Journal of Clinical Epidemiology 121 (2020) 20−28

ORIGINAL ARTICLE

Single-reviewer abstract screening missed 13 percent of relevant studies: a crowd-based, randomized controlled trial

Gerald Gartlehner[a,b,*], Lisa Affengruber[a,c], Viktoria Titscher[a], Anna Noel-Storr[d], Gordon Dooley[e], Nicolas Ballarini[f], Franz König[f]

[a]Department for Evidence-based Medicine and Evaluation, Cochrane Austria, Danube University Krems, Krems, Austria
[b]RTI-University of North Carolina Evidence-based Practice Center, RTI International, Research Triangle Park, NC, USA
[c]Department of Family Medicine, Care and Public Health Research Institute, Maastricht University, Maastricht, The Netherlands
[d]Radcliffe Department of Medicine, University of Oxford, Oxford, UK
[e]Metaxis Ltd, Curbridge, UK
[f]Section of Medical Statistics, Medical University of Vienna, Vienna, Austria

Accepted 14 January 2020; Published online 21 January 2020

Abstract

Objectives: To determine the accuracy of single-reviewer screening in correctly classifying abstracts as relevant or irrelevant for literature reviews.

Study Design and Setting: We conducted a crowd-based, parallel-group randomized controlled trial. Using the Cochrane Crowd platform, we randomly assigned eligible participants to 100 abstracts each of a pharmacological or a public health topic. After completing a training exercise, participants screened abstracts online based on predefined inclusion and exclusion criteria. We calculated sensitivities and specificities of single- and dual-reviewer screening using two published systematic reviews as reference standards.

Results: Two hundred and eighty participants made 24,942 screening decisions on 2,000 randomly selected abstracts from the reference standard reviews. On average, each abstract was screened 12 times. Overall, single-reviewer abstract screening missed 13% of relevant studies (sensitivity: 86.6%; 95% confidence interval [CI], 80.6%−91.2%). By comparison, dual-reviewer abstract screening missed 3% of relevant studies (sensitivity: 97.5%; 95% CI, 95.1%−98.8%). The corresponding specificities were 79.2% (95% CI, 77.4%−80.9%) and 68.7% (95% CI, 66.4%−71.0%), respectively.

Conclusions: Single-reviewer abstract screening does not appear to fulfill the high methodological standards that decisionmakers expect from systematic reviews. It may be a viable option for rapid reviews, which deliberately lower methodological standards to provide decision makers with accelerated evidence synthesis products.

Keywords: Randomized controlled trial; Literature screening; Systematic reviews; Rapid reviews; Accuracy

1. Background

Systematic reviews use explicit and predefined scientific methods to synthesize large amounts of information that address a specific question of interest [1]. The goal of any systematic review is to detect all relevant (published and unpublished) studies that meet inclusion criteria. Because of their comprehensiveness and methodological rigor, systematic reviews have become an invaluable tool to guide evidence-based healthcare, improve patient information, and support policy decision making. A search in Epistemonikos (www.epistemonikos.org) reveals that in 2018, more than 41,000 systematic reviews were published worldwide.

Funding: This work was supported by funding from Cochrane to the Cochrane Rapid Reviews Methods Group and by internal funds of Cochrane Austria.

Ethics approval and consent to participate: The local institutional review board at Danube University Krems concluded that no ethical approval was necessary for our study. All participants provided electronic consent to be part of the study.

Potential conflicts of interest: Gerald Gartlehner is a co-convener of the Cochrane Rapid Reviews Methods Group. All authors declare that they have no financial conflicts of interest regarding the topic of this study.

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

Trial registration: Open Science Framework: https://osf.io/3jyqt.

* Corresponding author. Danube University Krems, Dr. Karl Dorrekstrasse 30, Krems 3500, Austria. Tel.: +43 2732 893 2911; fax: +43 2732 893 4910.
E-mail address: gartlehner@cochrane.at (G. Gartlehner).

https://doi.org/10.1016/j.jclinepi.2020.01.005

What is new?

Key findings

- Single-reviewer abstract screening missed 13.4% of relevant studies; by comparison, dual-reviewer abstract screening missed 2.5% of relevant studies.
- Single-reviewer screening missed more relevant studies for public health than the pharmacological topic (16.8% vs. 10.5%).
- Regression analyses had not detected the statistically significant impact of native speaker status, domain knowledge, or experience with literature reviews on the correctness of decisions.

What this adds to what was known?

- Previous studies that assessed the same research question involved only two to four reviewers. Estimates substantially varied because they were highly dependent on the individual skill levels of those who screened abstracts.
- With 280 reviewers, our study is the largest attempt to date to quantify the accuracy of single- and dual-reviewer abstract screening. The large pool of reviewers mitigated the impact of individual screening experience and topic expertise on the accuracy of screening decisions.

What is the implication and what should change now?

- Single-reviewer abstract screening does not appear to fulfill the high methodological standards that decision makers expect from systematic reviews. Institutions that use single reviewers to screen abstracts for systematic reviews should reconsider this approach.

Despite the significance of systematic reviews, the labor intensity involved in conducting a systematic review is a considerable drawback. In recent years, rapid reviews have evolved as a new method of evidence synthesis that streamlines some methodological steps of the systematic review process to provide quick answers to time-sensitive questions of decision makers.

A crucial step in any rapid or systematic review is the screening of abstracts to select potentially relevant studies. No consensus exists in method guides regarding whether abstract screening should be conducted in duplicate by two independent investigators (dual-reviewer screening) or whether screening by a sole investigator is sufficient (single-reviewer screening). To reduce the risk of falsely excluding relevant studies, institutions, such as the University of York Center for Reviews and Dissemination [2], the United States (US) National Academy of Medicine [1], or the German Institute for Quality and Efficiency in Health Care [3] advocate dual-reviewer screening. Their rationale is that dual screening minimizes bias, reduces human error, and can help identify ambiguous inclusion criteria [1]. Critical appraisal tools for systematic reviews also list dual-reviewer screening of the literature as a quality criterion [4,5].

Other institutions that produce systematic reviews, such as the U.S. Agency for Healthcare Research and Quality (AHRQ) [6], the Campbell Collaboration [7], Cochrane [8], and the National Institute for Health and Care Excellence (NICE) [9], acknowledge that dual-reviewer screening is the optimal approach, but view single-reviewer screening as an acceptable alternative. For example, the new version of the Cochrane Handbook for Systematic Reviews of Interventions states that "*…it is acceptable that this initial screening of titles and abstracts is undertaken by only one person.*" [10]. In rapid reviews, single-reviewer screening is also a common approach [11].

To date, few methodological evaluations have determined the proportion of relevant studies that will be missed when titles and abstracts are screened by a sole investigator. A recent systematic review detected four studies of the accuracy of single-reviewer compared with dual-reviewer screening [12]. Across these publications, the median proportion of missed studies was 5%. A recent study that was still not included in this systematic review reported similar findings. Single-reviewer screening of abstracts by two investigators missed 4.4% and 5.3% of relevant studies, respectively [13].

A substantial limitation of these evaluations, however, is that results were based on a few reviewers. Across all five studies, only 11 individual reviewers were involved. Consequently, results are highly dependent on the screening skills and the experience of each of the 11 individuals, which is reflected in a wide variability of the accuracy of their screening efforts. The proportion of missed studies among the 11 screeners ranged from 0% to 21% in experienced reviewers and from 0% to 58% in reviewers with less experience [12]. Although the theoretical rationale for dual-reviewer screening in systematic reviews is compelling, precise quantification of the proportion of relevant studies that will be missed with single-reviewer screening is prudent.

The objective of our methodical study was to assess the accuracy of single- and dual-reviewer screening with a reviewer base sufficiently large enough to eliminate the impact of the subjective skills and expertise of individual screeners. In addition, we strove to assess the impact of experience and domain knowledge of investigators on the accuracy of the abstract screening process in a pharmacological and a public health topic.

2. Methods

The protocol of our study was registered in the Open Science Framework (https://osf.io/3jyqt). We made one

modification to the registered protocol; originally, we had planned a third study group for which the Cochrane RCT (randomized controlled trial) Classifier would have prescreened abstracts. Because the RCT Classifier eliminated only 10% of abstracts during the pilot phase of our study, we dropped this study group because the body of abstracts would have been too similar to the pharmacological group.

2.1. Study design and general approach

Our study design was an online, parallel-group RCT with the intention to create two similar groups of screeners, one for a pharmacological and the other for a public health topic. Within each group, we randomly assigned 100 abstracts to each participant. Fig. 1 depicts the design of the study.

2.2. Data sources for abstract screening

We selected two published systematic reviews, one on a pharmacological [14] (pharmacological versus nonpharmacological interventions for depression), and the other on public health [15] (environmental interventions to reduce the consumption of sugar-sweetened beverages) intervention as sources for abstract screening. The reviews had to meet the following eligibility criteria: (1) key questions focused on efficacy or effectiveness of interventions, (2) reviews contained at least 2,000 records with at least 40 included studies, (3) decisions regarding inclusions and exclusions of abstracts were based on decisions from two independent screeners, and (4) decisions regarding inclusions or exclusions at abstract and full-text levels were consistently coded in a bibliographic database. The pharmacological review included RCTs only; the public health review both randomized and nonrandomized studies. We chose these reviews because the institutions that produced the reviews (AHRQ and Cochrane) are known to employ high methodological standards.

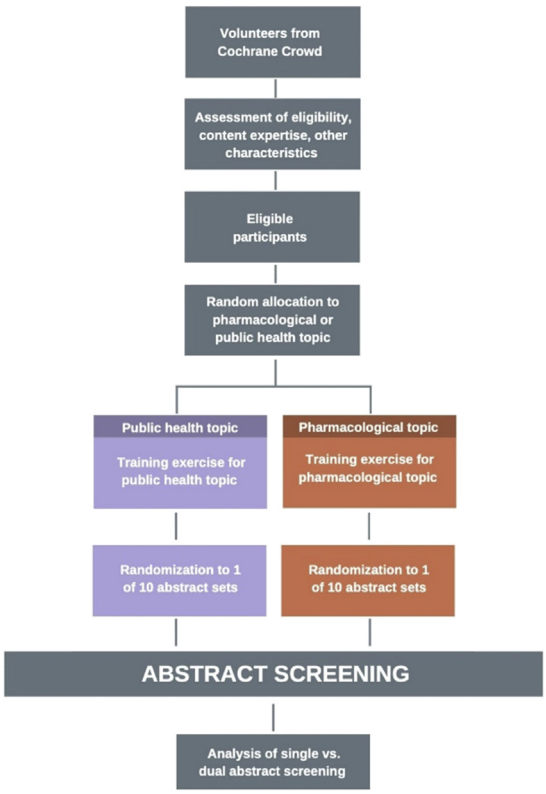

Fig. 1. Design of the study.

From each systematic review, we randomly selected 10 times 100 abstracts (10 abstract sets for each topic). We used a stratified sampling technique so that each set of 100 abstracts contained at least 15% of abstracts coded as ''Includes'' and at least 4% of abstracts of studies that were included in the reference standard systematic reviews. To achieve independent samples, we sampled without replacement, which means that a single abstract could be part of only one abstract set.

2.3. Study participants and sample size

To recruit potential study participants, we used professional networks and Cochrane Crowd, Cochrane's citizen science platform, which hosts a global community of more than 14,000 citizens and researchers who undertake identification and classification tasks to support Cochrane [16]. We approached potential participants by email, explaining the objective of the study. To be eligible, participants had to meet the following criteria: (1) ability to understand English, (2) prior experience with abstract screening, (3) willingness to donate 2 to 3 hours of their time, and (4) be aged 18 years or older. Participants were not allowed to be authors of the reference standard systematic reviews. Participant eligibility was determined by a short questionnaire with a forced-choice, closed answering format. In addition, we assessed participants' personal characteristics about age, gender, review experience, and domain knowledge. A 5-point Likert scale was applied to questions addressing personal expertise and domain knowledge (see Supplementary File 1). Participants who did not meet the eligibility criteria were assigned to an exploratory group to pilot-test Cochrane's Screen 4 Me approach, which combines known assessments, Cochrane's RCT Classifier, and Cochrane Crowd to screen abstracts. We will report the results of this pilot study in a separate publication.

We determined a priori that a sample size of 250 participants would render a two-sided 95% confidence interval for a single mean that extends 0.036 from the observed average proportion of correct decisions, assuming that the standard deviation is 0.29 and the confidence interval is based on a large sample z-statistic. To derive a conservative estimate for the standard deviation, we assumed that the proportion of correct decisions over all the raters is uniformly distributed, which gives a standard deviation of 0.29 = (Sqrt(1/12)). The observed standard deviation in

our study sample was 0.25, which resulted in higher precision than anticipated.

2.4. Allocation of participants

To implement abstract screening, we used the Cochrane Crowd platform, which employed simple randomization to assign eligible participants in a 1:1 ratio to the pharmacological or the public health topic. Investigators were kept unaware of the randomization sequence due to the online nature of the study and the built-in randomization algorithm of Cochrane Crowd.

Before screening abstracts, all participants underwent topic-specific training about inclusion and exclusion criteria. In Cochrane Crowd, participants received background information about the topic and a table with inclusion and exclusion criteria, presented in a downloadable format during the actual screening (See Supplementary File 2). Each participant went through an interactive training exercise of 15 abstracts with feedback about the correctness of the inclusion or exclusion decision. The goal of the training was to ''calibrate'' participants for the task. To proceed to abstract screening, participants had to answer 80% (12 out of 15 abstracts) of the training set correctly. Participants had unlimited attempts to achieve the training set goal.

Once they had successfully completed the training exercise, participants proceeded to abstract screening. Cochrane Crowd randomly allocated an abstract set of 100 abstracts to each participant. Participants had to either include or exclude each abstract. No neutral options, such as ''maybe'' or ''do not know'' were provided. Participants were instructed to include abstracts, in cases of uncertainty or insufficient information for a definitive decision.

We initially pilot-tested the training exercise and the abstract screening with 10 volunteers and adjusted the process based on their feedback.

2.5. Outcomes

Our primary outcome of interest was the accuracy (sensitivity and specificity) to correctly classify relevant and irrelevant studies during single- or dual-reviewer abstract screening. As secondary outcomes, we explored the extent of variability of decisions of reviewers on an abstract level, and the impact of experience and domain knowledge on the correctness of decisions.

2.6. Data collection and statistical analysis

We pseudonymized data, which was then stored securely and protected from unauthorized use. Decisions about inclusions and exclusions of studies in the two selected systematic reviews (i.e., the final decisions about inclusions or exclusions of publications) served as reference standards. We defined *true positive* (TP) decisions as the number of publications correctly identified as includes, *true negative* (TN) decisions as the number of publications correctly identified as excludes, *false negative* (FN) decisions as the number of publications incorrectly classified as excludes (FN decisions are also referred to as ''missed studies''), and *false positive* (FP) decisions as the number of publications incorrectly classified as includes. Sensitivity is the ability to correctly classify the relevant publications as includes:

$$\frac{TP}{(\mathrm{TP}+\mathrm{FN})}$$

Specificity is the ability to correctly classify the irrelevant publications as excludes:

$$\frac{TN}{(\mathrm{TN}+\mathrm{FP})}$$

Because of the multilevel structure of the data, which implies that observations are not independent, we calculated estimates and confidence intervals using logistic mixed models. We modeled the probability of inclusion per abstract, including ''reviewer'' and ''abstract'' as a random effect and study group (pharmacological or public health topic) as a fixed effect. The random effect ''reviewer'' models differences in the overall inclination of reviewers to include abstracts, the random effect ''abstract'' models differences between abstracts with respect to the probability to be included. For participants who only partially completed the task, we included any abstract with a screening decision in the model.

We estimated sensitivity and specificity by using two separate models. For sensitivity, we modeled the probability of inclusion using the subset of abstracts for which the correct decision was inclusion (TP + FN). For specificity, we modeled the probability of exclusion using the subset of abstracts for which the correct decision was exclusion (TN + FP). The estimates for single-reviewer screening were obtained directly from the model by performing the inverse logit transformation of the response and integrating over the distribution of the random effects. For dual-reviewer screening, we evaluated all pairwise combinations of screeners for a given abstract and assumed that the inclusion of an abstract by one screener of a given pair, resulted in inclusion. We first calculated the probability that two randomly chosen reviewers excluded an abstract given the random factor ''abstract.'' This probability was then averaged over the distribution of the random factor ''abstract.''

To explore the impact of native speaker status, topic expertise, or experience with literature reviews on the correctness of decisions, we conducted a logistic regression with native speaker status, topic expertise, gender, and experience with systematic reviews as covariates. We used R version 3.5.1 (2018-07-02) for all statistical analyses.

3. Results

We screened the eligibility of 491 volunteers between May 23 and June 30, 2019, of whom 379 (77%) met

inclusion criteria. The Cochrane Crowd algorithm randomly allocated 202 participants to the pharmacological topic and 177 to the public health topic. All participants had to first complete the topic-specific training exercises successfully before advancing to the abstract screening stage. With an unlimited number of attempts, participants were required to correctly classify 80% of the abstracts in the training set. Forty-three participants (21%) in the pharmacological group and 52 (29%) in the public health group did not complete the training exercises successfully and could not proceed with the study. Overall, 280 reviewers (74% of those randomized) advanced to the abstract screening stage (159 on the pharmacological topic and 121 on the public health topic). Of these, 239 (85%) completed the review of all 100 assigned abstracts. Partial completers screened between 2 and 99 abstracts. All screened abstracts were included in the final analyses. Fig. 2 presents the flow of participants from enrollment to analysis. Table 1 summarizes the characteristics of the participants.

3.1. Sensitivity and specificity of single-reviewer and dual-reviewer screening

In total, reviewers made 24,942 screening decisions about the inclusion or exclusion of abstracts. On average, each of the 2,000 randomly selected abstracts was screened 12 times (range 5 to 22 times). For the pharmacological topic, 87% of reviewers screened all 100 assigned abstracts, for the public health topic, 83% of reviewers screened all 100. On average, reviewers chose to include 23% of the pharmacological abstracts and 26% of public health abstracts.

Fig. 3 summarizes sensitivities and specificities of single- and dual-reviewer abstract screening overall, and by topic. Overall, single-reviewer abstract screening missed 13.4% of relevant studies (sensitivity: 86.6%; 95% confidence interval [CI], 80.6%–91.2%). By comparison, dual-reviewer abstract screening missed 2.5% of relevant studies (sensitivity: 97.5%; 95% CI, 95.1%–98.8%). Corresponding specificities were 79.2% (95% CI, 77.4%–80.9%) for single-reviewer screening and 68.7% (95% CI, 66.4%–71.0%) for dual-reviewer screening. Single-reviewer screening missed more relevant studies for the public health topic than the pharmacological topic (16.8% vs. 10.5%).

Across all 10 abstract sets of the pharmacological topic, the proportion of falsely excluded studies (i.e., false negative decisions) ranged from 0% to 36% for single-reviewer abstract screening and from 0% to 23% for dual-reviewer abstract screening. The corresponding numbers for the public health topic ranged from 26% to 55% of missed studies for single-reviewer screening and from 5% to 38% for dual-reviewer screening (see Supplementary File 3).

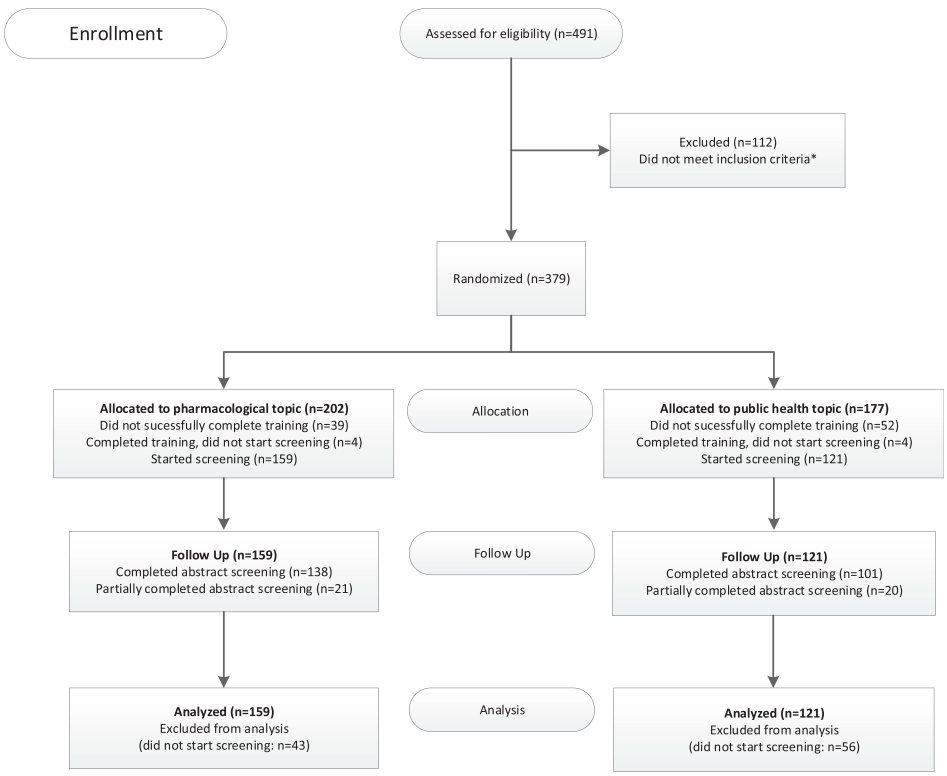

Fig. 2. Flow diagram of participants of the single-reviewer screening trial. * Participants who did not meet eligibility criteria were assigned to pilot-test Cochrane's Screen 4 Me approach.

Table 1. Characteristics of participants by assigned topic

Participants characteristics	Pharmacological topic (n = 159)	Public health topic (n = 121)
Age (mean [SD], years; n = 280)	38.2 (12.0)	39.3 (11.1)
Gender (proportion, female; n = 149[a])	60.2%	65.6%
English native speaker (n = 280)	25.8%	25.6%
Professional background (n = 208[a])		
Researcher	38.6%	34.4%
Health professional	29.5	34.4%
Student	11.4%	4.9%
Other	20.5%	26.2%
Number of reviews, participants have screened abstracts for (median [IQR]; n = 144[a])	8.5 (3.0–20.0)	9.0 (3.3–25.0)
Self-rating of experience with systematic reviews (n = 144[a])		
Very good	36.0%	19.0%
Good	45.3%	48.3%
Average	14.0%	24.1%
Below average	4.7%	8.6%
Self-rating of topic expertise (n = 149[a])		
Very good	8%	3%
Good	22%	28%
Average	40%	36%
Below average	31%	33%

Abbreviations: IQR, interquartile range; n, number of participants; SD, standard deviation.
[a] Because of a technical problem, data were not recorded for all participants.

3.2. Secondary outcomes

Regression analyses had not detected the statistically significant impact of native speaker status, topic expertise, or experience with literature reviews on the correctness of decisions.

As a secondary outcome, we explored the extent of variability of decisions by individual reviewers compared with decisions during abstract review in the two systematic reviews that served as reference standards. On average, the decisions of reviewers to include or exclude abstracts were

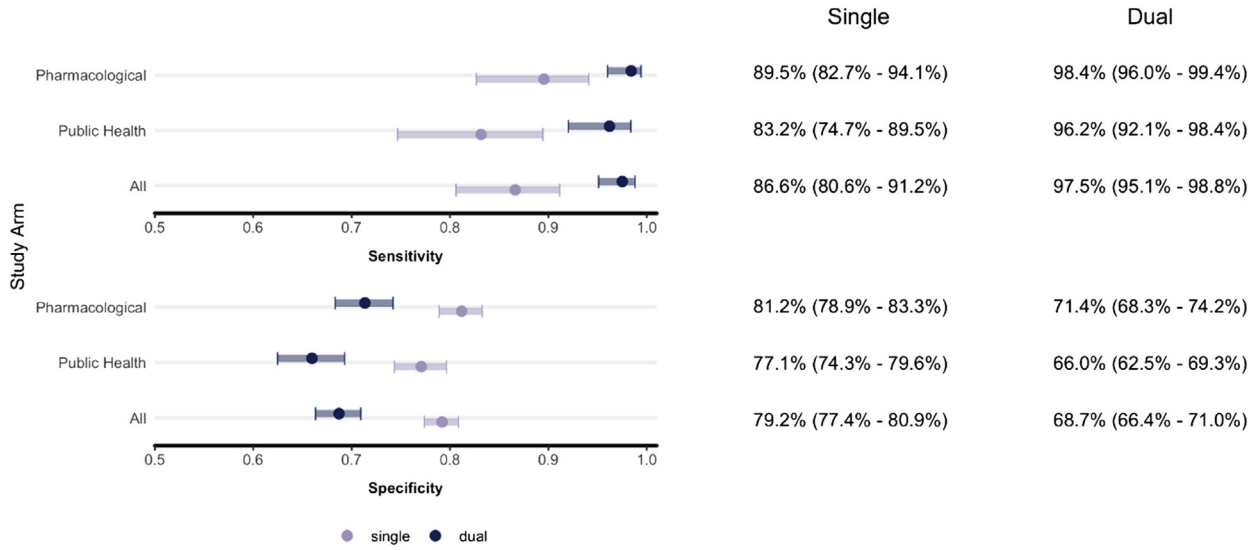

Fig. 3. Sensitivities and specificities for single- and dual-reviewer abstract screening.

consistent with 64.9% (95% CI, 61.1%−68.6%) of decisions to include and 84.1% (95% CI, 82.4%−85.7%) of decisions to exclude abstracts in the reference standard reviews. On average, single reviewers classified 15% of pharmacological abstracts and 16% of public health abstracts differently than the reference standard systematic reviews.

4. Discussion

Our study is the largest attempt to date, to quantify the accuracy of single- and dual-reviewer abstract screening. We employed a crowd-based RCT with 280 participants to mitigate the impact of individual screening experience and topic expertise on the accuracy of screening decisions. Previous studies that assessed the same research question involved only two to four reviewers, which makes findings highly dependent on individual skill levels of those who screen abstracts [13,17−20]. Each abstract in our study was classified by a mean of 12 reviewers. In our study, single-reviewer screening missed, on average, 13.4% of relevant studies. By comparison, dual-reviewer screening falsely excluded only 2.5% of the relevant studies. The reason why dual-reviewer screening failed to achieve perfect sensitivity is that published systematic reviews served as reference standards. In systematic reviews, cross-checking reference lists of other publications and external peer review are important methodological steps to rectify false exclusions during literature screening. Accuracy results were lower for the more complex public health topic compared with the clinical topic that included only RCTs as eligible study design.

Our findings are less optimistic than those reported by the systematic review by Waffenschmidt et al., who calculated a median proportion of missed studies for a single-reviewer screening of 5%, albeit with a wide range of 0%−58% [12]. Even the lower bound of the confidence interval of our results (95% CI, 9%−19% of missed studies), does not encompass their estimate.

Given these results, single-reviewer abstract screening does not appear to fulfill the high methodological standards that decision makers expect from systematic reviews. Even under the best-case scenario, taking the lower bound of the confidence interval into account, single-reviewer screening would still miss at least 9% of relevant studies.

The acceptability of this trade-off, however, might be different for rapid reviews, which have evolved as a pragmatic alternative to systematic reviews. Rapid reviews are knowledge syntheses that streamline standard systematic review methods to accelerate the production of the review [21]. Decision makers using rapid reviews are willing to give up some degree of certainty regarding the correctness of results. An international survey of guideline developers and decision makers reported that users of rapid reviews are willing to accept 10% of incorrect results in exchange for a more rapid evidence synthesis product [22].

Our study has several limitations. First, results are based on only two topics and two systematic reviews as reference standards. Although the two institutions (AHRQ and Cochrane) that conducted the reviews employ high methodological standards, the two reviews might have still missed relevant studies (i.e., they are imperfect reference standards). Furthermore, it is unclear how generalizable our findings are to other topics or to other types of reviews, such as diagnostic or prognostic reviews. Second, it is difficult for any study to create a realistic environment that is generalizable to real-world situations. Although participants in our study received brief topic-specific training, their average domain knowledge might have still been less than that of an investigator who screens abstracts for a systematic review on these topics. Only 8% of participants in the pharmacological group and 3% in the public health group rated their domain knowledge as very good. By comparison, authors of both reference standard reviews presumably had very high content expertise. The author group of the pharmacological review consisted of professional systematic reviewers of an AHRQ-funded Evidence-based Practice Center with a focus on mental health; authors of the public health review were German public health experts. A lack of domain knowledge might cause reviewers to perceive more decisions as unclear. Because systematic reviewers generally err on the side of inclusion if decisions are unclear, the fairly high proportion of included abstracts in our study (23% for the pharmacological topic and 26% for the public health topic) might be a consequence. Higher inclusion rates during abstract screening, however, favor sensitivity and might have led to slightly more optimistic estimates for sensitivity than would be observed in a real-life systematic review. Third, for calculating estimates for dual-reviewer screening, we assumed that an inclusion decision by a single reviewer would include an abstract. In reality, inclusion decisions on an abstract level are sometimes made dually to reduce the workload during a full-text review. Our approach might additionally favor sensitivity because abstracts can also be falsely excluded during the adjudication process when screeners have to agree about the inclusion or exclusion of an abstract. Fourth, 24% of randomized participants did not successfully complete their training sets and, thus, were not allowed to proceed to the actual screening exercise. The proportion was higher for the public health topic (29%) than for the pharmacological topic (19%). It is unclear how such postrandomization exclusions affect accuracy results. The training exercise for the public health topic might have been more challenging than the one for the pharmacological topic. Since our intent was not to compare results of the two randomized groups, any differential bias introduced by these postrandomization exclusions, would not have had an effect on our results. Fifth, because of a programming error, participants' characteristics were not recorded for all study participants. Although this loss of data was entirely at random, it could have affected the precision of participants' characteristics

in Table 1 and the power of the regression analysis. It did not have an effect on determining participants' eligibility for the study or the main analyses. Finally, we did not assess the impact of missed studies on meta-analyses or conclusions. It is conceivable that studies that were missed by single-reviewer screening would have little impact on the conclusions of the systematic reviews. Previous work by Nussbaumer-Streit et al. on the impact of abbreviated literature searches reported that missed studies rarely changed the conclusions of systematic reviews [23].

Because studies that are missed through abbreviated literature searches might be different than studies missed through single-reviewer screening, future research needs to explore the impact of such studies on the results of meta-analyses and conclusions. In addition, methodological studies need to explore how studies that were falsely excluded during literature screening, can best be recovered again through searching nonbibliographic information sources, such as reference lists of published reviews, included studies, or editorials after literature screening. More research is also required on how to use machine-learning tools to reduce false decisions during single-reviewer screening. Web applications that employ natural language processing technologies to support systematic reviewers during abstract screening have become more user friendly and more common. To date, however, the results of studies assessing whether semi-automated screening tools can improve the accuracy of single-reviewer screening are conflicting.

5. Conclusions

Given that single-reviewer abstract screening misses about 13% of relevant studies, this approach should probably not be used for systematic reviews. Institutions, such as AHRQ, Cochrane, or NICE that allow single-reviewer abstract screening in their methods documents, should reconsider their methodological guidance in light of these new findings. Single-reviewer screening, however, might be a viable option for rapid reviews, which deliberately lower methodological standards to provide decision makers with accelerated evidence synthesis products.

CRediT authorship contribution statement

Gerald Gartlehner: Conceptualization, Methodology, Project administration, Funding acquisition, Writing - original draft, Writing - review & editing. **Lisa Affengruber:** Conceptualization, Methodology, Project administration, Writing - review & editing. **Viktoria Titscher:** Methodology, Writing - review & editing. **Anna Noel-Storr:** Conceptualization, Methodology, Project administration, Writing - review & editing. **Gordon Dooley:** Conceptualization, Investigation, Data curation, Writing - review & editing. **Nicolas Ballarini:** Formal analysis, Supervision, Writing - review & editing. **Franz König:** Formal analysis, Supervision, Writing - review & editing.

Acknowledgments

We would like to thank all our colleagues from around the world who donated their time to participate in our study. We are grateful to Jan Stratil and colleagues who gave us access to the bibliographic database of their Cochrane review. We also want to thank our colleagues from the Department for Evidence-based Medicine and Evaluation at Danube University Krems for piloting the study and for their invaluable feedback, as well as Martin Posch from the Medical University of Vienna for statistical input. Many thanks also to Sandra Hummel from Cochrane Austria for her administrative support throughout the project and to Emma Persad for help with the figures.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2020.01.005.

References

[1] Institute of Medicine. Finding what works in health care: standards for systematic reviews. Washington, DC: The National Academies Press; 2011. Available at http://www.nationalacademies.org/hmd/Reports/2011/Finding-What-Works-in-Health-Care-Standards-for-Systematic-Reviews.aspx. Accessed December 26, 2019.

[2] University of York CfRaD. CRD's guidance for undertaking reviews in health care. 2008. Available at https://www.york.ac.uk/media/crd/Systematic_Reviews.pdf. Accessed December 26, 2019.

[3] Institute for Quality and Efficiency in Health. General Methods Version 5.0 2019. Available at https://www.iqwig.de/en/methods/methods-paper.3020.html. Accessed December 26, 2019.

[4] Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. Amstar 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017;358:j4008.

[5] Whiting P, Savovic J, Higgins JP, Caldwell DM, Reeves BC, Shea B, et al. ROBIS: a new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol 2016;69:225–34.

[6] Agency for Healthcare Research and Quality. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Chapters available at: Rockville, MD. 2014 [AHRQ Publication No. 10(4)-EHC063-EF]. Available at https://effectivehealthcare.ahrq.gov/products/methods-guidance-bias/methods. Accessed December 26, 2019.

[7] Campbell Systematic Reviews. Campbell Policies and Guidelines Series No. 3. Methodological expectations of Campbell Collaboration intervention reviews: Conduct standards. 2019. Available at https://onlinelibrary.wiley.com/page/journal/18911803/homepage/author-guidelines. Accessed December 26, 2019.

[8] JHiggins JPT, Lasserson T, Chandler J, Tovey D, Thomas J, Flemyng E, et al. Methodological Expectations of Cochrane Intervention Reviews. London: Cochrane; 2019. Available at https://community.cochrane.org/sites/default/files/uploads/inline-files/MECIR%20October%202019%20Final%20Online%20version.pdf. Accessed December 26, 2019.

[9] National Institute for Health and Care Excelllence. Developing NICE guidelines: the manual 2014. Available at https://www.nice.org.uk/process/pmg20/chapter/reviewing-research-evidence#identifying-and-selecting-relevant-evidence. Accessed December 26, 2019.

[10] Higgins JPTTJ, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. Cochrane Handbook for Systematic Reviews of Interventions, Version 6 (updated July 2019); [Chapter 4]: Searching for and selecting studies. Cochrane. Available at https://training.cochrane.org/handbook/current. Accessed December 26, 2019.

[11] 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.

[12] Waffenschmidt S, Knelangen M, Sieben W, Buhn S, Pieper D. Single screening versus conventional double screening for study selection in systematic reviews: a methodological systematic review. BMC Med Res Methodol 2019;19:132.

[13] Stoll CRT, Izadi S, Fowler S, Green P, Suls J, Colditz GA. The value of a second reviewer for study selection in systematic reviews. Res Synth Methods 2019;10:539−45.

[14] Gartlehner G, Gaynes BN, Forneris C, Lohr KN. Comparative benefits and harms of antidepressant, psychological, complementary, and exercise treatments for major depression. Ann Intern Med 2016;165:454.

[15] von Philipsborn P, Stratil JM, Burns J, Busert LK, Pfadenhauer LM, Polus S, et al. Environmental interventions to reduce the consumption of sugar-sweetened beverages and their effects on health. Cochrane Database Syst Rev 2019;(6):CD012292.

[16] Cochrane Collaboration. Cochrane Crowd. 2019. Available at https://community.cochrane.org/help/data-management-tools/cochrane-crowd. Accessed February 6, 2020.

[17] 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.

[18] 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.

[19] Pham MT, Waddell L, Rajić A, Sargeant JM, Papadopoulos A, McEwen SA. Implications of applying methodological shortcuts to expedite systematic reviews: three case studies using systematic reviews from agri-food public health. Res Synth Methods 2016;7(4):433−46.

[20] 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(1):140.

[21] Khangura S, Konnyu K, Cushman R, Grimshaw J, Moher D. Evidence summaries: the evolution of a rapid review approach. Syst Rev 2012;1:10.

[22] Wagner G, Nussbaumer-Streit B, Greimel J, Ciapponi A, Gartlehner G. Trading certainty for speed - how much uncertainty are decisionmakers and guideline developers willing to accept when using rapid reviews: an international survey. BMC Med Res Methodol 2017;17:121.

[23] Nussbaumer-Streit B, Klerings I, Wagner G, Heise TL, Dobrescu AI, Armijo-Olivo S, et al. Abbreviated literature searches were viable alternatives to comprehensive searches: a meta-epidemiological study. J Clin Epidemiol 2018;102:1−11.