources of bias	✓	✓		Specific synthesis of studies
Hayden *et al*[38]	Quality In Prognosis Studies (QUIPS) tool	Multiple sources of bias		✓		Study
Higgins *et al*[21 39 40]	Cochrane risk of bias tool for randomised trials (RoB 1.0)	Multiple sources of bias		✓	✓	Study
Higgins *et al*[41 42]	RoB 2.0 revised tool for assessing risk of bias in randomised trials	Multiple sources of bias			✓	Specific outcome/result in a study
Hoojimans *et al*[43]	SYstematic Review Centre for Laboratory animal Experimentation (SYRCLE) RoB tool	Multiple sources of bias		✓	✓	Study
Kim *et al*[44]	Risk of Bias Assessment Tool for Nonrandomized Studies (RoBANS)	Multiple sources of bias		✓	✓	Study
Kirkham *et al*[3 32]	Outcome Reporting Bias In Trials I (ORBIT-I) classification system for benefit outcomes	Reporting bias only		✓		Specific outcome/result in a study
Meader *et al*[45 46]	Semi-Automated Quality Assessment Tool (SAQAT)	Multiple sources of bias	✓	✓		Specific synthesis of studies
Reid *et al*[47]	Selective reporting bias algorithm	Reporting bias only		✓	✓	Study
Saini *et al*[48]	ORBIT-II classification system for harm outcomes	Reporting bias only		✓		Specific outcome/result in a study
Salanti *et al*[49 50]	Framework for evaluating the quality of evidence from a network meta-analysis	Multiple sources of bias	✓	✓		Specific synthesis of studies
Sterne *et al*[4]	Risk Of Bias In Non-randomized Studies of Interventions I (ROBINS-I) tool	Multiple sources of bias			✓	Specific outcome/result in a study
Viswanathan and Berkman[51]	Research Triangle Institute (RTI) item bank for assessment of risk of bias and precision for observational studies of interventions or exposures	Multiple sources of bias		✓		Study
Viswanathan et al[52]	RTI item bank for assessing risk of bias and confounding for observational studies of interventions or exposures	Multiple sources of bias		✓		Study

*Level of assessment classified as: 'study' when assessments are directed at a study as a whole (eg, tool used to assess whether *any* outcomes in a study were not reported); 'specific outcome/result in a study' when assessments are directed at a specific outcome or result within a study (eg, tools used to assess whether a particular outcome, such as pain, was not reported) or 'specific synthesis of studies' when assessments are directed at a specific synthesis (eg, tool used to assess whether a particular synthesis, such as a meta-analysis of pain, is missing unpublished studies).

users to perform multiple outcome-level assessments of the risk of reporting bias (rather than one assessment for the study as a whole). In total, 15 tools include a mechanism for assessing risk of bias due to selective non-reporting in studies, but assessing the corresponding risk of bias in a synthesis that is missing the

Table 2 Summary of general characteristics of included tools

Characteristic	Summary data (n=18 tools)
Type of tool	
Domain-based	16 (89%)
Checklist	1 (6%)
Scale	1 (6%)
Scope of tool	
Assessment of reporting bias only	5 (28%)
Assessment of multiple sources of bias/ quality	13 (72%)
Types of reporting bias assessed	
Bias due to selective publication only	0 (0%)
Bias due to selective non-reporting only	6 (33%)
Bias in selection of the reported result only	3 (17%)
Bias due to selective publication and bias due to selective non-reporting	4 (22%)
Bias due to selective non-reporting and bias in selection of the reported result	5 (28%)
Total number of items in the tool	7 (5–13)
Number of items relevant to risk of reporting bias	1 (1–2)
Number of response options for risk of reporting bias judgement	3 (3–3)
Types of study designs to which the tool applies	
Randomised trials only	5 (28%)
Systematic reviews only	3 (17%)
Non-randomised studies of interventions only	2 (11%)
Randomised trials and non-randomised studies of interventions	2 (11%)
Non-randomised studies of interventions or exposures	2 (11%)
Other (cross-sectional studies, animal studies, network meta-analyses, prognosis studies)	4 (22%)
Level of assessment of risk of reporting bias	
Study as a whole	9 (50%)
Specific outcome/result in a study	5 (28%)
Specific synthesis of studies	4 (22%)
Data sources used to inform tool content*	
Literature review (eg, of items in existing tools or empirical evidence)	9 (50%)
Ideas generated at expert consensus meeting	8 (44%)
Pilot feedback on preliminary version of the tool	7 (39%)
Data from psychometric or cognitive testing†	5 (28%)
Other (eg, adaptation of existing tool)	5 (28%)
Delphi study responses	2 (11%)
No methods stated	2 (11%)
Guidance available	
Brief annotation per item/response option	9 (50%)
Detailed guidance manual	4 (22%)
Worked example for each response option	2 (11%)
Detailed annotation per item/response option	1 (6%)
None	2 (11%)

Summary data given as number (%) or median (IQR).
*The percentages in this category do not sum to 100% since the development of some tools was informed by multiple data sources.
†Psychometric testing includes any evaluation of the measurement properties (eg, construct validity, inter-rater reliability, test–retest reliability) of a draft version of the tool. Cognitive testing includes use of qualitative methods (eg, interview) to explore whether assessors who are using the tool for the first time were interpreting the tool and guidance as intended.

non-reported outcomes is not within the scope of 11 of these tools.[3 21 28 30 38 43 44 47 48 51 52]

A variety of criteria are used in existing tools to inform a judgement of 'high' risk of bias due to selective publication (table 3), selective non-reporting (table 4), and selection of the reported result (table 5; more detail is provided in online supplementary table S4). In the four tools with an assessment of risk of bias due to selective publication, 'high' risk criteria include evidence of funnel plot asymmetry, discrepancies between published and unpublished studies, use of non-comprehensive searches and presence of small, 'positive' studies with for-profit interest (table 3). However, not all of these criteria appear in all tools (only evidence of funnel plot asymmetry does), and the relative weight assigned to each criterion in the overall risk of reporting bias judgement is clear for only one tool (the Semi-Automated Quality Assessment Tool; SAQAT).[45 46]

All 15 tools with an assessment of the risk of bias due to selective non-reporting suggest that the risk of bias is 'high' when it is clear that an outcome was measured but no results were reported (table 4). Fewer of these tools (n=8; 53%) also recommend a 'high' risk judgement when results for an outcome are partially reported (eg, it is stated that the result was non-significant, but no effect estimate or summary statistics are presented).

The eight tools that include an assessment of the risk of bias in selection of the reported result recommend various criteria for a 'high' risk judgement (table 5). These include when some outcomes that were not

Table 3 Criteria used in existing tools to inform a judgement of 'high' risk of bias due to selective publication

'High' risk of bias criteria proposed in existing tools	AHRQ RRB	GRADE	SAQAT	NMA-Quality	Total, n (%)
Assessment directed at a specific synthesis (eg, meta-analysis)					
Evidence of funnel plot asymmetry (based on visual inspection of funnel plot or statistical test for funnel plot asymmetry)	✓	✓	✓	✓	4 (100)
Smaller studies tend to demonstrate more favourable results (based on visual assessment, without funnel plot)	✓				1 (25)
Clinical decision would differ for estimates from a fixed-effect versus a random-effects model because the findings from a fixed-effect model are closer to the null	✓				1 (25)
Substantial heterogeneity in the meta-analysis cannot be explained by some clinical or methodological factor	✓				1 (25)
At least one study is affected by non-publication or non-accessibility	✓				1 (25)
Presence of small (often 'positive') studies with for-profit interest in the synthesis		✓		✓	2 (50)
Presence of early studies (ie, set of small, 'positive' trials addressing a novel therapy) in the synthesis		✓		✓	2 (50)
Discrepancy in findings between published and unpublished trials		✓	✓	✓	3 (75)
Search strategies were not comprehensive		✓	✓	✓	3 (75)
Methods to identify all available evidence were not comprehensive		✓		✓	2 (50)
Grey literature were not searched			✓		1 (25)
Restrictions to study selection on the basis of language were applied			✓		1 (25)
Industry influence may apply to studies included in the synthesis			✓		1 (25)

AHRQ RRB, AHRQ tool for evaluating the risk of reporting bias[29]; GRADE, GRADE rating of quality of evidence[34–37]; NMA-Quality, Framework for evaluating the quality of evidence from a network meta-analysis[49]; SAQAT, Semi-Automated Quality Assessment Tool.[45 46]

prespecified are added post hoc (in 4 (50%) tools), or when it is likely that the reported result for a particular outcome has been selected, on the basis of the findings, from among multiple outcome measurements or analyses within the outcome domain (in 2 (25%) tools).

General characteristics of studies evaluating measurement properties of included tools

Despite identifying 17 studies that evaluated measurement properties of an included tool, psychometric statistics for the risk of reporting bias component were available only from 12 studies[43 44 54–60 62 64 66] (the other five studies include only data on properties of the multidimensional tool as a whole[31 53 61 63 65]; online supplementary table S5). Nearly all 12 studies (11; 92%) evaluated inter-rater agreement between two assessors; eight of these studies reported weighted κ values, but only two described the weighting scheme.[55 62] Eleven studies[43 44 54–60 64 66] evaluated the measurement properties of tools for assessing risk of bias in a study due to selective non-reporting or risk of bias in selection of the reported result; in these 11 studies, a median of 40 (IQR 32–109) studies were assessed. One study[62] evaluated a tool for assessing risk of bias in a synthesis due to selective publication, in which 44 syntheses were assessed. In the studies evaluating inter-rater agreement, all involved two assessors.

Results of evaluation studies

Five studies[54 56–58 60] included data on the inter-rater agreement of assessments of risk of bias due to selective non-reporting using the Cochrane risk of bias tool for randomised trials[21] (table 6). Weighted κ values in four studies[54 56–58] ranged from 0.13 to 0.50 (sample size ranged from 87 to 163 studies), suggesting slight to moderate agreement.[27] In the other study,[60] the per cent agreement in selective non-reporting assessments in trials that were included in two different Cochrane reviews was low (43% of judgements were in agreement). Two other studies found that inter-rater agreement of selective non-reporting assessments were substantial for SYRCLE's RoB tool (κ=0.62, n=32),[43] but poor for the RoBANS tool (κ=0, n=39).[44] There was substantial agreement between

Table 4 Criteria used in existing tools to inform a judgement of 'high' risk of bias due to selective non-reporting

'High' risk of bias criteria proposed in existing tools	AHRQ ORB	AHRQ RRB	AXIS	GRADE	QUIPS	RoB 1.0	SYRCLE RoB	RoBANS	ORBIT-I	SAQAT	Reid	ORBIT-II	NMA-Quality	RTI 2012	RTI 2013	Total, n (%)
Assessment directed at study as a whole																
One or more outcomes of interest were clearly measured, but no results were reported			✓		✓	✓	✓	✓			✓			✓	✓	8 (53)
One or more outcomes of interest were reported incompletely so that they could not be entered in a meta-analysis						✓		✓								2 (13)
The study report fails to include results for a key outcome that would be expected to have been reported for such a study						✓	✓	✓						✓	✓	5 (33)
Assessment directed at a specific outcome																
Particular outcome clearly measured but no results were reported	✓	✓		✓					✓			✓	✓			6 (40)
Particular outcome of interest is reported incompletely so that it cannot be entered in a meta-analysis (typically stating only that P>0.05)	✓	✓		✓					✓			✓	✓			6 (40)

Continued

Table 4 Continued

'High' risk of bias criteria proposed in existing tools	AHRQ ORB	AHRQ RRB	AXIS	GRADE	QUIPS	RoB 1.0	SYRCLE RoB	RoBANS	ORBIT-I	SAQAT	Reid	ORBIT-II	NMA-Quality	RTI 2012	RTI 2013	Total, n (%)
Judgement says particular outcome is likely to have been measured and analysed but not reported on the basis of its results	✓	✓		✓					✓			✓	✓			6 (40)
Composite outcomes are presented without the individual component outcomes				✓								✓				2 (13)
Result reported globally across all groups												✓				1 (7)
Result reported for some groups only												✓				1 (7)
Data were not reported consistently for the outcome of interest										✓						1 (7)
Assessment directed at a specific synthesis																
Selective non-reporting suspected in a number of included studies		✓		✓						✓			✓			4 (27)

AHRQ ORB, AHRQ outcome and analysis reporting bias framework[28]; AHRQ RRB, AHRQ tool for evaluating the risk of reporting bias[29]; AXIS, Appraisal tool for Cross-Sectional Studies[30]; GRADE, GRADE rating of quality of evidence[34–37]; NMA-Quality, Framework for evaluating the quality of evidence from a network meta-analysis[49]; ORBIT-I, Outcome Reporting Bias In Trials classification system for benefit outcomes[3 32]; ORBIT-II, Outcome Reporting Bias In Trials classification system for harm outcomes[48]; QUIPS, Quality In Prognosis Studies tool[38]; Reid, Reid *et al* selective reporting bias algorithm[47]; RoB 1.0, Cochrane risk of bias tool for randomised trials[21 39 40]; RoBANS, Risk of Bias Assessment Tool for Nonrandomized Studies[44]; RTI 2012, RTI Item Bank for Assessment of Risk of Bias and Precision for Observational Studies of Interventions or Exposures[51]; RTI 2013, RTI Item Bank for Assessing Risk of Bias and Confounding for Observational Studies of Interventions or Exposures[52]; SAQAT, Semi-Automated Quality Assessment Tool[45 46]; SYRCLE RoB, SYstematic Review Centre for Laboratory animal Experimentation risk of bias tool.[43]

Table 5 Criteria used in existing tools to inform a judgement of 'high' risk of bias in selection of the reported result

'High' risk of bias criteria proposed in existing tools	AHRQ ORB	Downs-Black	RoB 1.0	RoB 2.0	SYRCLE RoB	RoBANS	Reid	ROBINS-I	Total, n (%)
Assessment directed at study as a whole									
One or more reported outcomes were not prespecified (unless clear justification for their reporting is provided, such as an unexpected adverse event)			✓		✓	✓	✓		4 (50)
One or more outcomes were reported using measurements, analysis methods or subsets of the data (eg, subscales) that were not prespecified			✓		✓				2 (15)
One or more retrospective, unplanned, subgroup analyses were reported		✓							1 (13)
Any analyses that had not been planned at the outset of the study were not clearly indicated		✓							1 (13)
Assessment directed at a specific outcome/result									
Particular outcome was not prespecified but results were reported	✓								1 (13)
Reported result for a particular outcome is likely to have been selected, on the basis of the findings, from multiple outcome measurements (eg, scales, definitions, time points) within the outcome domain				✓				✓	2 (25)
Reported result for a particular outcome is likely to have been selected, on the basis of the findings, from multiple analyses of the data				✓				✓	2 (25)
Reported result for a particular outcome is likely to have been selected, on the basis of the findings, from different subgroups								✓	1 (13)

AHRQ ORB, AHRQ outcome and analysis reporting bias framework[28]; Downs-Black, Downs Black tool[31]; Reid, Reid *et al* selective reporting bias algorithm[47]; RoB 1.0, Cochrane risk of bias tool for randomised trials[21 39 40]; RoB 2.0, Revised tool for assessing risk of bias in randomised trials[41 42]; RoBANS, Risk of Bias Assessment Tool for Nonrandomized Studies[44]; ROBINS-I, Risk Of Bias In Non-randomized Studies of Interventions tool[4]; SYRCLE RoB, SYstematic Review Centre for Laboratory animal Experimentation risk of bias tool.[43]

Table 6 Reported measurement properties of tools with an assessment of the risk of reporting bias

Study ID	Tool	Measurement property	Sample size	Areas of healthcare addressed	Weighted kappa (95% CI)	Weighting scheme	Interpretation of kappa*
Armijo-Olivo *et al*[54]	RoB 1.0	Inter-rater agreement of assessments of risk of bias due to selective non-reporting (between two external reviewers)	87	Musculoskeletal, cardiorespiratory, neurological and gynaecological conditions.	0.5 (CI not reported)	Not described	Moderate agreement
Armijo-Olivo *et al*[54]	RoB 1.0	Inter-rater agreement of assessments of risk of bias due to selective non-reporting (between two external reviewers and Cochrane reviewers)	87	See above	0.13 (CI not reported)	Not described	Slight agreement
Hartling *et al*[56]	RoB 1.0	Inter-rater agreement of assessments of risk of bias due to selective non-reporting	163	Child health	0.13 (95% CI −0.05 to 0.31)	Not described	Slight agreement
Hartling *et al*[57]	RoB 1.0	Inter-rater agreement of assessments of risk of bias due to selective non-reporting	107	Asthma	0.4 (95% CI 0.14 to 0.67)	Not described	Fair agreement
Hartling *et al*[58 59]	RoB 1.0	Inter-rater agreement of assessments of risk of bias due to selective non-reporting (between two reviewers, all trials)	124	Varied	0.27 (95% CI 0.06 to 0.49)	Not described	Fair agreement
Hartling *et al*[58 59]	RoB 1.0	Inter-rater agreement of assessments of risk of bias due to selective non-reporting (between pairs of reviewers across different centres, all trials)	30	Varied	0.08 (95% CI −0.09 to 0.26)	Not described	Slight agreement
Jordan *et al*[60]	RoB 1.0	Inter-rater agreement of assessments of risk of bias due to selective non-reporting (between judgements of trials appearing in two SRs)	28	Subfertility	Not reported†	Not applicable	Not applicable
Vale *et al*[66]	RoB 1.0	Agreement between selective non-reporting assessments performed using published article only versus published article and data collected during the individual participant data process	95	Cancer pain	Not reported†	Not applicable	Not applicable

Continued

Table 6 Continued

Study ID	Tool	Measurement property	Sample size	Areas of healthcare addressed	Weighted kappa (95% CI)	Weighting scheme	Interpretation of kappa*
Hoojimans *et al*[43]	SYRCLE RoB	Inter-rater agreement of assessments of risk of bias due to selective non-reporting	32	Animal studies (not specified)	0.62 (CI not reported)	Not described	Substantial agreement
Kim *et al*[44]	RoBANS	Inter-rater agreement of assessments of risk of bias due to selective non-reporting	39	Depression, myocardial infarction, postpartum haemorrhage, chronic non-cancer pain	0 (CI not reported)	Not described	Poor agreement
Llewellyn *et al*[62]	SAQAT	Inter-rater agreement of assessments of risk of bias due to selective publication (between two SAQAT raters)	29	Varied	0.63 (95% CI 0.17 to 1)	Quadratic	Substantial agreement
Llewellyn *et al*[62]	SAQAT	Inter-rater agreement of assessments of risk of bias due to selective publication (between one rater using SAQAT and one using the standard GRADE approach)	15	Varied	Not reported†	Not applicable	Not applicable
Norris *et al*[64]	ORBIT-I	Inter-rater agreement of ORBIT-I classifications of risk of bias due to selective non-reporting	40	Varied	Not calculated, as too little variation in judgements	Not applicable	Not applicable
Bilandzic *et al*[55]	ROBINS-I	Inter-rater agreement of assessments of risk of bias in selection of the reported result	16	Thiazolidinediones and cardiovascular events	0.78 (CI not reported)	Linear	Substantial agreement
Bilandzic *et al*[55]	ROBINS-I	Inter-rater agreement of assessments of risk of bias in selection of the reported result	21	COX-2 inhibitors and cardiovascular events	0.45 (CI not reported)	Linear	Moderate agreement

*Interpretation of kappa based on categorisation system defined by Landis *et al*.[27]
†Data presented as per cent agreement, not weighted kappa.
COX-2, cyclooxygenase-2; ORBIT-I, Outcome Reporting Bias In Trials classification system for benefit outcomes[3 32]; RoB 1.0, Cochrane risk of bias tool for randomised trials[21 39 40]; RoBANS, Risk of Bias Assessment Tool for Nonrandomized Studies[44]; ROBINS-I, Risk Of Bias In Non-randomized Studies of Interventions tool[4]; SAQAT, Semi-Automated Quality Assessment Tool[45 46]; SRs, systematic reviews; SYRCLE RoB, SYstematic Review Centre for Laboratory animal Experimentation risk of bias tool.[43]

raters in the assessment of risk of bias due to selective publication using the SAQAT (κ=0.63, n=29).[62] The inter-rater agreement of assessments of risk of bias in selection of the reported result using the ROBINS-I tool[4] was moderate for NRSI included in a review of the effect of cyclooxygenase-2 inhibitors on cardiovascular events (κ=0.45, n=21), and substantial for NRSI included in a review of the effect of thiazolidinediones on cardiovascular events (κ=0.78, n=16).[55]

DISCUSSION

From a systematic search of the literature, we identified 18 tools designed for use by individuals performing evidence syntheses to assess risk of reporting biases in the included studies or in their synthesis of studies. The tools varied with regard to the type of reporting bias assessed (eg, bias due to selective publication, bias due to selective non-reporting), and the level of assessment (eg, for the study as a whole, a particular outcome within a study or a particular synthesis of studies). Various criteria are used across tools to designate a synthesis as being at 'high' risk of bias due to selective publication (eg, evidence of funnel plot asymmetry, use of non-comprehensive searches). However, the relative weight assigned to each criterion in the overall judgement is not clear for most of these tools. Tools for assessing risk of bias due to selective non-reporting guide users to assess a study, or an outcome within a study, as 'high' risk of bias if no results are reported for an outcome. However, assessing the corresponding risk of bias in a synthesis that is missing the non-reported outcomes is outside the scope of most of these tools. Inter-rater agreement estimates were available for five tools,[4 21 43 44 62] and ranged from poor to substantial; however, the sample sizes of most evaluations were small, and few described the weighting scheme used to calculate κ.

Strengths and limitations

There are several strengths of this research. Methods were conducted in accordance with a systematic review protocol (https://osf.io/9ea22/). Published articles were identified by searching several relevant databases using a search string developed in conjunction with an information specialist,[17] and by contacting experts to identify tools missed by the search. Detailed information on the content and measurement properties of existing tools was collected, providing readers with pertinent information to help decide which tools to use in future reviews. However, the findings need to be considered in light of some limitations. Screening of articles and data collection were performed by one author only. It is therefore possible that some relevant articles were missed, or that errors in data collection were made. The search for unpublished tools was not comprehensive (only Google Scholar was searched), so it is possible that other tools for assessing risk of reporting biases exist. Further, restricting the search to articles in English was done to expedite the review process, but may have resulted in loss of information about tools written in other languages, and additional evidence on measurement properties of tools.

Comparison with other studies

Other systematic reviews of risk of bias tools[12–17] have restricted inclusion to tools developed for particular study designs (eg, randomised trials, diagnostic test accuracy studies), where the authors recorded all the sources of bias addressed. A different approach was taken in the current review, where all tools (regardless of study design) that address a particular source of bias were examined. By focusing on one source of bias only, the analysis of included items and criteria for risk of bias judgements was more detailed than that recorded previously. Some of the existing reviews of tools[15] considered tools that were developed or modified in the context of a specific systematic review. However, such tools were excluded from the current review as they are unlikely to have been developed systematically,[15 67] and are difficult to find (all systematic reviews conducted during a particular period would need to have been examined for the search to be considered exhaustive).

Explanations and implications

Of the 18 tools identified, only four (22%) included a mechanism for assessing risk of bias due to selective publication, which is the type of reporting bias that has been investigated by methodologists most often.[2] This is perhaps unsurprising given that hundreds of statistical methods to 'detect' or 'adjust' for bias due to selective publication have been developed.[18] These statistical methods may be considered by methodologists and systematic reviewers as the tools of choice for assessing this type of bias. However, application of these statistical methods without considering other factors (eg, existence of registered but unpublished studies, conflicts of interest that may influence investigators to not disseminate studies with unfavourable results) is not sufficiently comprehensive, and could lead to incorrect conclusions about the risk of bias due to selective publication. Further, there are many limitations of these statistical approaches, in terms of their underlying assumptions, statistical power, which is often low because most meta-analyses include few studies,[7] and the need for specialist statistical software to apply them.[19 68] These factors may have limited their use in practice and potentially explain why a large number of systematic reviewers currently ignore the risk of bias due to selective publication.[7–9 69]

Our analysis suggests that the factors that need to be considered to assess risk of reporting biases adequately (eg, comprehensiveness of the search, amount of data missing from the synthesis due to unpublished studies and under-reported outcomes) are fragmented. A similar problem was occurring a decade ago with the assessment of risk of bias in randomised trials. Some authors assessed only problems with randomisation, while others focused on whether trials were not

'double blinded' or had any missing participant data.[70] It was not until all the important bias domains were brought together into a structured, domain-based tool to assess the risk of bias in randomised trials,[21] that systematic reviewers started to consider risk of bias in trials comprehensively. A similar initiative to link all the components needed to judge the risk of reporting biases into a comprehensive new tool may improve the credibility of evidence syntheses.

In particular, there is an emergent need for a new tool to assess the risk that a synthesis is affected by reporting biases. This tool could guide users to consider risk of bias in a synthesis due to both selective publication and selective non-reporting, given that both practices lead to the same consequence: evidence missing from the synthesis.[11] Such a tool would complement recently developed tools for assessing risk of bias within studies (RoB 2.0[41] and ROBINS-I[4] which include a domain for assessing the risk of bias in selection of the reported result, but no mechanism to assess risk of bias due to selective non-reporting). Careful thought would need to be given as to how to weigh up various pieces of information underpinning the risk of bias judgement. For example, users will need guidance on how evidence of known, unpublished studies (as identified from trial registries, protocols or regulatory documents) should be considered alongside evidence that is more speculative (eg, funnel plots suggesting that studies may be missing). Further, guidance for the tool will need to emphasise the value of seeking documents other than published journal articles (eg, protocols) to inform risk of bias judgements. Preparation of a detailed guidance manual may enhance the usability of the tool, minimise misinterpretation and increase reliability in assessments. Once developed, evaluations of the measurement properties of the tool, such as inter-rater agreement and construct validity, should be conducted to explore whether modifications to the tool are necessary.

CONCLUSIONS

There are several limitations of existing tools for assessing risk of reporting biases in studies or syntheses of studies, in terms of their scope, guidance for reaching risk of bias judgements and measurement properties. Development and evaluation of a new, comprehensive tool could help overcome present limitations.

Contributors MJP conceived and designed the study, collected data, analysed the data and wrote the first draft of the article. JEM and JPTH provided input on the study design and contributed to revisions of the article. All authors approved the final version of the submitted article.

Funding MJP is supported by an Australian National Health and Medical Research Council (NHMRC) Early Career Fellowship (1088535). JEM is supported by an NHMRC Australian Public Health Fellowship (1072366). JPTH is funded in part by Cancer Research UK Programme Grant C18281/A19169; is a member of the MRC Integrative Epidemiology Unit at the University of Bristol, which is supported by the UK Medical Research Council and the University of Bristol (grant MC_UU_12013/9); and is a member of the MRC ConDuCT-II Hub (Collaboration and innovation for Difficult and Complex randomised controlled Trials In Invasive procedures; grant MR/K025643/1).

Competing interests JPTH led or participated in the development of four of the included tools (the current Cochrane risk of bias tool for randomised trials, the RoB 2.0 tool for assessing risk of bias in randomised trials, the ROBINS-I tool for assessing risk of bias in non-randomised studies of interventions and the framework for assessing quality of evidence from a network meta-analysis). MJP participated in the development of one of the included tools (the RoB 2.0 tool for assessing risk of bias in randomised trials). All authors are participating in the development of a new tool for assessing risk of reporting biases in systematic reviews.

Patient consent Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The study protocol, data collection form, and the raw data and statistical analysis code for this study are available on the Open Science Framework: https://osf.io/3jdaa/

REFERENCES

1. Chan AW, Song F, Vickers A, *et al*. Increasing value and reducing waste: addressing inaccessible research. *Lancet* 2014;383:257–66.
2. Song F, Parekh S, Hooper L, *et al*. Dissemination and publication of research findings: an updated review of related biases. *Health Technol Assess* 2010;14:8.
3. Kirkham JJ, Dwan KM, Altman DG, *et al*. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ* 2010;340:c365.
4. Sterne JA, Hernán MA, Reeves BC, *et al*. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 2016;355:i4919.
5. Schmucker C, Schell LK, Portalupi S, *et al*. Extent of non-publication in cohorts of studies approved by research ethics committees or included in trial registries. *PLoS One* 2014;9:e114023.
6. Jones CW, Keil LG, Holland WC, *et al*. Comparison of registered and published outcomes in randomized controlled trials: a systematic review. *BMC Med* 2015;13:282.
7. Page MJ, Shamseer L, Altman DG, *et al*. Epidemiology and reporting characteristics of systematic reviews of biomedical research: a cross-sectional study. *PLoS Med* 2016;13:e1002028.
8. Koletsi D, Valla K, Fleming PS, *et al*. Assessment of publication bias required improvement in oral health systematic reviews. *J Clin Epidemiol* 2016;76:118–24.
9. Hedin RJ, Umberham BA, Detweiler BN, *et al*. Publication Bias and Nonreporting Found in Majority of Systematic Reviews and Meta-analyses in Anesthesiology Journals. *Anesth Analg* 2016;123:1018–25.
10. Ziai H, Zhang R, Chan AW, *et al*. Search for unpublished data by systematic reviewers: an audit. *BMJ Open* 2017;7:e017737.
11. Page MJ, Higgins JP. Rethinking the assessment of risk of bias due to selective reporting: a cross-sectional study. *Syst Rev* 2016;5:108.
12. Moher D, Jadad AR, Nichol G, *et al*. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Control Clin Trials* 1995;16:62–73.
13. Olivo SA, Macedo LG, Gadotti IC, *et al*. Scales to assess the quality of randomized controlled trials: a systematic review. *Phys Ther* 2008;88:156–75.
14. Bai A, Shukla VK, Bak G, *et al*. *Quality Assessment Tools Project Report*. Ottawa: Canadian Agency for Drugs and Technologies in Health, 2012.
15. Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. *Int J Epidemiol* 2007;36:666–76.

16. Whiting P, Rutjes AW, Dinnes J, *et al*. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. *J Clin Epidemiol* 2005;58:1–12.
17. Whiting P, Davies P, Savovic J, *et al*. Evidence to inform the development of ROBIS, a new tool to assess the risk of bias in systematic reviews, September 2013. 2013 https://www.researchgate.net/publication/303312018_Evidence_to_inform_the_development_of_ROBIS_a_new_tool_to_assess_the_risk_of_bias_in_systematic_reviews (accessed 1 Aug 2017).
18. Mueller KF, Meerpohl JJ, Briel M, *et al*. Methods for detecting, quantifying, and adjusting for dissemination bias in meta-analysis are described. *J Clin Epidemiol* 2016;80:25–33.
19. Jin ZC, Zhou XH, He J. Statistical methods for dealing with publication bias in meta-analysis. *Stat Med* 2015;34:343–60.
20. Sterne JA, Sutton AJ, Ioannidis JP, *et al*. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ* 2011;343:d4002.
21. Higgins JP, Altman DG, Gøtzsche PC, *et al*. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011;343:d5928.
22. Sterne JAC, Egger M, Moher D. Chapter 10: Addressing reporting biases. In: Higgins JPT, Green S, eds. *Cochrane handbook for systematic reviews of interventions Version 510*. Chichester, UK: John Wiley & Sons, 2011.
23. Whiting P, Savović J, Higgins JP, *et al*. ROBIS: A new tool to assess risk of bias in systematic reviews was developed. *J Clin Epidemiol* 2016;69:225–34.
24. Shea BJ, Grimshaw JM, Wells GA, *et al*. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Med Res Methodol* 2007;7:10.
25. Haddaway NR, Collins AM, Coughlin D, *et al*. The Role of Google Scholar in Evidence Reviews and Its Applicability to Grey Literature Searching. *PLoS One* 2015;10:e0138237.
26. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;20:37–46.
27. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.
28. Balshem H, Stevens A, Ansari M, *et al*. *Finding grey literature evidence and assessing for outcome and analysis reporting biases when comparing medical interventions: AHRQ and the Effective Health Care Program. (Prepared by the Oregon Health and Science University and the University of Ottawa Evidence-based Practice Centers under Contract Nos. 290-2007-10057-I and 290-2007-10059-I.) AHRQ Publication No. 13(14)-EHC096-EF*. Rockville, MD: Agency for Healthcare Research and Quality, 2013.
29. Berkman ND, Lohr KN, Ansari M, *et al*. *Chapter 15 Appendix A: A Tool for Evaluating the Risk of Reporting Bias (in Chapter 15: Grading the Strength of a Body of Evidence When Assessing Health Care Interventions for the Effective Health Care Program of the Agency for Healthcare Research and Quality: An Update). Methods Guide for Comparative Effectiveness Reviews (Prepared by the RTI-UNC Evidence-based Practice Center under Contract No. 290-2007-10056-I). AHRQ Publication No. 13(14)-EHC130-EF*. Rockville, MD: Agency for Healthcare Research and Quality, 2013.
30. Downes MJ, Brennan ML, Williams HC, *et al*. Development of a critical appraisal tool to assess the quality of cross-sectional studies (AXIS). *BMJ Open* 2016;6:e011458.
31. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J Epidemiol Community Health* 1998;52:377–84.
32. Dwan K, Gamble C, Kolamunnage-Dona R, *et al*. Assessing the potential for outcome reporting bias in a review: a tutorial. *Trials* 2010;11:52.
33. Guyatt GH, Oxman AD, Vist GE, *et al*. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;336:924–6.
34. Guyatt GH, Oxman AD, Vist G, *et al*. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). *J Clin Epidemiol* 2011;64:407–15.
35. Guyatt GH, Oxman AD, Montori V, *et al*. GRADE guidelines: 5. Rating the quality of evidence--publication bias. *J Clin Epidemiol* 2011;64:1277–82.
36. Schünemann H, Brożek J, Guyatt G, *et al*. Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach. http://gdt.guidelinedevelopment.org/app/handbook/handbook.html.
37. Santesso N, Carrasco-Labra A, Langendam M, *et al*. Improving GRADE evidence tables part 3: detailed guidance for explanatory footnotes supports creating and understanding GRADE certainty in the evidence judgments. *J Clin Epidemiol* 2016;74.
38. Hayden JA, van der Windt DA, Cartwright JL, *et al*. Assessing bias in studies of prognostic factors. *Ann Intern Med* 2013;158:280–6.
39. Higgins JPT, Altman DG, Sterne JAC. Chapter 8: Assessing risk of bias in included studies. In: Higgins JPT, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester (UK: John Wiley & Sons, 2008:187–.
40. Higgins JPT, Altman DG, Sterne JAC. Chapter 8: Assessing risk of bias in included studies. In: Higgins JPT, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0*. London: The Cochrane Collaboration, 2011.
41. Higgins JPT, Savović J, Page MJ, *et al*. Revised Cochrane risk of bias tool for randomized trials (RoB 2.0), Version 20. 2016 https://sites.google.com/site/riskofbiastool/ (accessed 19 Sep 2017).
42. Higgins JPT, Sterne JAC, Savović J, *et al*. A revised tool for assessing risk of bias in randomized trials. *Cochrane Methods Cochrane Database of Systematic Reviews* 2016;10(Suppl 1):29–31.
43. Hooijmans CR, Rovers MM, de Vries RB, *et al*. SYRCLE's risk of bias tool for animal studies. *BMC Med Res Methodol* 2014;14:43.
44. Kim SY, Park JE, Lee YJ, *et al*. Testing a tool for assessing the risk of bias for nonrandomized studies showed moderate reliability and promising validity. *J Clin Epidemiol* 2013;66:408–14.
45. Meader N, King K, Llewellyn A, *et al*. A checklist designed to aid consistency and reproducibility of GRADE assessments: development and pilot validation. *Syst Rev* 2014;3:82.
46. Stewart GB, Higgins JP, Schünemann H, *et al*. The use of Bayesian networks to assess the quality of evidence from research synthesis: 1. *PLoS One* 2015;10:e0114497.
47. Reid EK, Tejani AM, Huan LN, *et al*. Managing the incidence of selective reporting bias: a survey of Cochrane review groups. *Syst Rev* 2015;4:85.
48. Saini P, Loke YK, Gamble C, *et al*. Selective reporting bias of harm outcomes within studies: findings from a cohort of systematic reviews. *BMJ* 2014;349:g6501.
49. Salanti G, Del Giovane C, Chaimani A, *et al*. Evaluating the quality of evidence from a network meta-analysis. *PLoS One* 2014;9:e99682.
50. Higgins JP, Del Giovane C, Chaimani A, *et al*. Evaluating the Quality of Evidence from a Network Meta-Analysis. *Value Health* 2014;17:A324.
51. Viswanathan M, Berkman ND. Development of the RTI item bank on risk of bias and precision of observational studies. *J Clin Epidemiol* 2012;65:163–78.
52. Viswanathan M, Berkman ND, Dryden DM, *et al*. *AHRQ Methods for Effective Health Care. Assessing Risk of Bias and Confounding in Observational Studies of Interventions or Exposures: Further Development of the RTI Item Bank*. Rockville (MD: Agency for Healthcare Research and Quality (US), 2013.
53. Armijo-Olivo S, Stiles CR, Hagen NA, *et al*. Assessment of study quality for systematic reviews: a comparison of the Cochrane Collaboration Risk of Bias Tool and the Effective Public Health Practice Project Quality Assessment Tool: methodological research. *J Eval Clin Pract* 2012;18:12–18.
54. Armijo-Olivo S, Ospina M, da Costa BR, *et al*. Poor reliability between Cochrane reviewers and blinded external reviewers when applying the Cochrane risk of bias tool in physical therapy trials. *PLoS One* 2014;9:e96920.
55. Bilandzic A, Fitzpatrick T, Rosella L, *et al*. Risk of Bias in Systematic Reviews of Non-Randomized Studies of Adverse Cardiovascular Effects of Thiazolidinediones and Cyclooxygenase-2 Inhibitors: Application of a New Cochrane Risk of Bias Tool. *PLoS Med* 2016;13:e1001987.
56. Hartling L, Ospina M, Liang Y, *et al*. Risk of bias versus quality assessment of randomised controlled trials: cross sectional study. *BMJ* 2009;339:b4012.
57. Hartling L, Bond K, Vandermeer B, *et al*. Applying the risk of bias tool in a systematic review of combination long-acting beta-agonists and inhaled corticosteroids for persistent asthma. *PLoS One* 2011;6:e17242.
58. Hartling L, Hamm M, Milne A, *et al*. *AHRQ Methods for Effective Health Care. Validity and Inter-Rater Reliability Testing of Quality Assessment Instruments*. Rockville (MD: Agency for Healthcare Research and Quality (US), 2012.
59. Hartling L, Hamm MP, Milne A, *et al*. Testing the risk of bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. *J Clin Epidemiol* 2013;66:973–81.
60. Jordan VM, Lensen SF, Farquhar CM. There were large discrepancies in risk of bias tool judgments when a randomized controlled trial appeared in more than one systematic review. *J Clin Epidemiol* 2017;81:72–6.
61. Kumar A, Miladinovic B, Guyatt GH, *et al*. GRADE guidelines system is reproducible when instructions are clearly operationalized even

among the guidelines panel members with limited experience with GRADE. *J Clin Epidemiol* 2016;75:115–8.
62. Llewellyn A, Whittington C, Stewart G, *et al*. The Use of Bayesian Networks to Assess the Quality of Evidence from Research Synthesis: 2. Inter-Rater Reliability and Comparison with Standard GRADE Assessment. *PLoS One* 2015;10:e0123511.
63. Mustafa RA, Santesso N, Brozek J, *et al*. The GRADE approach is reproducible in assessing the quality of evidence of quantitative evidence syntheses. *J Clin Epidemiol* 2013;66:736–42.
64. Norris SL, Holmer HK, Ogden LA, *et al*. *AHRQ Methods for Effective Health Care. Selective Outcome Reporting as a Source of Bias in Reviews of Comparative Effectiveness*. Rockville (MD): Agency for Healthcare Research and Quality (US), 2012.
65. O'Connor SR, Tully MA, Ryan B, *et al*. Failure of a numerical quality assessment scale to identify potential risk of bias in a systematic review: a comparison study. *BMC Res Notes* 2015;8:224.
66. Vale CL, Tierney JF, Burdett S. Can trial quality be reliably assessed from published reports of cancer trials: evaluation of risk of bias assessments in systematic reviews. *BMJ* 2013;346:f1798.
67. Whiting PF, Rutjes AW, Westwood ME, *et al*. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. *J Clin Epidemiol* 2013;66:1093–104.
68. Sterne JAC, Egger M, Moher D, *et al*. Chapter 10: Addressing reporting biases. In: Higgins JPT, Churchill R, Chandler J, eds. *Cochrane Handbook for Systematic Reviews of Interventions version 5.2.0*. Chichester, UK: The Cochrane Collaboration, 2017.
69. Atakpo P, Vassar M. Publication bias in dermatology systematic reviews and meta-analyses. *J Dermatol Sci* 2016;82:69–74.
70. Lundh A, Gøtzsche PC. Recommendations by Cochrane Review Groups for assessment of the risk of bias in studies. *BMC Med Res Methodol* 2008;8:22.