Journal of Clinical Epidemiology 106 (2019) 121−135

REVIEW

Few studies exist examining methods for selecting studies, abstracting data, and appraising quality in a systematic review

Reid C. Robson[a], Ba' Pham[a], Jeremiah Hwee[b], Sonia M. Thomas[a], Patricia Rios[a], Matthew J. Page[c], Andrea C. Tricco[a,b,*]

[a] *Li Ka Shing Knowledge Institute of St Michael's Hospital, 209 Victoria Street, East Building, Room 716, Toronto, Ontario M5B 1W8, Canada*
[b] *Epidemiology Division, Dalla Lana School of Public Health, University of Toronto, 155 College Street, 6th floor, Toronto, Ontario M5T 3M7, Canada*
[c] *School of Public Health and Preventive Medicine, Monash University, 553 St Kilda Road, Melbourne, Victoria 3004, Australia*

Accepted 1 October 2018; Published online 9 October 2018

Abstract

Objectives: The aim of the article was to identify and summarize studies assessing methodologies for study selection, data abstraction, or quality appraisal in systematic reviews.

Study Design and Setting: A systematic review was conducted, searching MEDLINE, EMBASE, and the Cochrane Library from inception to September 1, 2016. Quality appraisal of included studies was undertaken using a modified Quality Assessment of Diagnostic Accuracy Studies 2, and key results on accuracy, reliability, efficiency of a methodology, or impact on results and conclusions were extracted.

Results: After screening 5,600 titles and abstracts and 245 full-text articles, 37 studies were included. For screening, studies supported the involvement of two independent experienced reviewers and the use of Google Translate when screening non-English articles. For data abstraction, studies supported involvement of experienced reviewers (especially for continuous outcomes) and two independent reviewers, use of dual monitors, graphical data extraction software, and contacting authors. For quality appraisal, studies supported intensive training, piloting quality assessment tools, providing decision rules for poorly reported studies, contacting authors, and using structured tools if different study designs are included.

Conclusion: Few studies exist documenting common systematic review practices. Included studies support several systematic review practices. These results provide an updated evidence-base for current knowledge synthesis guidelines and methods requiring further research.

Keywords: Knowledge synthesis; Systematic reviews; Methodology; Study selection; Data abstraction; Quality appraisal

1. Introduction

Systematic reviews (SRs)—the gathering of all evidence relevant to a research question in a transparent and unbiased way—are considered the gold standard for synthesizing health care evidence because of their methodological rigor [1]. Guidance for their conduct and reporting are readily available and produced by several well-known organizations [2−8]. The conduct of an SR comprises six main steps: defining a clear research question and literature search strategy, selecting relevant studies, assessing their methodological quality or risk of bias (RoB), abstracting relevant data, synthesizing results, and reporting findings [6].

Much research has been conducted on optimal literature search strategies, developing tools for assessing RoB, and establishing components to assess the quality of reporting [9−15]. However, there is much less information to support current standards on how to select studies for inclusion in

Competing interest: A.C.T. is an associate editor of the journal but was not involved with the publication process.

Ethics approval and consent to participate: not applicable.

Consent for publication: not applicable.

Informed consent and patient details: not applicable.

Submission declaration and verification: This article has not been published previously.

Availability of data and materials: The datasets used during this study are available from the corresponding author on reasonable request.

Funding: This review was funded by an Ontario Ministry of Research, Innovation, and Science Early Researcher Award (2015 to 2020) awarded to A.C.T. A.C.T. is also funded by a Tier 2 Canada Research Chair in Knowledge Synthesis (2016 to 2021). J.H. is supported by the Canadian Institutes of Health Research Doctoral Award. M.J.P. is supported by an Australian National Health and Medical Research Council (NHMRC) Early Career Fellowship (1088535).

* Corresponding author. Tel.: (416) 864-6060x77521; fax: (416) 864-6057.
E-mail address: triccoa@smh.ca (A.C. Tricco).

https://doi.org/10.1016/j.jclinepi.2018.10.003
0895-4356/© 2018 Elsevier Inc. All rights reserved.

an SR, abstract their data, and appraise their quality (or RoB) [16]. This information should also be valuable to those conducting rapid SRs because rapid reviews necessarily must streamline the SR process while attempting to maintain the integrity of an SR [17].

As the knowledge synthesis community advocates for evidence-based practice, it is imperative that our knowledge synthesis methods are informed by research evidence. We thus aimed to conduct an SR to determine the accuracy, reliability, impact, and efficiency of different methods for study selection, data abstraction, and quality appraisal in SRs.

What is new?

Key findings

- For screening of titles and abstracts, we recommend using reviewers with content expertise and two independent reviewers. If resources preclude screening by two independent reviewers, we suggest having one person screen, and another verify the list of excluded studies. Google Translate to English may be considered for some languages (e.g., German).
- For data abstraction, we recommend using data abstractors with experience conducting reviews (especially for continuous outcomes), using two computer screens, contacting authors for additional data, using computer-assisted programs for graphical data, and using two independent reviewers. If resources preclude data abstraction by two independent reviewers, we recommend that one person abstracts, and another verifies the outcome data when a statistical analysis is being considered (e.g., meta-analysis and network meta-analysis).
- For quality appraisal, we recommend intensive training (including piloting and calibration), providing decision rules for studies reporting insufficient detail, contacting authors for details regarding quality assessment, and using a structured quality assessment tool if different types of study designs are included. For qualitative reviews, a structured or unstructured tool can be considered.

What this adds to what was known?

- This review provides updated evidence for methods commonly cited in systematic review guidelines, for study selection, data abstraction, and quality appraisal, confirming several current practices, encouraging others, while cautioning against a few.

What is the implication and what should change now?

- It may be prudent to update guidelines for systematic and rapid reviews, particularly with respect to our findings regarding resources and efficiency.

2. Materials and methods

2.1. Study protocol

We registered the protocol for our SR with PROSPERO (CRD42016047877) [18].

2.2. Eligibility criteria

Studies examining methodological approaches for the selection of studies according to defined eligibility criteria, abstraction of their data, or their quality appraisal were included [19]. Specifically, studies were included if they compared or evaluated the accuracy or reliability of a method or described factors that affect the method's accuracy or reliability.

We defined accuracy studies as those that compared the examined method to a "gold standard" method (as may be depicted in the methodologically rigorous Cochrane Handbook for Systematic Reviews [2]). For example, a study may examine whether a single reviewer was able to identify all the studies eligible for inclusion according to those included using the gold standard two independent reviewers (this aspect of a method's accuracy is referred to as sensitivity, or recall; specificity would refer to how many studies were correctly excluded).

We defined reliability studies as those examining concordance between reviewers, for example, how similar were the study selection decisions made by two reviewers using a "review titles only" methodology. Studies evaluating the impact of a method on the results, conclusions, or efficiency (in terms of timeline or reviewer-person-time) of an SR were also eligible. Primary research studies providing quantitative data were eligible, and studies could evaluate one approach or compare several. As we were interested in accuracy, reliability, and impact on efficiency, qualitative studies were excluded.

Studies examining adherence to SR guidelines (e.g., Preferred Reporting Items for Systematic Reviews and Meta-Analyses), assessing quality (e.g., Assessing the Methodological Quality of Systematic Reviews), evaluating search strategies, or assessing whether appropriate study selection criteria were defined (as opposed to followed) were excluded. Also excluded were studies evaluating the concordance between two different SRs addressing the same question, unless the study specifically examined differences in results because of the methods used for screening (e.g., one vs. two reviewers), data abstraction, or RoB appraisal. Studies developing tools for RoB appraisal in SRs were also excluded. Finally, although we acknowledge the importance of text mining and automation in the advancement of SR methodology [20,21], this has been addressed elsewhere [22], and we did not include studies evaluating these methods.

2.3. Search strategy and information sources

An experienced librarian compiled comprehensive literature searches of MEDLINE, EMBASE, and the Cochrane Library from inception to September 1, 2016. The main (MEDLINE) literature search was peer-reviewed by another experienced librarian using the Peer Review of Electronic Search Strategies Statement [9]. Details of the search can be found in Appendix A. We scanned the reference lists of included studies, relevant SRs, and guidance on conducting SRs [2–5] for potentially relevant articles. Experts and authors of seminal articles were also contacted, and authors of conference abstracts were contacted to obtain complete reports. We consulted our personal files to ensure studies were not missed.

2.4. Study selection and data collection

Using our predefined eligibility criteria, a standardized form for screening studies for inclusion was developed and pilot-tested on a random sample of 25 titles and abstracts by five team members (R.R., B.P., J.H., S.M.T., and A.C.T.). This calibration was repeated for the screening of full-text articles. Pilot-testing was repeated until 80% agreement between reviewers was achieved, with one pilot test required for title and abstract screening, and two pilot tests for full-text article screening. Pairs of reviewers independently screened citations and full-text articles for inclusion. Discrepancies between reviewers were resolved by discussion or a third reviewer. All levels of screening were conducted using Synthesi.SR, proprietary online software of the Knowledge Translation Program of St. Michael's Hospital [23].

A data abstraction form and guidance document were developed and piloted using the same process on a random sample of three studies (one study for each method—selection, abstraction, and quality appraisal). Because of resource limitations, data abstraction was conducted with one abstractor and one verifier, and discrepancies were resolved by discussion or involvement of a third reviewer. Study characteristics, results, key findings, and conclusions were abstracted.

2.5. RoB assessment

A modified version of the Quality Assessment of Diagnostic Accuracy Studies (QUADAS)-2 checklist was used for the assessment of RoB of included studies. The QUADAS-2 consists of four domains (participant selection, the index test, reference standard, and flow and timing) that are assessed using "signaling questions." We modified the signaling questions pertaining to participant selection to recognize that both reviewers and records were selected in our studies (two domains) and removed the flow and timing domain. The checklist was pilot-tested using the same sample of three randomly selected articles used in the data abstraction pilot and was completed by one reviewer and verified by another. Discrepancies were resolved through discussion or involvement of a third reviewer.

2.6. Synthesis

Characteristics of the selection, abstraction, and appraisal methodologies from the included studies were examined. The classification of the methodologies for presenting key findings was agreed by the team. Results for studies supporting individual methodologies were tabulated and summarized. No meta-analyses were performed.

3. Results

3.1. Literature search

After screening 5,602 titles and abstracts, and 245 potentially relevant full-text articles, 37 studies (Fig. 1) describing 12 methods (Table 1) for the selection (11 studies), abstraction (13 studies), or appraisal (15 studies) of studies were eligible for inclusion. A list of key excluded studies can be found in Appendix B.

3.2. Study characteristics

Table 1 summarizes the characteristics of the 37 included studies. A high proportion of studies were published between 2010 and 2014 (45.9%). The most common study designs were non-randomized controlled trials (non-RCTs; 27.0%), RCTs (21.6%), and cross-sectional studies (21.6%). Studies were conducted most commonly in the United States (27.0%), Canada (24.3%), and the United Kingdom (16.2%). The most common methods examined were the use of inexperienced reviewers for both data abstraction ($n = 5$) and appraisal ($n = 4$), the use of two independent reviewers for selection ($n = 4$), and blinding of reviewers for appraisal ($n = 4$; Table 1). The overall RoB is summarized in Fig. 2.

3.3. Methods for conducting the selection of studies in an SR

Eleven studies provided evidence on six study selection methodologies. Details can be found in Table 2 and Appendices C and D. Only one study [32] had low RoB across the four modified QUADAS-2 domains. Four studies [26,30,31,34] had high RoB across some domains, and the remainder had one or more domains with an unclear RoB (Appendix E).

3.3.1. Two independent reviewers for study selection (four studies)

In an SR of postal questionnaires, where the search produced over 22,000 records, Edwards et al. [24] compared one vs. two independent reviewers for accuracy and

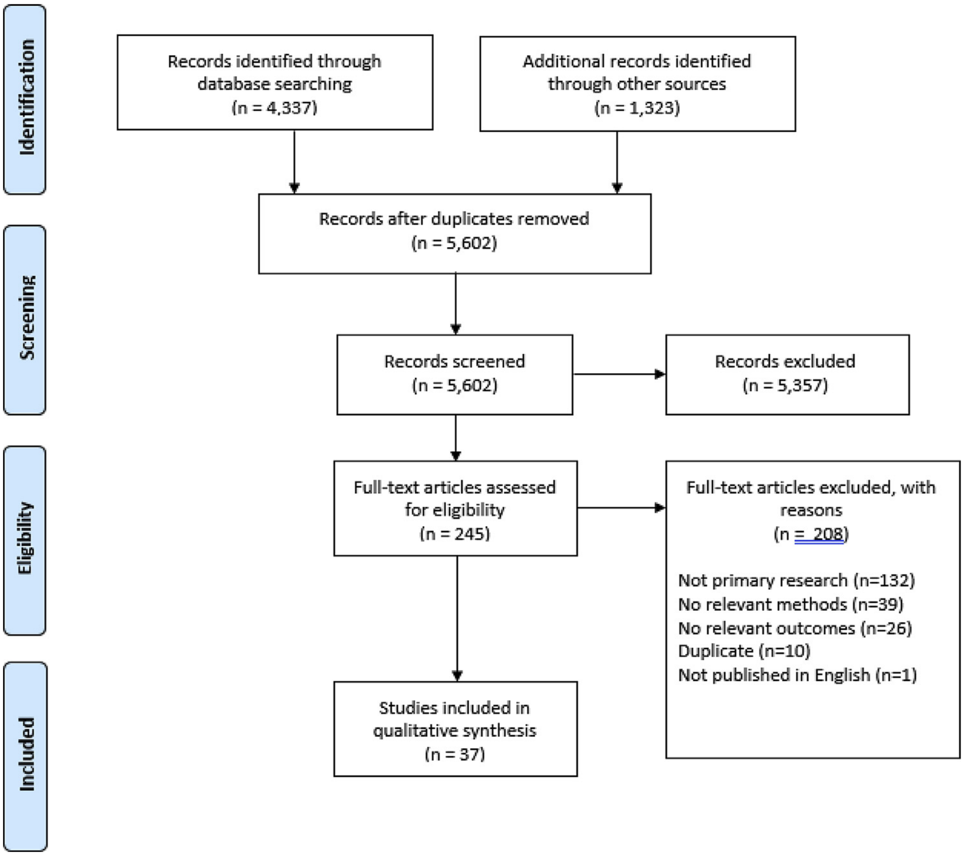
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 4,337)
Additional records identified through other sources (n = 1,323)
Records after duplicates removed (n = 5,602)
Records screened (n = 5,602)
Records excluded (n = 5,357)
Full-text articles assessed for eligibility (n = 245)
Full-text articles excluded, with reasons (n = 208)
Not primary research (n=132)
No relevant methods (n=39)
No relevant outcomes (n=26)
Duplicate (n=10)
Not published in English (n=1)
Studies included in qualitative synthesis (n = 37)

Fig. 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow diagram.

inter-rater reliability when screening records for inclusion. The authors found good reliability between raters but estimated that single reviewers missed on average 8% (range 0%–24%) of eligible reports, whereas reviewer pairs did not miss any (range 0%–1%). A second reviewer increased the number of eligible studies identified by 9% (range across pairs 0%–32%). Edwards et al. concluded that two reviewers should screen records for eligibility to ensure that relevant trials are not missed. In two smaller studies of diagnostic accuracy, Doust et al. [25] found similar results and showed that three reviewers were generally unnecessary (Table 2).

In 2013, Yip et al. [26] assessed an SR that used two independent reviewers for study selection. Yip et al. reviewed the selection criteria against the 21 included studies and the available literature, and identified 13 errors in study selection (see Appendix D). The authors argued that the extent of these errors, in addition to errors found in abstraction, raised concerns about the credibility of the conclusions of the original review.

Using process data (e.g., time use, screening decisions) from an SR on medical education, Shemilt [27] constructed a decision-analytic model for a cost-effectiveness analysis comparing the relative efficiency of screening methods, including "double" and "single" screening. Shemilt found that while double screening did not miss any of the eight eligible studies, single screening missed one, but cost 30% less, saving an estimated 22K pounds.

3.3.2. *Use of inexperienced reviewers (two studies)*

In an SR examining hypothermia for brain injury, Ng et al. [28] randomly assigned 58 medical students with no experience in SRs or the content area to each screen 650 citations using different screening modalities. The medical students' performance was compared to two experienced content experts who had conducted a previous review on this topic. Students received minimal training—a one-page summary of the review protocol, along with inclusion criteria, but no further support. Student performance was highly variable, and below that of the experienced reviewers (median sensitivity range: 47%–67%; median specificity range: 93%–97%). No student identified all 14 articles selected by the experts for inclusion. Ng et al. noted that opportunities exist for improvement, and recommended the use of nonexpert groups be further investigated.

In an SR of dietary guidelines, Cooper et al. [29] compared the inter-rater reliability and accuracy of screening by six graduate students vs. six content experts (nutrition professionals), both individually and among pairs. Reviewers were provided a training manual that included the research purpose and question, instructions on inclusion/exclusion criteria, a study design booklet, tip

Table 1. Summary study characteristics of included studies on selection, abstraction, and appraisal methods

Characteristic	Overall (N = 37[a])	Selection studies (N = 11)	Abstraction studies (N = 13)	Appraisal studies (N = 15)
Year of publication				
1995–1999	5[a]	1	1	4
2000–2004	3	1	1	1
2005–2009	10	3	5	2
2010–2014	17[a]	5	6	7
2015–2017	2	1	0	1
Study design				
RCT	8[a]	2	2	5
Non-RCT	10	2	2	6
Cross-sectional	8	1	6	1
Controlled before–after	1[a]	1	1	0
Repeated measures study	3	2	0	1
Case study	6	2	2	2
Simulation study	1	1	0	0
Country				
United States of America	10[a]	4	6	2
Canada	9	2	2	5
United Kingdom	6	2	0	4
Australia	4	2	1	1
Brazil	2	1	1	0
Denmark/Switzerland/The Netherlands	3	0	1	2
France	1	0	0	1
NR	2	0	2	0
Methodology				
Two reviewers	5	4	1	0
Inexperienced reviewers	11	2	5	4
Translation to English	3	2	1	0
Computer-assisted extraction	2	NA	2	NA
Blinding of reviewers	5	1[a]	1[a]	4
Titles-first screening	1	1	NA	NA
Dual computer monitors	1	1[a]	1[a]	0
Contacting authors for additional data	4	0	2	2
Structured assessment	2	0	0	2
Use of additional guidance	3	0	0	3

Abbreviations: RCT, randomized controlled trial; NR, not reported, NA, not available.

[a] Two of the 37 included studies (Berlin 1997; Wang 2014) report on both selection and abstraction methods; therefore, these numbers do not add up to the overall total.

sheets and frequently asked questions, and received a 1-hour group training session with 10 examples. Two of the authors assessed every title (n = 185) and abstract (n = 90) and formed the reference standard for the assessment of accuracy. Accuracy was similar among pairs, where every pair had false negatives for both title and abstract screening, and sensitivity was generally low (sensitivity for abstract screening: 67%–89%; specificity: 76%–91%). Graduate student pairs had greater inter-rater agreement than nutrition professional pairs for title screening ($P < 0.05$), but no differences were observed for abstract screening. Cooper et al. concluded that graduate students and professionals were comparable in their (inter-rater) agreement, while both had important differences with the expert reference standard.

3.3.3. Translation to English (two studies)

A pilot study, presented as a poster at a Cochrane Collaboration meeting [30], used Google Translate on 11 German articles from a Cochrane review, so that English reviewers could evaluate their potential for inclusion. The agreement between the inclusion decisions in the Cochrane review and the decisions with Google Translate was assessed using kappa. Inter-rater agreement was 73% (kappa = 0.38), and the authors concluded that Google Translate could be a potential tool when screening German articles.

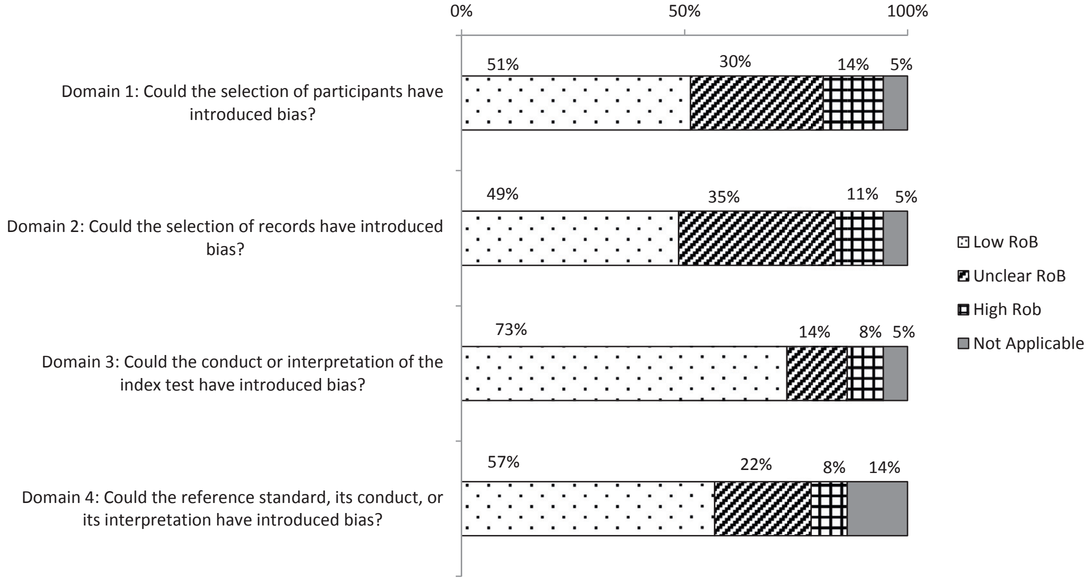

Fig. 2. Overall risk of bias across Quality Assessment of Diagnostic Accuracy Studies (QUADAS)-2 domains.

Busse et al. [31] evaluated a 10-question guide for English reviewers to assess the inclusion of non-English articles. The guide was applied to 133 articles in 19 non-English languages in an SR of fibromyalgia. Reviewers mistakenly judged six trials as ineligible, and eight as eligible (sensitivity = 0.89 and specificity = 0.90), while a strategy developed post-hoc (exclude languages with less than three articles, articles where the title/abstract suggests ineligibility, or articles that lack a clearly reported statistical analysis) had higher sensitivity (0.96). Busse et al. concluded that these strategies showed promise for limiting the need for translators in SRs with many non-English articles.

3.3.4. Dual computer monitors (one study)

In a study conducted at a single evidence-synthesis center, Wang et al. [34] evaluated the efficiency of conducting SRs after implementing dual monitors (two screens linked to a single computer) compared with reviews that did not use dual monitors in the period prior. Sixty studies and 54 reviewers were included. There were no significant differences in time spent on abstract or full-text screening.

3.3.5. Blinding of reviewers (one study)

In a single study, Berlin et al. [32] chose five randomly selected published meta-analyses, and assessed each study in the meta-analyses for inclusion, using two pairs of reviewers, based on the eligibility criteria in the publications. One pair was blinded to identifying information (author identity, institution, journal, and treatment group), whereas the other was not. There was considerable disagreement between blinded and unblinded reviewers and between reviewers and the published meta-analyses regarding which studies were eligible for inclusion. However, the inclusion decisions had little impact on the summary odds ratio for the five meta-analyses studied, whereas the effort to blind was time-consuming (average of 7.7 hours per article).

3.3.6. Titles-first screening (one study)

In an SR examining an association between night-shift work and breast cancer, Mateen et al. [33] compared a titles-first screening approach with the more common titles-and-abstract simultaneously approach. Inter-rater agreement was slightly lower for the titles-first approach, but sensitivity was 100% in both cases (all 13 articles included in the final SR were identified). By immediately rejecting 86% of the citations based on title alone, the authors concluded that titles-first screening may offer a more efficient yet equally effective option, if future studies confirm these findings.

3.4. Methods for conducting the abstraction of studies included in an SR

Thirteen studies provided evidence on seven study selection methodologies. Details can be found in Table 3 and Appendices C and D. Three studies [32,37,41] had low RoB across the four modified QUADAS-2 domains, one study had an instance of high RoB regarding study conduct [34], and the remainder had one or more domains with an unclear RoB (Appendix E).

3.4.1. Double-data abstraction (one study)

Buscemi et al. [35] randomly assigned four reviewers to single-data abstraction (with verification by a second reviewer) or independent double-data abstraction for each of 30 studies in an SR of melatonin for sleep disorders. Discrepancies were resolved by consensus. The authors found

Table 2. Summary of key findings from studies evaluating study selection methods

Author (year)	Study size	Risk of bias[a]	Study objectives	Key conclusions related to study selection methods
Edwards (2002) [24]	One SR; 22,571 citations; four reviewers	U/U/L/U	Estimate accuracy and reliability of reviewers when screening for eligible records	Two reviewers should screen records for eligibility to maximize ascertainment of relevant trials.
Doust (2005) [25]	Two SRs; 638 + 373 citations; two reviewer pairs	U/U/L/U	Assess sensitivity and precision of five search strategies and reliability and accuracy of reviewers screening results	Two reviewers should screen the initial list of citations; lack of information in the abstract makes it difficult to assess eligibility for studies of diagnostic accuracy.
Yip (2014) [26]	One SR; 591 citations; four reviewers	L/U/U/H	Assess the accuracy of results reported in the recent SR on CT screening for lung cancer	13 study selection errors among 21 studies, which raises concern about credibility of original review conclusions; thorough data checking process required.
Shemilt (2016) [27]	One SR; 12,477 citations	NA	Compare costs and effects of four SR approaches for screening eligible studies	Alternatives to the ''double screening'' approach, such as the ''safety first'' approach or integrating text mining, may improve efficiency but require further study.
Ng (2014) [28]	One SR; 650 citations; 58 students	L/U/L/L	Provide preliminary data on accuracy of title/abstract screening by students and effects of screening modality	Medical students are feasible option for screening studies. A Web-based SR platform was more accurate (sensitive, specific) than three other screening modalities.
Cooper (2006) [29]	One SR; 185 titles, 90 abstracts; 12 reviewers	U/U/L/L	Compare literature screening completed by experts vs. nonexperts	Experts and nonexperts had comparable agreement on title + abstract screening, but some differences remained.
Freitas de Souza (2009) [30]	One Cochrane review with 11 articles	H/H/U/H	Evaluate accuracy of Google Translate for selection of trials in non-English languages for inclusion in Cochrane reviews	Google Translator tool could be a potential tool for Cochrane authors to translate German articles.
Busse (2014) [31]	133 articles in 19 languages; eight reviewer pairs	U/U/H/L	Assess whether English-speaking reviewers can identify eligible foreign-language articles for an SR	Strategies to limit the need for non-English language reviewers when screening potentially eligible non-English language articles show potential.
Berlin (1997) [32]	Five MAs; two reviewer pairs	L/L/L/L	Determine the impact of blinding reviewers during study selection and data extraction	Blinding during study selection and data abstraction did not have a clinically or statistically significant effect on results.
Mateen (2013) [33]	One SR; 2,965 citations; five reviewers	L/U/U/L	Compare two methods of screening citations for an	Titles-first screening may be more efficient than

(*Continued*)

Table 2. Continued

Author (year)	Study size	Risk of bias[a]	Study objectives	Key conclusions related to study selection methods
			SR	titles and abstracts together, but further study required.
Wang (2014) [34]	Two screens: 32 SRs, 18 reviewers; One screen: 30 SRs, 36 reviewers	L/L/H/U	Evaluate effectiveness of using two computer screens on the efficiency of conducting SRs	No significant difference in time spent on abstract screening, full-text screening, or inter-rater agreement.

Abbreviations: H, high risk of bias; L, low risk of bias; MA, meta-analyses; SR, systematic review; U, unknown.

[a] Risk of bias was evaluated in four domains; domain 1 = participant selection; domain 2 = record selection; domain 3 = evaluated SAA methods; domain 4 = reference standard. For complete risk of bias assessment results, refer to Appendix E.

that single-data abstraction with verification was 36% faster ($P = 0.03$), yet led to 22% more errors ($P = 0.019$). However, the errors did not have an impact on the overall meta-analysis treatment effect estimates and conclusions. Buscemi et al. suggested that single-data abstraction may be best suited to reviews with multiple outcomes and employing experienced reviewers, because the potential negative effects of single extraction may be diluted if conclusions are based on multiple outcomes.

3.4.2. Use of inexperienced reviewers (five studies)

Horton et al. [36] categorized 87 reviewers' experience with SRs and data abstraction as minimal, moderate, or substantial, based on years of SR experience (<2, 4–6, >7), number of reviews conducted (<2, 4–6, >7), and number of studies abstracted (<50, 51–300, >300). They then compared error rates among these groups in the abstraction of three studies on insomnia treatment. Detailed written data abstraction instructions were provided. There were no significant differences in error rates (range 28%–31%) or study effect estimates between reviewer groups. Minimally experienced reviewers took 22% more time to abstract data, compared with reviewers with substantial experience, but this difference was not statistically significant.

Tendal et al. [37] examined inter-reviewer reliability among five PhD students and five methodologists with substantial experience in meta-analyses. Each reviewer independently extracted continuous data (n, mean, standard deviation) for the calculation of the standardized mean difference (SMD) in 45 trials that were used in 10 meta-analyses. Reviewers were given trial protocols, reports, and a copy of the Cochrane Handbook for Systematic Reviews [46]. The methodologist pairs more often agreed than PhD student pairs at the study trial level (61% vs. 46% of pairs). There was a smaller difference at the meta-analysis level (33% vs. 27%). In 14 of the 100 SMDs calculated at the meta-analysis level (10 meta-analyses × 10 reviewers), individual reviewers reached different conclusions than the published review, with most differences coming from PhD student reviewers. The authors noted that meta-analyses using SMDs are prone to reviewer variation and suggested that reliability might be improved by having more detailed review protocols, more than one reviewer, and statistical expertise.

Gresham et al. [40] compared data abstraction by four experienced with four less experienced reviewers for four trial reports, and found the less experienced reviewers were faster but made 14.2% more errors (95% confidence interval: 7.2%–21.2%). For every 5-minute increase in time abstracting, 2.4% fewer errors occurred on average ($P < 0.001$), regardless of experience level. Two other studies [38,39] involving inexperienced reviewers met our inclusion criteria, but did not provide sufficient data for abstraction.

3.4.3. Translation to English (one study)

In a study involving 15 reviewers and 10 RCTs in five languages, Balk et al. [41] used Google Translate to translate articles before double-data extraction by English reviewers. Two speakers fluent in the relevant language also independently abstracted data from the original articles (differences were resolved by discussion). Results were compared for the two approaches. Although Google Translate performed reasonably well for extraction of some types of data (e.g., study design characteristics), others, specifically outcomes data, were poorly extracted. Accuracy depended on the language translated, with greatest accuracy for Spanish (7% of items had less than 50% correct abstractions), and the least for Chinese (22%). Google Translate required fewer resources (average translation time 30 minutes), and Balk et al concluded it may be appropriate, and time-saving, to run Google Translate and have a native spearker review the translation. Balk et al also felt it was reasonable to conclude Google Translate was still insufficiently accurate for use in data abstraction.

3.4.4. Dual computer monitors (one study)

Although no differences were found in time spent selecting studies in Wang et al.'s study of dual computer monitors (Section 3.3), the authors did find that dual monitors reduced the time to abstract data by 37% (24 minutes, $P = 0.04$) compared with single monitors, with no

Table 3. Summary of key findings from studies evaluating data abstraction methods

Author (year)	Study size	Risk of bias[a]	Study objectives	Key conclusions related to data abstraction methods
Buscemi (2006) [35]	30 studies; four abstractors	U/L/L/U	Compare single- vs. double-data extraction in terms of frequency of errors, treatment effect, and time required	Double-data abstraction is recommended when resources permit. Single-data extraction saves time but generates more errors than double-data abstraction.
Horton (2010) [36]	Three studies; abstractor experience: 28 low, 19 moderate, 23 high	L/U/U/L	Assess impact of experience on accuracy and efficiency of data extraction in SRs	High error rates regardless of reviewer experience; however, experience reduced extraction time. Review-specific strategies to master data extraction should be developed.
Tendal (2009) [37]	10 MAs; five experts; five PhD students	L/L/L/L	Study interobserver variation related to extraction of data for calculation of standardized mean differences used in MAs	Disagreements were common; improving MA reliability may require more detailed review protocols, >1 observer, and statistical expertise.
Jayaram (2014) [38]	10 studies; two abstractors	L/U/L/U	Assess whether training inexperienced reviewers can produce a high-quality Cochrane review	Inexperienced reviewers can complete high-quality SRs quickly in an error-free, readable, and scientific manner.
Florence (2005) [39]	One study; 48 abstractors	U/U/L/L	Examine synthesized findings to establish consistency between reviewer pairs for this emerging synthesis methodology	A systematic process of extracting, categorizing, and synthesizing findings of qualitative studies can lead to reproducible results.
Gresham (2014) [40]	Four studies; eight abstractors	U/U/L/U	Determine data abstraction time using *SRDR* and factors associated with data abstraction errors	Less experienced abstractors were faster, but produced more errors. The faster data are abstracted, the more errors for both experienced and less experienced abstractors.
Balk (2013) [41]	60 studies; 15 abstractors	L/L/L/L	Conduct a rigorous analysis of translations of articles from five languages	While Google Translate may reduce language bias (by including non-English studies), there is a tradeoff between including all articles and risk of error.
Wang (2014) [34]	Two screens: 32 SRs, 18 reviewers; One screen: 30 SRs, 36 reviewers	L/L/H/U	Evaluate effectiveness of using two computer screens on the efficiency of conducting SRs	Significant reduction of time spent on data extraction were observed when using two screens.
Berlin (1997) [32]	Five MAs; two reviewer pairs	L/L/L/L	Determine whether blinding reviewers produces different MA results than not blinding	Blinding is not necessary when conducting MAs of RCTs.
Selph (2014) [42]	66 authors; 77 studies	U/L/L/L	Explore yield of contacting authors of diagnostic accuracy studies and impact on SR findings	Contacting authors was time-consuming. Additional data did not meaningfully impact summary estimates, but did impact the assessment the clinical utility of two tests.
Gibson (2006) [43]	146 authors	U/U/U/U	Examine response levels of	Contacting authors did not help

(*Continued*)

Table 3. Continued

Author (year)	Study size	Risk of bias[a]	Study objectives	Key conclusions related to data abstraction methods
			authors asked to provide missing or incomplete data for an SR	acquire missing data. Improved reporting standards are needed to minimize unsuccessful attempts to contact authors
Cahill (2007) [44]	Two abstractors	L/L/L/U	Test precision and accuracy of graphical data extraction software compared with manual "by hand and eye" methods	Graphical data extraction software offers increased consistency and replicability; however, it requires transparent decision-making and an accurate set-up procedure.
de Oliveira (2003) [45]	Three studies; two abstractors	U/U/L/L	To report on reliable methods for abstracting graphical data	The proposed method is reliable in obtaining numerical data from graphs and makes use of often discarded data for MAs.

Abbreviations: H, high risk of bias; L, low risk of bias; MA, meta-analyses; SR, systematic review; SRDR, systematic review data repository; U, unknown.

[a] Risk of bias was evaluated in four domains; domain 1 = participant selection; domain 2 = record selection; domain 3 = evaluated SAA methods; domain 4 = reference standard. For complete risk of bias assessment results, refer to Appendix E.

significant difference in inter-rater reliability. Wang et al. concluded that dual monitors may improve the efficiency of conducting data abstraction in SRs [34].

3.4.5. Blinding of reviewers (one study)

Berlin et al. [32] compared blinded and unblinded reviewers for both study selection and abstraction, and as noted earlier (Section 3.3), found that blinding did not affect the summary odds ratio for the five meta-analyses examined. Berlin concluded that reviewer blinding was unnecessary.

3.4.6. Contacting study authors for additional data (two studies)

In a review of diagnostic studies, Selph et al. [42] requested information from 66 international authors for 77 studies. Authors were initially contacted by e-mail and then by telephone if there was no response. Sixty-eight percent (68%) were successfully contacted and provided additional data for 29 (38%) studies. Response rates were not affected by country responding, and more extensive attempts (contacting more than three times, using telephone) were of low yield. The additional data obtained from the authors impacted conclusions regarding the utility of two (of 32) diagnostic blood tests, and allowed estimation of the accuracy of a third.

In a study of weight loss, Gibson et al. [43] contacted 146 authors from 19 countries by e-mail and/or letter-mail to obtain additional data. E-mail (47%) was more likely to elicit a response than letter (24%), and letter plus e-mail had the highest response (73%), but this difference was not statistically significant (counts were not provided). The combination of methods was found to be more effective than multiple contacts using the same methods. Response rates did not depend on the author's country nor the number of items requested, but did tend to decline with the age of the article ($P < 0.05$). Average response times differed significantly ($P < 0.05$) and were shortest for e-mail (3 days), followed by e-mail plus letter (13 days), and were longest for letter-only (27 days). Gibson et al. called for improved reporting standards to minimize the relatively unsuccessful attempt to contacting authors, given the considerable resource requirements.

3.4.7. Computer-assisted extraction of graphical data (two studies)

de Oliveira et al. [45] used the cross-hair facility of Adobe Photoshop 7.0 to determine graphical coordinates for graphs in three published reports, and found nearly perfect agreement between the two reviewers using Adobe, as well as with the tables representing the graphs in the original publications. Similarly, using two reviewers, Cahill et al. [44] compared "hand and eye" graphical extraction with two commercially available software packages (DigitizeIT and Grab It!), and found that the software packages had some benefits in terms of accuracy and precision but required care in setting up the process.

3.5. Methods for conducting quality appraisal of studies included in an SR

Fifteen studies provided evidence on five quality appraisal methodologies. Details can be found in Table 4 and Appendices C and D. Six studies [48,49,51,55,60,61] had low RoB across four domains, and seven studies

Table 4. Summary of key findings from studies evaluating appraisal methods

Author (year)	Study size	Risk of bias[a]	Study objectives	Key conclusions related to appraisal methods
da Costa (2017) [47]	Two reviewers; 56 articles	H/L/L/L	Investigate whether training raters on how to assess RoB can improve the reliability of the Cochrane RoB tool	Intensive, standardized training on RoB assessment may improve the reliability of the Cochrane RoB tool.
Sands (1996) [48]	Four trainees; one expert; 97 articles	L/L/L/L	Determine the validity and reliability of training graduate students to read and evaluate methodologic content of articles	Valid, reliable results were obtained by training students to screen articles for methodologic content.
Fourcade (2007) [49]	78 reviewers; 39 articles	L/L/L/L	Develop and evaluate a training tool for a checklist that evaluates nonpharmacologic trials (CLEAR NPT).	There was no difference in effect between treatment groups with or without CLEAR NPT tool training.
Oremus (2012) [50]	10 reviewers; 78 articles	L/H/L/NA	Investigate inter-rater and test—retest reliability for quality assessments conducted by inexperienced student raters	Training followed by a pilot rating phase may improve agreement.
Jadad (1996) [51]	Seven reviewers; 36 articles	L/L/L/L	Determine effect of rater blinding on assessments of quality using the Jadad scale	Blind assessments produced significantly lower and more consistent scores than open assessments.
Berard (2000) [52]	Four reviewers; 20 articles	H/L/L/L	Estimate the inter-rater and test—retest reliability of Chalmers' quality score scale	Reliability varied between subscales and is dependent on articles' blinding status (potential bias when not blinded).
Clark (1999) [53]	Four reviewers; 76 articles	H/L/H/H	Determine reliability of the Jadad scale and effect of blinding on inter-rater agreement	Blinding did not significantly affect the Jadad scale scores.
Verhagen (1998) [54]	20 reviewers; 12 articles	L/H/L/L	Investigate reliability of the Maastricht criteria list for quality assessment in SRs and if blinded reviewing limits review bias	There was no evidence that blinding is necessary to prevent review bias.
Armijo-Olivo (2014) [55]	Six reviewers; 109 articles	L/L/L/L	Test inter-rater reliability of the RoB tool by comparing ratings from Cochrane review authors with blinded external reviewers	Improved guidelines, and revisions to the RoB tool for different health areas, are needed to improve consistency of RoB assessments.
Hartling (2013) [56]	12 reviewers; 124 articles	U/L/L/NA	Assess reliability of the Cochrane RoB tool for RCTs between individual raters and consensus agreements of reviewer pairs	Low observed agreement suggests the need for detailed guidance in RoB assessment.
Robertson (2014) [57]	Five reviewers; 48 articles	L/H/L/NA	Assess inter-rater agreement of the RoB tool for nonrandomized studies and explore the association between RoB and treatment effect size	Only fair agreement observed between reviewers demonstrates the urgent need for further validation to improve inter-rater agreement.
Vale (2013) [58]	Two reviewers; 95 articles	H/L/L/L	Evaluate the reliability of RoB assessments based on published trial reports, for determining trial inclusion in meta-analyses	It may be unreliable to use trial publications alone to assess RoBs; authors should be contacted to inform assessment.

(*Continued*)

Table 4. Continued

Author (year)	Study size	Risk of bias[a]	Study objectives	Key conclusions related to appraisal methods
Littlewood (2012) [59]	Seven SRs	NA	Reflect upon the study appraisal process for SRs in physiotherapy	Study authors should be contacted to inform the quality appraisal process.
Crowe (2011) [60]	Five reviewers; five articles	L/L/L/L	Evaluate whether Crowe critical appraisal tool (CCAT), subject matter/research design knowledge affects appraisal of articles	The CCAT provided better score reliability than informal assessment.
Dixon-Woods (2007) [61]	Six reviewers; 12 articles	L/L/L/L	Compare three methods for appraising qualitative articles that were candidates for inclusion in an SR	Structured approaches may not produce greater consistency of judgments regarding qualitative article inclusion in an SR.

Abbreviations: CCAT, Crowe Critical Appraisal Tool; L, low risk of bias; H, high risk of bias; L, low risk of bias; SR, systematic review; U, unknown.

[a] Risk of bias was evaluated in four domains; domain 1 = participant selection; domain 2 = record selection; domain 3 = evaluated SAA methods; domain 4 = reference standard. For complete risk of bias assessment results, refer to Appendix E.

[47,50,52–54,57,58] had high RoB across one or more of the domains (Appendix E).

3.5.1. *Use of inexperienced reviewers (four studies)*

Standardized intensive training of inexperienced raters was associated with high inter-rater reliability in RoB assessment in two studies [47,48]. In a study by da Costa et al. [47], raters inexperienced in RoB assessment were randomized to two training modalities using the Cochrane RoB tool for RCTs [10]. Raters in the minimal training group received guidance on the tool and attended a one-hour lecture on the definition and importance of each of the domains of bias. Raters in the intensive training group received the minimal training, piloted the tool using a purposively selected sample of 10 articles, and calibrated their assessment to that of an experienced rater. Compared with minimal training, inter-rater reliability substantially improved for all items with intensive training. Reliability between inexperienced raters and two experienced RoB assessors who served as a reference was also higher for the intensive training group, especially for incomplete outcome data and allocation concealment.

In a study by Sands et al. [48], extensive training was provided to graduate students before assessment of 97 studies regarding the health effects of electromagnetic field radiation. Training included a checklist and coding manual, piloting, and weekly meetings with a senior methodologist to resolve discrepancies. Moderate to good inter-rater reliability was obtained for 15 of 23 items included in the quality assessment tool.

Two other studies evaluated comparatively less intensive training modalities. Training involved an online pedagogical tool to enhance the understanding of a checklist that evaluates nonpharmacological trial reports [49] and a 90-min didactic session on the Jadad tool for RCTs and the Newcastle-Ottawa tool for cohort and case–control studies [50]. No differences between trained and not-trained groups were reported.

3.5.2. *Blinding of reviewers (four RCTs)*

One study concluded that blind assessments produced significantly lower and more consistent scores than open assessments [51], but this was not reproduced in three subsequent studies [52–54].

3.5.3. *Use of additional guidance (three studies)*

Two studies examining RoB assessment of RCTs found that provision of guidance and decision rules improved inter-rater reliability, particularly when they were tailored to the specific review topic and when rules were designed to facilitate the assessment of reports not providing adequate information on study design, methods, and conduct [55,56]. In a third study, additional guidance was used in a modified version of the Cochrane RoB tool for nonrandomized comparative studies, but this did not improve the inter-rater reliability of 13 of the 21 items in the modified tool [57].

3.5.4. *Obtaining additional data from study authors (two studies)*

In an examination of 95 trials from 13 IPD (individual patient data) meta-analyses, Vale et al. [58] evaluated the reliability of basing Cochrane RoB assessments on published trial reports alone. They examined the inter-rater agreement between assessments made with and without supplementary information from trial protocols, data collection forms, and summary results. The supplementary information reduced the proportion of unclear ratings for all domains and increased the number of trials assessed as low RoB from 23% to 66%. Five of the 13 meta-analyses had no trials assessed as low RoB based on publications alone, yet when additional information was used, all five had some trials with low RoB. Another study (which did not provide usable data) argued that reasonable attempts to contact study authors should be made to inform the quality appraisal process [59].

3.5.5. Structured vs. unstructured critical appraisal of studies in qualitative reviews (two studies)

Crowe et al. [60] conducted a randomized trial evaluating the effects of the Crowe Critical Appraisal Tool (CCAT) vs. informal appraisal (IA) for assessing health studies. Eight research staff and two postgraduate students were randomly assigned to the IA or CCAT groups to independently appraise five research articles, each with a different study design. The intraclass correlation coefficient for absolute agreement was 0.76 for the IA group and 0.88 for the CCAT group, suggesting that structured assessment (CCAT) provided better score reliability and should help readers with different levels of experience reach similar appraisal results.

Dixon-Woods et al. [61] compared three methods for appraising qualitative research articles: judgment based on expert opinion, a UK Cabinet Office quality framework, and a Critical Appraisal Skills Programme tool. A sample of 12 research articles on support for breastfeeding was appraised by six qualitative reviewers, who selected studies to be included in the SR based on the appraisal results. Although structured approaches did allow reviewers to be more explicit about the reasons for their decisions, the authors concluded that structured approaches may not produce greater consistency in study selection.

4. Discussion

To our knowledge, this is the first SR of methods for the conduct of several essential steps in the SR process. Our study focused on methods relevant to study selection decisions, data abstraction, and quality appraisal, and findings confirm several current practices and provide evidence for some new or alternative practices while discouraging a few. Our results can be used to update guidance on the conduct of SRs [2−4] and rapid reviews. In addition, SR teams can use our results to identify areas where they can revise their SR methodological processes.

For screening, we recommend reviewers with content expertise, Google Translate to screen German citations, and use of two independent reviewers. If resources preclude screening by two independent reviewers, we suggest having one reviewer screen and another verify the list of excluded studies. Although we found that screening titles first was more efficient than screening titles and abstracts simultaneously, and produced similar results, we recommend screening both titles and abstracts simultaneously because of the way literature search results are exported to include titles and abstracts, and there was not a substantial gain in time by screening titles first. We do not recommend the use of two computer screens for screening or blinding of reviewers during screening.

For data abstraction, we recommend data abstractors who are experienced in reviews (especially for continuous outcomes), two computer screens (for efficiency purposes), contacting authors for additional data, computer-assisted programs to extract graphical data, and using two independent abstractors. If resources preclude abstraction by two independent reviewers, we recommend that one person abstracts, and another verifies the outcome data if statistical analysis is planned (e.g., meta-analysis, network meta-analysis). We do not recommend the use of Google Translate for abstracting data from articles in languages other than English, or blinding reviewers for data abstraction.

For quality appraisal, we recommend intensive training, piloting the quality assessment tool, and providing decision rules for studies that report insufficient detail, contacting authors for details regarding quality assessment, and using a structured quality assessment tool if different types of study designs are included. For qualitative reviews, a structured or unstructured tool can be considered. We do not recommend blinding reviewers to appraise study quality.

Mathes et al. [16] recently published a methodological review of data extraction errors and methods to increase data extraction quality. Our study identified the same three methods studies as Mathes et al. [35−37] and several others covering other extraction approaches. Notably, our findings and conclusions are consistent. In particular, Mathes et al. concluded the evidence base for established methods of data abstraction is weak, and we note that few studies exist documenting current practices.

There are several limitations for the included studies. There was limited evidence for most of the methodologies studied. For example, we found only one study that examined the value of double-data abstraction. This limits the strength of our findings, as well as generalizability. It also suggests that further research is required to examine many of the methods recommended for the conduct of SRs [2]. Most studies included a small set of reviewers and a convenience sample of records or reviews that may limit the generalizability of our findings. As well, the quality of the included studies was generally low, which may impact the confidence in our interpretation of results.

There are also several limitations to our SR process that are worth noting. Two reviewers were not used to conduct data abstraction or quality appraisal because of resource limitations. Instead, data were abstracted by one person and verified by a second. We modified the QUADAS-2 instrument for assessment of studies in our review, but we did not validate the modified version. Our categorization of methodologies was achieved through consensus among members of our team; other categorizations could be equally valid. Although we considered gray literature for inclusion in our review, we did not identify any unpublished studies.

5. Conclusion

Few studies exist documenting common SR practices. However, limited evidence was identified supporting several practices. Our review of methodologies for the selection, abstraction, and appraisal of studies for SR provides an updated evidence-base for current guidelines for SRs, considerations for rapid reviews, as well as methods that warrant further research.

Acknowledgments

The authors are grateful for the assistance from Dr. Jessie McGowan for developing our search strategy, Elise Cogo for peer-reviewing the search strategy, Alissa Epworth for running the search and obtaining full-text articles, and Dr. Sharon Straus for reviewing and providing helpful feedback on the article. The authors also thank Krystle Amog and Shazia Siddiqui for formatting the appendices and article.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2018.10.003.

References

[1] Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev 2015;4:1.

[2] Higgins J, Green S. Cochrane handbook for systematic reviews of interventions version 5.1.0. The Cochrane Collaboration; 2011. Available at http://handbook-5-1.cochrane.org/.

[3] Systematic reviews: CRD's guidance for undertaking reviews in health care. University of York, Centre for Reviews & Dissemination; 2009. Available at. https://www.york.ac.uk/media/crd/Systematic_Reviews.pdf.

[4] Owens D, Lohr KN, Atkins D, Treadwell J, Reston J, Bass E, et al. Methods guide for effectiveness and comparative effectiveness reviews. Rockville, MD: Agency for Healthcare Research and Quality; 2014.

[5] Institute of medicine committee on standards for systematic reviews of comparative effectiveness R. In: Eden J, Levit L, Berg A, Morton S, editors. Finding what works in health care: standards for systematic reviews. Washington (DC): National Academies Press (US) Copyright 2011 by the National Academy of Sciences. All rights reserved.; 2011.

[6] Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 2009;4:264−9.

[7] Joanna Briggs Institute Reviewers' Manual: 2015 edition/supplement. Australia: Methodology for JBI Scoping Reviews; 2015.

[8] Methodological expectations of Campbell Collaboration intervention reviews: conduct standards. Available at. https://campbellcollaboration.org/library/campbell-methods-conduct-standards.html.

[9] McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS peer review of electronic search strategies: 2015 guideline statement. J Clin Epidemiol 2016;75:40−6.

[10] Higgins JP, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;343:d5928.

[11] Wells G, Shea B, O'Connell D, Peterson J, Welch V, Losos M, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. The Ottawa Hospital Research Institute: Ottawa, ON.

[12] 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.

[13] Lucas NP, Macaskill P, Irwig L, Bogduk N. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). J Clin Epidemiol 2010;63:854−61.

[14] Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, et al. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. Int J Surg 2012;10:28−55.

[15] Vandenbroucke JP, Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the reporting of observational studies in epidemiology (strobe): explanation and elaboration. Ann Intern Med 2007;147:W163−94.

[16] Mathes T, Klassen P, Pieper D. Frequency of data extraction errors and methods to increase data extraction quality: a methodological review. BMC Med Res Methodol 2017;17:152.

[17] Tricco AC, Antony J, Zarin W, Strifler L, Ghassemi M, Ivory J, et al. A scoping review of rapid review methods. BMC Med 2015;13:224.

[18] Tricco A, Robson R, Thomas S, Pham B, Page M. Accuracy, reliability, impact, and efficiency of different methods for selecting studies, abstracting data, and appraising quality in a systematic review: a systematic review protocol. PROSPERO; 2016. Available at http://www.crd.york.ac.uk/PROSPERO/display_record.php?ID=CRD42016047877.

[19] Moher D, Stewart L, Shekelle P. All in the family: systematic reviews, rapid reviews, scoping reviews, realist reviews, and more. Syst Rev 2015;4:183.

[20] Wallace BC, Dahabreh IJ, Schmid CH, Lau J, Trikalinos TA. Modernizing the systematic review process to inform comparative effectiveness: tools and methods. J Comp Eff Res 2013;2:273−82.

[21] O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Erratum to: using text mining for study identification in systematic reviews: a systematic review of current approaches. Syst Rev 2015;4:59.

[22] Tsafnat G, Dunn A, Glasziou P, Coiera E. The automation of systematic reviews. BMJ 2013;346:f139.

[23] Newton D. Synthesi.SR. Toronto, Ontario: Knowledge Translation Program. Toronto, ON: Li Ka Shing Knowledge Institute, St. Michael's Hospital; 2012.

[24] Edwards P, Clarke M, DiGuiseppi C, Pratap S, Roberts I, Wentz R. Identification of randomized controlled trials in systematic reviews: accuracy and reliability of screening records. Stat Med 2002;21:1635−40.

[25] Doust JA, Pietrzak E, Sanders S, Glasziou PP. Identifying studies for systematic reviews of diagnostic tests was difficult due to the poor sensitivity and precision of methodologic filters and the lack of information in the abstract. J Clin Epidemiol 2005;58:444−9.

[26] Yip R, Islami F, Zhao S, Tao M, Yankelevitz DF, Boffetta P. Errors in systematic reviews: an example of computed tomography screening for lung cancer. Eur J Cancer Prev 2014;23:43−8.

[27] Shemilt I, Khan N, Park S, Thomas J. Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. Syst Rev 2016;5:140.

[28] Ng L, Pitt V, Huckvale K, Clavisi O, Turner T, Gruen R, et al. Title and Abstract Screening and Evaluation in Systematic Reviews (TASER): a pilot randomised controlled trial of title and abstract screening by medical students. Syst Rev 2014;3:121.

[29] Cooper M, Ungar W, Zlotkin S. An assessment of inter-rater agreement of the literature filtering process in the development of evidence-based dietary guidelines. Public Health Nutr 2006;9:494−500.

[30] Freitas de Souza R, Sequeira P, Nasser M. Is Google Translate useful for the selection of studies to be included in Cochrane reviews. 17th Cochrane Colloquium, Singapore 2009;:11−4.

[31] Busse JW, Bruno P, Malik K, Connell G, Torrance D, Ngo T, et al. An efficient strategy allowed English-speaking reviewers to identify foreign-language articles eligible for a systematic review. J Clin Epidemiol 2014;67:547−53.

[32] Berlin JA, On behalf of University of Pennsylvania Meta-analysis Blinding Study Group. Does blinding of readers affect the results of meta-analyses? Lancet 1997;350:185–6.

[33] Mateen FJ, Oh J, Tergas AI, Bhayani NH, Kamdar BB. Titles versus titles and abstracts for initial screening of articles for systematic reviews. J Clin Epidemiol 2013;5:89–95.

[34] Wang Z, Asi N, Elraiyah TA, Abu Dabrh AM, Undavalli C, Glasziou P, et al. Dual computer monitors to increase efficiency of conducting systematic reviews. J Clin Epidemiol 2014;67:1353–7.

[35] Buscemi N, Hartling L, Vandermeer B, Tjosvold L, Klassen TP. Single data extraction generated more errors than double data extraction in systematic reviews. J Clin Epidemiol 2006;59:697–703.

[36] Horton J, Vandermeer B, Hartling L, Tjosvold L, Klassen TP, Buscemi N. Systematic review data extraction: cross-sectional study showed that experience did not increase accuracy. J Clin Epidemiol 2010;63:289–98.

[37] Tendal B, Higgins JP, Juni P, Hrobjartsson A, Trelle S, Nuesch E, et al. Disagreements in meta-analyses using outcomes measured on continuous or rating scales: observer agreement study. BMJ 2009;339:b3128.

[38] Jayaram MK, Mansi K, Haynes E, Adams CE, Furtado VA. Catch 22: is the future of systematic reviewing only for the experienced? [unknown]. Cochrane Schizophrenia Group; 2014.

[39] Florence Z, Schulz T, Pearson A. Inter-reviewer agreement: an analysis of the degree to which agreement occurs when using tools for the appraisal, extraction and meta-synthesis of qualitative research findings. Abstracts of the 13th Cochrane Colloquium. Melbourne, Australia 2005. p. 69.

[40] Gresham G, Matsumura S, Li T. Faster may not be better: data abstraction for systematic reviews. U. S.: Cochrane Eyes and Vision Group. US Cochrane Center; 2014.

[41] Balk EM, Chung M, Chen ML, Trikalinos TA, Kong Win Chang L. Assessing the accuracy of Google Translate to allow data extraction from trials published in non-English languages. Methods research report. Rockville (MD): Agency for Healthcare Research and Quality (US); 2013.

[42] Selph SS, Ginsburg AD, Chou R. Impact of contacting study authors to obtain additional data for systematic reviews: diagnostic accuracy studies for hepatic fibrosis. Syst Rev 2014;3:107.

[43] Gibson CA, Bailey BW, Carper MJ, LeCheminant JD, Kirk EP, Huang G, et al. Author contacts for retrieval of data for a meta-analysis on exercise and diet restriction. Int J Technol Assess Health Care 2006;22:267–70.

[44] Cahill K, Perera R, Selwood M. Electronic extraction of graphical data [abstract]. XV Cochrane Colloquium, Sao Paulo, Brazil 2007, 153-154p 2007 Oct 23-27.

[45] de Oliveira IR, Santos-Jesus R, Po AL, Poolsup N. Extracting numerical data from published reports of pharmacokinetics investigations: method description and validation. Fundam Clin Pharmacol 2003;17:471–2.

[46] Higgins JPTGS, editor. Cochrane Handbook for Systematic Reviews of Interventions 4.2.6. Chichester, UK: John Wiley & Sons, Ltd.; 2006. . Accessed September , 2006.

[47] da Costa BR, Beckett B, Diaz A, Resta NM, Johnston BC, Egger M, et al. Effect of standardized training on the reliability of the Cochrane risk of bias assessment tool: a prospective study. Syst Rev 2017;6:44.

[48] Sands ML, Murphy JR. Use of kappa statistic in determining validity of quality filtering for meta-analysis: a case study of the health effects of electromagnetic radiation. J Clin Epidemiol 1996;49:1045–51.

[49] Fourcade L, Boutron I, Moher D, Ronceray L, Baron G, Ravaud P. Development and evaluation of a pedagogical tool to improve understanding of a quality checklist: a randomised controlled trial. PLoS Clin Trials 2007;2.

[50] Oremus M, Oremus C, Hall GBC, McKinnon MC, Graham A, Gregory C, et al. Inter-rater and test-retest reliability of quality assessments by novice student raters using the Jadad and Newcastle-Ottawa Scales. BMJ Open 2012;2. https://www.ncbi.nlm.nih.gov/pubmed/22855629.

[51] Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 1996;17:1–12.

[52] Berard A, Andreu N, Tetrault J, Niyonsenga T, Myhal D. Reliability of Chalmers' scale to assess quality in meta-analyses on pharmacological treatments for osteoporosis. Ann Epidemiol 2000;10:498–503.

[53] Clark HD, Wells GA, Huet C, McAlister FA, Salmi LR, Fergusson D, et al. Assessing the quality of randomized trials: reliability of the Jadad scale. Control Clin Trials 1999;20:448–52.

[54] Verhagen AP, de Vet HC, de Bie RA, Kessels AG, Boers M, Knipschild PG. Balneotherapy and quality assessment: interobserver reliability of the Maastricht criteria list and the need for blinded quality assessment. J Clin Epidemiol 1998;51:335–41.

[55] Armijo-Olivo S, Ospina M, da Costa BR, Egger M, Saltaji H, Fuentes J, 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.

[56] Hartling L, Hamm MP, Milne A, Vandermeer B, Santaguida PL, Ansari M, 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.

[57] Robertson C, Ramsay C, Gurung T, Mowatt G, Pickard R, Sharma P, et al. Practicalities of using a modified version of the Cochrane Collaboration risk of bias tool for randomised and non-randomised study designs applied in a health technology assessment setting. Res Synth Methods 2014;5:200–11.

[58] 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.

[59] Littlewood C, Ashton J, Chance-Larsen K, May M, Sturrock B. The quality of reporting might not reflect the quality of the study: implications for undertaking and appraising a systematic review. J Man Manip Ther 2012;20:130–4.

[60] Crowe M, Sheppard L, Campbell A. Comparison of the effects of using the Crowe Critical Appraisal Tool versus informal appraisal in assessing health research: a randomised trial. Int J Evid Based Healthc 2011;9:444–9.

[61] Dixon-Woods M, Sutton A, Shaw R, Miller T, Smith J, Young B, et al. Appraising qualitative research for inclusion in systematic reviews: a quantitative and qualitative comparison of three methods. J Health Serv Res Policy 2007;12:42–7.