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

Reid C. Robson, Ba’ Pham, Jeremiah Hwee, Sonia M. Thomas, Patricia Rios, Matthew J. Page, and Andrea C. Tricco

Version  Post-print/Accepted Manuscript


Copyright/License  This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

How to cite TSpace items

Always cite the published version, so the author(s) will receive recognition through services that track citation counts, e.g. Scopus. If you need to cite the page number of the author manuscript from TSpace because you cannot access the published version, then cite the TSpace version in addition to the published version using the permanent URI (handle) found on the record page.

This article was made openly accessible by U of T Faculty. Please tell us how this access benefits you. Your story matters.
Few studies exist examining methods for selecting studies, abstracting data, and appraising quality in a systematic review.

Reid C. Robson\textsuperscript{a}  
Email: RobsonR@smh.ca

Ba’ Pham\textsuperscript{a} 
Email: ba.pham@theta.utoronto.ca

Jeremiah Hwee\textsuperscript{b} 
Email: jeremiah.hwee@mail.utoronto.ca

Sonia M. Thomas\textsuperscript{a} 
Email: ThomasSo@smh.ca

Patricia Rios\textsuperscript{a} 
Email: RiosP@smh.ca

Matthew J. Page\textsuperscript{c} 
Email: matthew.page@monash.edu

Andrea C. Tricco\textsuperscript{a, b*} 
Email: TriccoA@smh.ca

\textsuperscript{a}Li Ka Shing Knowledge Institute of St Michael’s Hospital, 209 Victoria Street, East Building, Room 716, Toronto, Ontario M5B 1 W8, Canada

\textsuperscript{b}Epidemiology Division, Dalla Lana School of Public Health, University of Toronto, 155 College Street, 6th floor, Toronto, Ontario M5T 3M7, Canada

\textsuperscript{c}School of Public Health and Preventive Medicine, Monash University, 553 St Kilda Road, Melbourne, Victoria 3004, Australia

*Corresponding Author:  
Scientist and Lead of the Knowledge Synthesis Team  
Knowledge Translation Program, Li Ka Shing Knowledge Institute of St. Michael’s Hospital  
Associate Professor, Dalla Lana School of Public Health, University of Toronto  
Tel: (416) 864-6060 ext. 77521  
Fax: (416) 864-6057 Email: triccoa@smh.ca
COMPETING INTEREST

Dr. Andrea C. Tricco is an associate editor of the journal but was not involved with the publication process.
ABSTRACT

Objective: 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 Sept 1, 2016. Quality appraisal of included studies was undertaken using a modified QUADAS-2 and key results on accuracy, reliability, or efficiency of a methodology, or impact on results and conclusions were extracted.

Results: After screening 5600 titles and abstracts and 245 full-text articles, 37 studies were included. For screening, studies supported 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

Running Title: Few studies exist examining methods for selecting studies, abstracting data, and appraising quality in a systematic review
**Word Count:** 197 (abstract: 200 max), 4991 (main text: 5000 max), 4 tables, 2 figures, 1 supplementary file.
WHAT IS NEW?

Key findings

• For screening of titles and abstracts, we recommend using reviewers with content expertise, and using 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, 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 is 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. More evaluation is needed with methods that are efficient or time-consuming.
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 healthcare 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 a 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, abstracting relevant data, synthesizing results, and reporting findings [6].

Much research has been conducted on optimal literature search strategies, developing tools for assessing risk of bias (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 a SR, abstract their data, and appraise their quality (or risk of bias) [16]. This information should also be valuable to those conducting rapid SRs, since rapid reviews necessarily must streamline the SR process, while attempting to maintain the integrity of a SR [17].

As the knowledge synthesis community advocates for evidence-based practice, it's imperative that our knowledge synthesis methods are informed by research evidence. We thus aimed to conduct a SR to determine the accuracy, reliability, impact, and efficiency of different methods for study selection, data abstraction, and quality appraisal in SRs.
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, e.g., 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 a 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 [PRISMA]), assessing quality (e.g., Assessing the Methodological Quality of Systematic Reviews [AMSTAR]), 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 risk of bias appraisal. Studies developing tools for risk of bias appraisal in SRs were also excluded. Finally, while 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 PRESS (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 papers. Experts and authors of seminal papers 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 [RR, BP, JH, SMT, ACT]. 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, quality appraisal). Due to 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. Risk of bias 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 upon 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 5601 titles and abstracts, and 244 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.
Table 1: Summary study characteristics of included studies on selection, abstraction, and appraisal methods

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

*2 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.

Abbreviations: RCT = Randomized Controlled Trial, NR = Not Reported
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 USA (27.0%), Canada (24.3%), and the United Kingdom (16.2%). The most common methods 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 a systematic review

Eleven studies provided evidence on six study selection methodologies. Details can be found in Table 2 and Appendices C and D. Only one study [24] had low RoB across the four modified-QUADAS-2 domains. Four studies [25-28] were rated at high RoB across some domains, and the remainder had one or more domains with an unclear RoB (Appendix E).
<table>
<thead>
<tr>
<th>Author (Year)</th>
<th>Study size</th>
<th>Risk of Bias*</th>
<th>Study objectives</th>
<th>Key conclusions related to study selection methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edwards (2002)</td>
<td>1 SR; 22,571 citations; 4 reviewers</td>
<td>U/U/L/U</td>
<td>Estimate accuracy and reliability of reviewers when screening for eligible records</td>
<td>2 reviewers should screen records for eligibility to maximize ascertainment of relevant trials</td>
</tr>
<tr>
<td>Doust (2005)</td>
<td>2 SRs; 638+373 citations; 2 reviewer pairs</td>
<td>U/U/L/U</td>
<td>Assess sensitivity and precision of 5 search strategies and reliability and accuracy of reviewers screening results</td>
<td>2 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</td>
</tr>
<tr>
<td>Yip (2014)</td>
<td>1 SR; 591 citations; 4 reviewers</td>
<td>L/U/U/H</td>
<td>Assess the accuracy of results reported in the recent SR on CT screening for lung cancer</td>
<td>13 study selection errors amongst 21 studies, which raises concern about credibility of original review conclusions; thorough data checking process required</td>
</tr>
<tr>
<td>Shemilt (2016)</td>
<td>1 SR; 12,477 citations</td>
<td>NA</td>
<td>Compare costs and effects of 4 SR approaches for screening eligible studies</td>
<td>Alternatives to the ‘double screening’ approach like the “safety first” approach, or integrating text mining, may improve efficiency but require further study</td>
</tr>
<tr>
<td>Ng (2014)</td>
<td>1 SR; 650 citations; 58 students</td>
<td>L/U/L/L</td>
<td>Provide preliminary data on accuracy of title/abstract screening by students, and effects of screening modality</td>
<td>Use of a web-based SR platform was associated with improved final sensitivity and lower specificity; No significant difference between the other 3 modalities.</td>
</tr>
<tr>
<td>Cooper (2006)</td>
<td>1 SR; 185 titles, 90 abstracts; 12 reviewers</td>
<td>U/U/L/L</td>
<td>Compare literature screening completed by experts vs. non-experts</td>
<td>Experts and non-experts had comparable agreement on title + abstract screening, but some differences remained</td>
</tr>
<tr>
<td>Freitas de Souza (2009)</td>
<td>1 Cochrane review with 11 articles</td>
<td>H/H/U/H</td>
<td>Evaluate accuracy of Google Translate for selection of trials in non-English languages for inclusion in Cochrane reviews</td>
<td>Google Translator tool could be a potential tool for Cochrane authors to translate German articles.</td>
</tr>
<tr>
<td>Busse (2014)</td>
<td>133 articles in 19 languages; 8 reviewer pairs</td>
<td>U/U/H/L</td>
<td>Assess whether English-speaking reviewers can identify eligible foreign-language articles for an SR</td>
<td>Strategies to limit the need for non-English-language reviewers when screening potentially eligible non-English-language articles show potential</td>
</tr>
<tr>
<td>Berlin (1997)</td>
<td>5 MAs; 2 reviewer pairs</td>
<td>L/L/L/L</td>
<td>Determine the impact of blinding reviewers during study selection and data extraction</td>
<td>Blinding did not have a clinically or statistically significant effect on results</td>
</tr>
<tr>
<td>Mateen (2013)</td>
<td>1 SR; 2965 citations; 5 reviewers</td>
<td>L/U/U/L</td>
<td>Compare 2 methods of screening citations for a SR</td>
<td>Titles-first screening may be more efficient than titles and abstracts together, but further study required</td>
</tr>
<tr>
<td>Wang (2014)</td>
<td>2 screens: 32 SRs, 18 reviewers; 1 screen: 30 SRs, 36 reviewers</td>
<td>L/L/H/U</td>
<td>Evaluate effectiveness of using 2 computer screens on the efficiency of conducting SRs</td>
<td>No significant difference in time spent on abstract screening, full-text screening, or inter-rater agreement</td>
</tr>
</tbody>
</table>

*Risk of bias was evaluated in 4 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.

Abbreviations: H = High risk of bias, L = Low risk of bias, MA = Meta-analyses, SR = Systematic Review, U = Unknown.
3.3.1. Two independent reviewers for study selection (4 studies)

In a SR of postal questionnaires, where a search produced over 22,000 records, Edwards et al. [29] compared one vs. two independent reviewers for accuracy and 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, while reviewer pairs did not miss any (range 0 to 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. [30] found similar results and showed that three reviewers were generally unnecessary (Table 2).

In 2013, Yip et al. [25] assessed a 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 a SR on medical education, Shemilt [31] 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 (2 studies)

In a SR examining hypothermia for brain injury, Ng et al. [32] 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 that the use of non-expert groups be 

In a SR of dietary guidelines, Cooper et al. [33] compared the inter-rater reliability and 
accuracy of screening by six graduate students versus six content experts (nutrition 
professionals), both individually and amongst pairs. Reviewers were provided a training manual 
that included the research purpose and question, instructions on inclusion/ exclusion criteria, a 
study design booklet, tip sheets, and frequently asked questions, and received a one-hour group 
training session with ten 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 
amongst 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 (2 studies)

A pilot study, presented as a poster at a Cochrane Collaboration meeting [27], 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.

Busse et al. [26] 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 a SR of fibromyalgia. Reviewers mistakenly judged six trials as ineligible, and eight as eligible (sensitivity= 0.89, 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 these strategies showed promise for limiting the need for translators in SRs with many non-English articles.

3.3.4. Dual computer monitors (1 study)

In a study conducted at a single evidence-synthesis center, Wang et al. [28] evaluated the efficiency of conducting SRs after implementing dual monitors (two screens linked to a single computer) compared to reviews that did not use dual monitors in the period prior. Sixty studies and 54 abstractors were included. There were no significant differences in time spent on abstract or full-text screening.
3.3.5. Blinding of reviewers (1 study)

In a single study, Berlin et al. [24] 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), while 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, while the effort to blind was time-consuming (average of 7.7 hours per paper).

3.3.6. Titles-first screening (1 study)

In an SR examining an association between night-shift work and breast cancer, Mateen et al. [34] compared a titles-first screening approach with the more common titles-and-abstract 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 a systematic review

Thirteen studies provided evidence on seven study selection methodologies. Details can be found in Table 3 and Appendices C and D. Three studies [24, 35, 36] had low RoB across the four modified-QUADAS-2 domains, and one study [28] had an instance of high RoB regarding study conduct [28] (Appendix E).
### Table 3. Summary of key findings from studies evaluating data abstraction methods

<table>
<thead>
<tr>
<th>Author (Year)</th>
<th>Study size</th>
<th>Risk of Bias*</th>
<th>Study objectives</th>
<th>Key conclusions related to data abstraction methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buscemi (2006)</td>
<td>30 studies; 4 abstractors</td>
<td>U/L/L/U</td>
<td>Compare single vs. double data extraction in terms of frequency of errors, treatment effect, and time required</td>
<td>Double data abstraction is recommended when resources permit. Single data extraction saves time but generates more errors than double data abstraction.</td>
</tr>
<tr>
<td>Horton (2010)</td>
<td>3 studies; abstractor experience: 28 low, 19 moderate, 23 high</td>
<td>L/U/U/L</td>
<td>Assess impact of experience on accuracy and efficiency of data extraction in SRs</td>
<td>High error rates regardless of reviewer experience; however, experience-reduced extraction time. Review-specific strategies to master data extraction should be developed.</td>
</tr>
<tr>
<td>Tendal (2009)</td>
<td>10 MAs; 5 experts; 5 PhD students</td>
<td>L/L/L/L</td>
<td>Study inter-observer variation related to extraction of data for calculation of standardized mean differences used in MAs</td>
<td>Disagreements were common; improving MA reliability may require more detailed review protocols, &gt; 1 observer, and statistical expertise.</td>
</tr>
<tr>
<td>Jayaram (2014)</td>
<td>10 studies; 2 abstractors</td>
<td>L/U/L/U</td>
<td>Assess whether training inexperienced reviewers can produce a high-quality Cochrane review</td>
<td>Inexperienced reviewers can complete high-quality SRs quickly in an error free, readable and scientific manner.</td>
</tr>
<tr>
<td>Florence (2005)</td>
<td>1 study; 48 abstractors</td>
<td>U/U/L/L</td>
<td>Examine synthesized findings to establish consistency between reviewer pairs for this emerging synthesis methodology</td>
<td>A systematic process of extracting, categorizing and synthesizing findings of qualitative studies can lead to reproducible results</td>
</tr>
<tr>
<td>Gresham (2014)</td>
<td>4 studies; 8 abstractors</td>
<td>U/U/U/L</td>
<td>Determine data abstraction time using SRDR and factors associated with data abstraction errors</td>
<td>The faster data are abstracted, the greater the proportion of errors for both experienced and less experienced abstractors.</td>
</tr>
<tr>
<td>Balk (2013)</td>
<td>60 studies; 15 abstractors</td>
<td>L/L/L/L</td>
<td>Conduct a rigorous analysis of translations of articles from 5 languages</td>
<td>While Google Translate may reduce language bias, there is a tradeoff between including all articles and risk of error.</td>
</tr>
<tr>
<td>Wang (2014)</td>
<td>30-32 SRs; 18-36 reviewers</td>
<td>L/A/H/U</td>
<td>Evaluate effectiveness of using 2 computer screens on the efficiency of conducting SRs</td>
<td>Significant reduction of time spent on data extraction when using 2 screens.</td>
</tr>
<tr>
<td>Berlin (1997)</td>
<td>5 MAs; 2 reviewer pairs</td>
<td>L/L/L/L</td>
<td>Determine whether blinding reviewers produces different MA results than not blinding</td>
<td>Blinding is not necessary when conducting MAs of RCTs.</td>
</tr>
<tr>
<td>Selph (2014)</td>
<td>66 authors; 77 studies</td>
<td>U/L/L/L</td>
<td>Explore yield of contacting authors of diagnostic accuracy studies and impact on SR findings</td>
<td>Contacting authors was time consuming. Additional data impacted assessing the clinical utility of 2 tests but not summary estimates</td>
</tr>
<tr>
<td>Gibson (2006)</td>
<td>146 authors</td>
<td>U/U/U/U</td>
<td>Examine response levels of authors asked to provide missing or incomplete data for an SR</td>
<td>Contacting authors did not help acquire missing data. Improved reporting standards are needed to minimize unsuccessful attempts to contact authors</td>
</tr>
<tr>
<td>Cahill (2007)</td>
<td>2 abstractors</td>
<td>L/A/L/U</td>
<td>Test precision and accuracy of graphical data extraction software compared to manual ‘by hand and eye’ methods</td>
<td>Graphical data extraction software provides increased consistency and replicability; however, it requires transparent decision-making, and accurate set-up procedure of the reviewer.</td>
</tr>
<tr>
<td>de Oliveira (2003)</td>
<td>3 studies; 2 abstractors</td>
<td>U/U/L/L</td>
<td>To report on reliable methods for abstracting graphical data</td>
<td>Proposed method is reliable in obtaining numerical data from graphs and makes use of discarded data for MAs</td>
</tr>
</tbody>
</table>

*Risk of bias was evaluated in 4 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.

Abbreviations: H = High risk of bias, L = Low risk of bias, MA = meta-analyses, SR = systematic review, SRDR = Systematic Review Data Repository, U = Unknown.
3.4.1. Double-data abstraction (1 study)

Buscemi et al. [37] 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 a SR of melatonin for sleep disorders. Discrepancies were resolved by consensus. The authors found 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, since the potential negative effects of single extraction may be diluted if conclusions are based on multiple outcomes.

3.4.2. Use of inexperienced reviewers (5 studies)

Horton et al. [38] 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 to reviewers with substantial experience, but this difference was not statistically significant.

Tendal et al. [35] examined inter-reviewer reliability amongst five PhD students and five methodologists with substantial experience in meta-analyses. Each reviewer independently extracted continuous data (n, mean, SD) for the calculation of the standardized mean difference (SMD) in 45 trials that were used in ten meta-analyses. Reviewers were given trial protocols,
reports, and a copy of the Cochrane Handbook for Systematic Reviews [39]. 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 x 10 reviewers), individual reviewers reached different conclusions than the published review, 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% CI: 7.2 to 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 [41, 42] involving inexperienced reviewers met our inclusion criteria, but did not provide sufficient data for abstraction.

3.4.3. Translation to English (1 study)

In a study involving 15 reviewers, and ten RCTs in five languages, Balk et al. [36] used Google Translate to translate articles prior to 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. While 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), but Balk et al. concluded it was still insufficiently accurate for use in data abstraction.

### 3.4.4. Dual computer monitors (1 study)

Though 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 to single monitors, with no significant difference in inter-rater reliability. Wang et al. concluded that dual monitors may improve the efficiency of conducting data abstraction in SRs [28].

### 3.4.5. Blinding of reviewers (1 study)

Berlin et al. [24] 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 (2 studies)

In a review of diagnostic studies, Selph et al. [43] requested information from 66 international authors for 77 studies. Authors were initially contacted by email, 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. [44] contacted 146 authors from 19 countries by e-mail and/or letter-mail to obtain additional data. Email (47%) was more likely to elicit a
response than letter (24%), and letter plus email 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 (2 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. [46] 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 a systematic review

Fifteen studies provided evidence on five quality appraisal methodologies, which are summarized below. Details can be found in Table 4 and Appendices C and D. Six studies [47-52], had low ROB across four domains, seven studies [53-59] had high ROB across one or more of the domains, and two could not be fully assessed [60, 61] (Appendix E).
Table 4. Summary of key findings from studies evaluating appraisal methods

<table>
<thead>
<tr>
<th>Author (Year)</th>
<th>Study size</th>
<th>Study objectives</th>
<th>Key conclusions related to appraisal methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>da Costa (2017)</td>
<td>2 reviewers; 56 articles</td>
<td>H/L/L/L Investigate whether training raters on how to assess RoB can improve the reliability of the Cochrane RoB tool</td>
<td>Intensive, standardized training on RoB assessment may improve the reliability of the Cochrane RoB tool.</td>
</tr>
<tr>
<td>Sands (1996)</td>
<td>4 trainees; 1 expert; 97 articles</td>
<td>L/L/L/L Determine the validity and reliability of training graduate students to read and evaluate methodologic content of papers</td>
<td>Valid, reliable results were obtained by training students to screen papers for methodologic content</td>
</tr>
<tr>
<td>Fourcade (2007)</td>
<td>78 reviewers; 39 articles</td>
<td>L/L/L/L Develop &amp; evaluate a tool for a checklist that evaluates non-phenomenological trials (CLEAR NPT)</td>
<td>No difference in effect between treatment groups with or without CLEAR NPT tool training</td>
</tr>
<tr>
<td>Oremus (2012)</td>
<td>10 reviewers; 78 articles</td>
<td>L/H/L/U Investigate inter-rater and test-retest reliability for quality assessments conducted by inexperienced student raters</td>
<td>Training followed by a pilot rating phase may improve agreement</td>
</tr>
<tr>
<td>Jadad (1996)</td>
<td>7 reviewers; 36 articles</td>
<td>L/L/L/L Determine effect of rater blinding on assessments of quality using the Jadad scale</td>
<td>Blind assessments produced significantly lower and more consistent scores than open assessments</td>
</tr>
<tr>
<td>Berard (2000)</td>
<td>4 reviewers; 20 articles</td>
<td>H/L/L/L Estimate the inter-rater and test-retest reliability of Chalmers' quality score scale</td>
<td>Reliability varied between sub-scales, and is dependent on blinding status (potential bias when not blinded)</td>
</tr>
<tr>
<td>Clark (1999)</td>
<td>4 reviewers; 76 articles</td>
<td>H/L/H/H Determine reliability of the Jadad scale and effect of blinding on inter-rater agreement</td>
<td>Blinding did not significantly affect the Jadad scale scores</td>
</tr>
<tr>
<td>Verhagen (1998)</td>
<td>20 reviewers; 12 articles</td>
<td>L/H/L/L Investigate reliability of the Maastricht criteria list for quality assessment in SRs, and if blinded reviewing limits review bias</td>
<td>No evidence that blinded reviewing is necessary to prevent bias</td>
</tr>
<tr>
<td>Armijo-Olivo (2014)</td>
<td>6 reviewers; 109 articles</td>
<td>L/L/L/L Test inter-rater reliability of the RoB tool by comparing ratings from Cochrane review authors with blinded external reviewers</td>
<td>Improved guidelines and revisions to the RoB tool for different health areas are needed due to improved consistency of RoB assessments</td>
</tr>
<tr>
<td>Hartling (2013)</td>
<td>12 reviewers; 124 articles</td>
<td>U/L/L/U Assess reliability of the Cochrane RoB tool for RCTs between individual raters and consensus agreements of reviewer pairs</td>
<td>Low agreement which suggests the need for detailed guidance in RoB assessment</td>
</tr>
<tr>
<td>Robertson (2013)</td>
<td>5 reviewers; 48 articles</td>
<td>L/H/U/H Assess inter-rater agreement of the RoB tool for non-randomized studies, and to explore the association between RoB and treatment effect size</td>
<td>Only fair agreement between reviewers, which demonstrates the urgent need for further validation to improve inter-rater agreement</td>
</tr>
<tr>
<td>Vale (2013)</td>
<td>2 reviewers; 95 articles</td>
<td>H/L/L/L Evaluate the reliability of RoB assessments based on published trial reports, for determining trial inclusion in meta-analyses</td>
<td>Unreliable to use trial publications alone to assess RoBs; authors may be contacted to inform assessment</td>
</tr>
<tr>
<td>Littlewood (2012)</td>
<td>7 SRs</td>
<td>-- Reflect upon the study appraisal process for SRs in physiotherapy</td>
<td>Study authors should be contacted in order to inform the quality appraisal process</td>
</tr>
<tr>
<td>Crowe (2011)</td>
<td>5 reviewers; 5 articles</td>
<td>L/L/L/L Evaluate whether critical appraisal tool, subject matter/research design knowledge affects appraisal of papers</td>
<td>CCAT provided better score reliability and consistency than informal assessment</td>
</tr>
<tr>
<td>Dixon-Woods (2007)</td>
<td>6 reviewers; 12 articles</td>
<td>L/L/L/L Compare 3 methods for appraising qualitative papers that were candidates for inclusion in an SR</td>
<td>Structured approaches may not produce greater consistency of judgments regarding qualitative paper inclusion in an SR</td>
</tr>
</tbody>
</table>

*Risk of bias was evaluated in 4 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.

Abbreviations: CCAT = Crowe Critical Appraisal Tool, L = Low risk of bias, H = High risk of bias, SR = systematic review, U = Unknown.
3.5.1. Use of inexperienced reviewers (4 studies)

Standardized intensive training of inexperienced raters was associated with high inter-rater reliability in RoB assessment in two studies [48, 53]. In a study by da Costa et al. [53], raters inexperienced in RoB assessment were randomized to two training modalities using the Cochrane risk of bias 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 ten articles, and calibrated their assessment to that of an experienced rater. Compared to 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 prior to 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 on-line pedagogical tool to enhance the understanding of a checklist that evaluates non-pharmacological trial reports [47], and a 90-minute didactic session on the Jadad tool for RCTs and the Newcastle-Ottawa tool for cohort and case-control studies [54]. No differences between trained and not-trained groups were reported.
3.5.2. Blinding of reviewers (4 RCTs)

One study concluded that blind assessments produced significantly lower and more consistent scores than open assessments [49], but this was not reproduced in three subsequent studies [55-57].

3.5.3. Use of additional guidance (3 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 [50, 60]. In a third study, additional guidance was used in a modified version of the Cochrane RoB tool for non-randomized comparative studies, but this did not improve the inter-rater reliability of 13 of the 21 items in the modified tool [58].

3.5.4. Obtaining additional data from study authors (2 studies)

In an examination of 95 trials from 13 IPD (individual patient data) meta-analyses, Vale et al. [59] evaluated the reliability of basing Cochrane RoB assessments upon 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 [61].
3.5.5. Structured versus unstructured critical appraisal of studies in qualitative reviews (2 studies)

Crowe et al. [52] conducted a randomized trial evaluating the effects of the Crowe Critical Appraisal Tool (CCAT) versus informal appraisal (IA) for assessing health studies. Eight research staff and two post-graduate students were randomly assigned to the IA or CCAT groups to independently appraise five research papers, each with a different study design. The intra-class 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-Wood et al. [51] compared three methods for appraising qualitative research papers: judgment based on expert opinion; a UK Cabinet Office quality framework; and a Critical Appraisal Skills Programme (CASP) tool. A sample of 12 research papers on support for breast-feeding were appraised by six qualitative reviewers, who selected studies to be included in the SR based upon the appraisal results. Though 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, 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 only was more efficient than screening titles and abstracts, and produced similar results, we recommend screening both titles and abstracts because of the way literature search results are exported to include titles and abstracts and there wasn’t a substantial gain in time by screening titles only. 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, 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 are planned (e.g., meta-analysis, network meta-analysis). We do not recommend the use of Google Translate for
abstracting data from papers 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, 38], 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 due to 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 grey literature for inclusion in our review, we did not identify any unpublished studies.
5. CONCLUSION

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

List of abbreviations

AMSTAR = Assessing the Methodological Quality of Systematic Reviews
CASP = Critical Appraisal Skills Programme
CCAT = Crowe Critical Appraisal Tool
EMBASE = Excerpta Medica dataBASE
IA = Informal appraisal
IPD = Individual Patient Data
MEDLINE = Medical Literature Analysis and Retrieval System Online
PRESS = Peer Review of Electronic Search Strategies
PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QUADAS = Quality Assessment of Diagnostic Accuracy Studies
RCT = Randomized Controlled Trial
RoB = Risk of Bias
SR = Systematic Review

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 ACT. ACT is also funded by a Tier 2 Canada Research Chair in Knowledge Synthesis (2016 to 2021). JH is supported by the Canadian Institutes of Health Research Doctoral Award. MJP is supported by an Australian National Health and Medical Research Council (NHMRC) Early Career Fellowship (1088535).

Authors' contributions

RCR, BP, ACT, MJP, and JH participated in the writing of the protocol. RCR, BP, JH, SMT, PR, and ACT contributed to the conduct of the study. The writing of the manuscript was led by RCR, with sections being written by BP, JH, SMT, PR, MJP, and ACT. All authors (RCR, BP, JH, SMT, PR, MCJ, and ACT) critically reviewed the manuscript.

Acknowledgements

We are grateful for the assistance from 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 Sharon Straus for reviewing and providing helpful feedback on the manuscript. We also thank Krystle Amog and Shazia Siddiqui for formatting the appendices and manuscript.
FIGURE LEGEND

Fig. 1. PRISMA Flow diagram

Fig. 2. Overall Risk of Bias across QUADAS-2 domains

ADDITIONAL FILES

File name: Supplementary File

Title of data: Supplementary Appendices

Description of Data: The appendices include all supplemental data and information.

Appendix A. Search Strategy

Appendix B. Key Excluded Studies

Appendix C. Characteristics of included studies: selection, abstraction and appraisal

Appendix D. Summary of findings for included studies: selection, abstraction, and appraisal

Appendix E. Risks of bias (RoB) - QUADAS Assessment and Results
Reference List


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


Figure 1: PRISMA Flow diagram

The flow diagram depicts the flow of studies during the systematic review process, highlighting the number of records identified in the search, and how many were included and excluded (along with reasons for exclusion) during screening.


For more information, visit www.prisma-statement.org.
Figure 2 – Overall Risk of Bias across QUADAS-2 domains

Figure 2. Risk of Bias across QUADAS-2 domains

Domain 1: Could the selection of participants have introduced bias?
- Low RoB: 51%
- Unclear RoB: 30%
- High RoB: 14%
- Not Applicable: 5%

Domain 2: Could the selection of records have introduced bias?
- Low RoB: 49%
- Unclear RoB: 35%
- High RoB: 11%
- Not Applicable: 5%

Domain 3: Could the conduct or interpretation of the index test have introduced bias?
- Low RoB: 73%
- Unclear RoB: 14%
- High RoB: 8%
- Not Applicable: 5%

Domain 4: Could the reference standard, its conduct, or its interpretation have introduced bias?
- Low RoB: 57%
- Unclear RoB: 22%
- High RoB: 8%
- Not Applicable: 14%

Figure 2 illustrates the overall risk of bias assessment of all included studies (n=37), across the 4 domains of the QUality Assessment of studies of Diagnostic Accuracy included in Systematic reviews – 2 (QUADAS-2) tool.
Declaration of Interest

Dr. Andrea C. Tricco is an associate editor of the journal but was not involved with the publication process.