Journal of Clinical Epidemiology 162 (2023) 72−80

ORIGINAL ARTICLE

Sensitivity and specificity of alternative screening methods for systematic reviews using text mining tools

Jimmy Li[a,b], Joudy Kabouji[c], Sarah Bouhadoun[d], Sarah Tanveer[e], Kristian B. Filion[f,g], Genevieve Gore[h], Colin Bruce Josephson[i,j,k,l,m], Churl-Su Kwon[n], Nathalie Jette[i,j], Prisca Rachel Bauer[o], Gregory S. Day[p], Ann Subota[j,q], Jodie I. Roberts[j], Sara Lukmanji[j], Khara Sauro[j,r,s], Adnane Alaoui Ismaili[t], Feriel Rahmani[u], Khadidja Chelabi[v], Yasmine Kerdougli[w], Nour Meryem Seulami[x], Aminata Soumana[w], Sarah Khalil[w], Noémie Maynard[t], Mark Robert Keezer[b,f,y,z,*]

[a]*Neurology Division, Centre Hospitalier de l'Université de Sherbrooke (CHUS), Sherbrooke, Canada*
[b]*Centre de Recherche du Centre Hospitalier de l'Université de Montréal (CRCHUM), Montreal, Canada*
[c]*Department of Pharmacy, University of Laval, Quebec City, Canada*
[d]*Department of Neurology, McGill University, Montreal, Canada*
[e]*Department of Pharmaceutical Health Services Research, University of Maryland, Baltimore, MD, USA*
[f]*Departments of Medicine and of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Canada*
[g]*Centre for Clinical Epidemiology, Jewish General Hospital − Lady Davis Institute, Montreal, Canada*
[h]*Schulich Library of Physical Sciences, Life Sciences, and Engineering, McGill University, Montreal, Canada*
[i]*Department of Clinical Neurosciences, University of Calgary, Calgary, Canada*
[j]*Department of Community Health Sciences, University of Calgary, Calgary, Canada*
[k]*O'Brien Institute for Public Health, University of Calgary, Calgary, Canada*
[l]*Hotchkiss Brain Institute, University of Calgary, Calgary, Canada*
[m]*Center for Health Informatics, University of Calgary, Calgary, Canada*
[n]*Department of Neurology, Epidemiology, Neurosurgery and the Gertrude H. Sergievsky Center, Columbia University, New York, NY, USA*
[o]*Department of Psychosomatic Medicine and Psychotherapy, Faculty of Medicine, Medical Center − University of Freiburg, Freiburg, Germany*
[p]*Department of Neurology, Mayo Clinic Florida, Jacksonville, FL, USA*
[q]*Department of Medicine, University of Calgary, Calgary, Canada*
[r]*Department of Surgery, University of Calgary, Calgary, Canada*
[s]*Department of Oncology & Arnie Charbonneau Cancer Institute, University of Calgary, Calgary, Canada*
[t]*Department of Internal Medicine, McGill University, Montreal, Canada*
[u]*Faculty of Medicine, McGill University, Montreal, Canada*
[v]*Department of Pediatrics, McGill University, Montreal, Canada*
[w]*Department of Family Medicine, McGill University, Montreal, Canada*
[x]*Department of Emergency Medicine, McGill University, Montreal, Canada*
[y]*Department of Neurosciences, Université de Montréal, Montreal, Canada*
[z]*School of Public Health, Université de Montréal, Montreal, Canada*

Accepted 19 July 2023; Published online 26 July 2023

Abstract

Objectives: To evaluate the impact of text mining (TM) on the sensitivity and specificity of title and abstract screening strategies for systematic reviews (SRs).

Study Design and Setting: Twenty reviewers each evaluated a 500-citation set. We compared five screening methods: conventional double screen (CDS), single screen, double screen with TM, combined double screen and single screen with TM, and single screen with TM. Rayyan, Abstrackr, and SWIFT-Review were used for each TM method. The results of a published SR were used as the reference standard.

Data availability: Data will be publicly available on in an online data repository upon manuscript acceptance (this statement will be updated).

* Corresponding author. Centre Hospitalier de l'Université de Montréal, 1000 rue Saint-Denis, Montréal, Quebec H2X 0C1, Canada. Tel.: +01 514 890 8233; fax: +01 514 412 7139.
E-mail address: mark.keezer@umontreal.ca (M.R. Keezer).

https://doi.org/10.1016/j.jclinepi.2023.07.010

Results: The mean sensitivity and specificity achieved by CDS were 97.0% (95% confidence interval [CI]: 94.7, 99.3) and 95.0% (95% CI: 93.0, 97.1). When compared with single screen, CDS provided a greater sensitivity without a decrease in specificity. Rayyan, Abstrackr, and SWIFT-Review identified all relevant studies. Specificity was often higher for TM-assisted methods than that for CDS, although with mean differences of only one-to-two percentage points. For every 500 citations not requiring manual screening, 216 minutes (95% CI: 169, 264) could be saved.

Conclusion: TM-assisted screening methods resulted in similar sensitivity and modestly improved specificity as compared to CDS. The time saved with TM makes this a promising new tool for SR.

Keywords: Knowledge synthesis; Rayyan; Abstrackr; SWIFT-Review; Artificial intelligence; Machine learning; Diagnostic study; Sensitivity; Specificity

1. Introduction

Systematic reviews (SRs) are a valuable means of synthesizing scientific knowledge and are regarded as one of the highest standards of evidence in evidence-based medicine [1]. A high-quality SR requires a rigorous, systematic, transparent screening of the scientific literature. This process requires, however, extensive resource allocation including substantial demands on researcher time and financial costs. One study estimated that an average of 332 hours is spent on the screening of titles, abstracts, and full-text articles for a single SR [2]. It is not uncommon for electronic database searches to begin with more than 5,000 titles and abstracts to screen [3]. As such, it is important to use screening methods that maximizes sensitivity so as to not miss any relevant articles but also specificity to minimize reviewer workload.

The conventional double screen (CDS) is a two-step selection process recommended by the Cochrane Collaboration to minimize the risk of missing relevant articles (i.e., optimize sensitivity) [4]. This process involves a first screen of all titles and abstracts by two independent reviewers, followed by a second screen of the full text of all citations identified as potentially relevant by either reviewer during the first screen. Relevant articles are then selected for inclusion in the SR.

More recent studies have examined whether incorporating text mining (TM) into the screening process is an accurate, efficient, cost-effective method for identifying eligible studies in SRs [5]. TM involves the identification of prediction patterns based on supervised machine learning algorithms, computational linguistics, and statistics to extract particular knowledge from textural data sources [6,7]. In the context of SRs, after observing a reviewer assess a training set of citations, the algorithm will predict whether subsequent citations are likely to meet the same inclusion and exclusion criteria. Numerous TM tools are available, with several of these being free-to-use and/or open-source [8].

It remains uncertain whether there are more time-efficient but equally accurate methods, as compared to the CDS method, to carry out the screening of citations in the context of an SR. Our aim was to compare different screening methods, including different approaches to using TM tools, evaluating the impact of each method on sensitivity and specificity. We also assessed the impact of reviewers' prior experience with SR screening on their performance as well as the impact of the method used on the time required to complete the screening process.

2. Methods

2.1. Citation database

We tested different screening methods on a citation database of 500 titles and abstracts that were used in a 2019 SR that examined the association between pregnancy complications and the subsequent risk of cardiovascular disease [9]. In this SR, 13,969 studies were screened and 83 were ultimately included. None of the reviewers were familiar with this topic. Our test set of 500 citations was sampled from the 2019 SR so that 10% of citations should be identified as relevant to the research question. This higher percentage of relevant citations (as opposed to that seen in many SRs) was to ensure that the performance (especially the sensitivity) of the reviewers was less affected by a random chance. Inclusion and exclusion criteria provided to reviewers were the same as those used in the published SR [9] and can be found in Table S1.

To train the TM tools used in this study, a ''training set'' of 200 articles (separate from the database of 500 citations) was randomly selected based on the results of the 2019 SR by a member of the research team who did not participate as a reviewer [9]. The training set size was consistent with previous studies evaluating TM in SRs [10,11]. We randomly chose 15 citations that were deemed relevant to the topic by the 2019 SR to be included in this training set. The test set and training set are available on OSF (https://osf.io/a5d6t/).

2.2. Screening methods

We tested five methods to screen titles and abstracts, including:

1) Conventional double screen (CDS): All titles and abstracts were screened by two reviewers.

What is new?

Key findings

- Using text mining tools to assist in screening titles and abstracts for a systematic review may yield sensitivity and specificity similar to the conventional double screening method.
- For every 500 citations not requiring manual screening, 216 minutes on average can be saved.

What this adds to what was known?

- How factors such as the training set size, the complexity of the research question, and the software at use may affect text mining performance in title and abstract screening represents an important avenue for future research.

What is the implication and what should change now?

- Integrating text mining into the title and abstract screening phase for systematic reviews can be cautiously considered on a case-by-case basis as a means of lowering workload while maintaining sensitivity and specificity.

2) Single screen (SS): All titles and abstracts were screened by one reviewer.
3) Double screen with TM (DSTM): A TM algorithm classified citations as having a high vs. low likelihood of inclusion. High likelihood citations were reviewed by two reviewers; low likelihood citations were excluded.
4) Combined double and single screen with TM (CDSSTM): Same method as DSTM, except low likelihood citations were reviewed by one of the two reviewers.
5) Single screen with TM (SSTM): After TM classification, high likelihood citations were reviewed by a single reviewer; low likelihood citations were excluded.

When a method called for two reviewers to screen the same titles and abstracts, a positive screen was defined as one where either reviewer deemed the article relevant to the research question.

2.3. Text mining

Rayyan [12], Abstrackr [13], and SWIFT-Review [14] were selected from the numerous TM platforms available for SR citation selection [8], due to their accessibility, popularity, and ease of use. Rayyan is a paid online service (although a free trial is available) that rates the applicability of yet-to-be screened citations using a five-star scheme, one being the least relevant and five being the most relevant [12]. In this study, citations with 2.5 stars or more were deemed high likelihood; citations with less than 2.5 stars were deemed low likelihood. This cut-off was chosen as it has been shown to maximize sensitivity [15]. Abstrackr is a free online service which attributes a prediction score from 0 to 1 to yet-to-be screened citations [13]. Once the prediction score drops below 0.40, all remaining citations are generally predicted to be irrelevant [11,13]. We considered high likelihood citations to have scores of at least 0.40 and low likelihood citations to have scores below 0.40. SWIFT-Review is a free program that ranks yet-to-be screened citations based on a prediction score from 0 to 1 [14]. Unlike Abstrackr, determining a cut-off likelihood of inclusion is more dynamic. When trained, SWIFT-Review generates a ranking performance graph from which the proportion of the highest ranked citations predicted to be relevant can be determined. In our case, all citations deemed relevant occurred in the top 10% of the ranked list (Figure S1). We chose a score of 0.55 (above which rested the top 20% of the ranked list; the additional 10% acting as a safety net) as our cut-off. Although Rayyan, Abstrackr, and SWIFT-Review were each used in the DSTM, CDSSTM, and SSTM screening methods, nine unique combinations of screening methods with TM software (e.g., DSTM with Rayyan) were evaluated. In each case, the same training set of 200 citations was fed into the program, and the likelihood of inclusion of the remaining 500 citations was assessed.

2.4. Reviewers

Twenty reviewers participated in the screening and were categorized according to their level of expertise. The 10 ‘‘expert reviewers’’ were healthcare experts who have previously participated in the screening process of at least two published SRs. The 10 ‘‘nonexpert reviewers’’ were medical students who had never carried out an SR. All reviewers attended an orientation session where their role in the study was outlined. Each reviewer reviewed the citation database only once and recorded the time taken to complete their review. The five different screening methods were reconstructed using this single review by each reviewer using methods outlined in Table S2.

2.5. Data analysis

We first calculated the sensitivity and specificity of each reviewer or reviewer duo for every screening method (i.e., CDS, SS, DSTM, CDSSTM, and SSTM). We then calculated the mean sensitivity and specificity across all reviewers or reviewer duos for every screening method, with corresponding 95% confidence intervals (CIs). Sensitivity was defined as the number of citations that were correctly included for a full-text review (true positives) divided by the total number identified by our reference standard. Specificity was defined as the number of citations that were correctly excluded (true negatives) divided by the

total number excluded by our reference standard. This reference standard was the final citations included in the 2019 SR [9]. We compared the mean sensitivity and specificity of the SS, DSTM, CDSSTM, and SSTM methods vs. CDS. To do so, we calculated the difference in means between each screening method and CDS, with corresponding 95% CIs. When comparing DSTM and CDSSTM to CDS, the CI was calculated presuming paired samples. Samples were presumed to be paired when comparing double screening methods (i.e., CDS, DSTM, and CDSSTM) because the same reviewer duos were used to construct these screening methods. When comparing single screening methods (i.e., SS and SSTM) to CDS, the CI was calculated presuming unpaired samples because a sample constructed from single reviewers could not be paired with a sample constructed from reviewer duos.

For our secondary analyses, we compared the mean sensitivity and specificity between experts and nonexperts for SS, providing differences in means with 95% CIs. We also calculated the sensitivity and specificity of Rayyan's, Abstrackr's, and SWIFT-Review's raw predictions by assuming that all high likelihood citations were included and all low likelihood citations were excluded.

The mean time taken to complete the review across all 20 reviewers was calculated. The average time savings per citation not requiring review were calculated assuming that each citation took the same time to review for one same reviewer. By multiplying the original time data by the proportion of citations identified as relevant by TM divided by the total number of citations needed to be screened without TM (i.e., 500), the theoretical time required for each screening method using TM was reconstructed.

Statistical analyses were carried out using R 4.2.2 [16]. This study did not require approval by a research ethics board.

3. Results

The mean sensitivity and specificity achieved by the 10 reviewer duos as part of the CDS were 97.0% (95% CI: 94.7, 99.3) and 95.0% (93.0, 97.1), respectively. When compared to SS, CDS provided an improvement in mean sensitivity of 6.1% (1.5, 10.7) but without a difference in mean specificity (−1.8% [−3.9, 0.5]) (Table 1). The mean sensitivity and specificity achieved by the 20 individual reviewers, recreating SS, were 90.9% (86.7, 95.1) and 96.8 (95.7, 97.8), respectively (Table S3). When comparing experts to nonexperts, the difference in mean sensitivity and specificity were −1.4% (−10.0, 7.5) and 0.90% (−1.2, 3.1), respectively. Experts took on average less time (141.5 minutes [64.0, 218.9]) to complete screening than nonexperts (Table S2).

When all citations deemed high likelihood by TM were automatically included and all other citations were excluded, without humans verifying or contesting the decisions made by the TM, the sensitivity for Rayyan, Abstrackr, and SWIFT-Review was consistently 100% (92.9%, 100%). In these conditions, the specificity for Rayyan, Abstrackr, and SWIFT-Review was 26.7% (22.6, 31.0), 82.2% (78.4, 85.6), and 86.2% (82.7, 89.3), respectively. Stated simply, these TM platforms identified all relevant citations as high likelihood but identified many irrelevant citations as high likelihood as well.

The sensitivity was identical between DSTM, CDSSTM, and CDS because TM tools identified all reference standard-relevant citations as having a high likelihood for inclusion. When compared with CDS, the difference in mean specificity for DSTM and for CDSSTM ranged from 0.30% to 2.2% (Table 2) and from 0.2% to 2.0% (Table 3), respectively. When compared with CDS, the difference in mean sensitivity for SSTM was consistently −6.1% (−10.7, −1.5) (Table 4). The difference in mean specificity for SSTM ranged from 1.9% to 2.9%.

The mean time taken to complete the review of 500 citations by our reviewers was 216 minutes (169, 264), corresponding to an average per citation time of 26 seconds (20.3, 31.6). The extrapolated time savings achievable using each screening method are presented in Table 5. In summary, incorporating TM could help reduce the time to screen 500 titles and abstracts by 12.9% to 91.5%, as compared to CDS.

4. Discussion

Our study assessed the sensitivity and specificity of alternative methods for title and abstract screening using the 3 TM tools: Rayyan, Abstrackr, and SWIFT-Review. Twenty reviewers screened the same database of 500 citations so that five screening methods (i.e., CDS, SS, DSTM, CDSSTM, SSTM) could be compared. The sensitivity of DSTM and CDSSTM were unchanged in comparison with CDS independently of the TM tool used because Rayyan, Abstrackr, and SWIFT-Review were able to identify all relevant citations as having a high likelihood for inclusion. The sensitivity of SSTM was lower than that of CDS, but this difference in sensitivity was the same as between SS and CDS, indicating that the use of TM played no part in this discrepancy. Specificity was modestly higher with most TM-assisted methods than with CDS. This is unsurprising as all 3 TM tools identified hundreds of citations as being irrelevant. In the cases of DSTM and SSTM, these irrelevant citations were by definition excluded, so reviewers did not have the ''opportunity'' to include a citation that should have been excluded. In the scenario where all citations deemed high likelihood by a TM tool were automatically included and all other citations were automatically excluded, without any human intervention, specificity was particularly poor. This suggests that TM tools in a vacuum may tend to be overinclusive and that human assistance is imperative to enhance specificity. Indeed, we have shown

Table 1. Conventional double screen, no TM, with comparison to single screen

Reviewer type	Reviewer duo	Sensitivity (%)	Specificity (%)	Time required (min)
NE	1	100.0	95.8	552.8
	2	98.0	92.9	720.4
	3	100.0	95.1	606.6
	4	96.0	90.0	468.7
	5	92.0	98.0	523.5
E	6	96.0	98.7	318.0
	7	100.0	97.1	285.5
	8	92.0	97.1	368.0
	9	96.0	92.0	336.0
	10	100.0	94.0	149.9
T	Mean (95% CI)	97.0 (94.7, 99.3)	95.0 (93.0, 97.1)	432.9 (326.5, 539.3)
	Δ mean vs. SS (95% CI)	6.1 (1.5, 10.7)	−1.8 (−3.9, 0.5)	216.4 (144.7, 288.2)

Abbreviations: CI, confidence interval; E, expert; NE, nonexpert; SS, single screen; T, total (both E and NE).

The time required to complete the screening by one duo was calculated by combining the times of each member of the duo. The sensitivity, specificity, and time required to complete the screening were not compared between NE and E because this has already been done in the supplementary material (Table S2).

that combining TM decisions with some form of human verification (i.e., DSTM, DSSTM, and SSTM) leads to high specificities. Importantly, we have shown that the automated exclusion of irrelevant citations diminished workload by up to almost 80%. In the same vein, incorporating TM into the screening process could help reduce the time to screen 500 titles and abstracts by 12.9–91.5%, as compared to CDS.

It is generally accepted that CDS is preferable to SS in most cases. A 2019 review identified four studies comparing CDS to SS, with a median proportion of missed studies of 5% with SS, substantially more than with CDS [17]. Our results also favor CDS due to its superior sensitivity. Surprisingly, little is known on the impact of reviewer experience on the title and abstract screening process. We found no appreciable difference in the sensitivity and specificity between 10 expert reviewers and 10 nonexpert reviewers. This said, our citation database was purposefully chosen to represent a topic unfamiliar to all reviewers. Experts still completed their screening faster than nonexperts by 141.5 minutes on average.

Much of the research on TM in SRs has focused on the accuracy of the TM tools' raw predictions. Fewer studies have evaluated the performance of TM tools in "real-world" applications, that is, when partnered with human input in a formal screening strategy [7]. The performance of these tools when used by human reviewers is especially relevant to research applications [7,18]. A 2019 study

Table 2. Double screen of high likelihood articles with no screen of low likelihood articles, Rayyan vs. Abstrackr vs. SWIFT-review, with comparison with conventional double screen

	Rayyan		Abstrackr		SWIFT-review	
Reviewer duo	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)
1	100.0	95.8	100.0	96.9	100.0	97.6
2	98.0	93.1	98.0	95.3	98.0	96.2
3	100.0	96.0	100.0	96.4	100.0	97.8
4	96.0	90.0	96.0	93.8	96.0	94.7
5	92.0	98.2	92.0	98.0	92.0	98.2
6	96.0	98.9	96.0	98.9	96.0	98.9
7	100.0	97.3	100.0	97.8	100.0	97.8
8	92.0	97.1	92.0	97.3	92.0	98.0
9	96.0	92.4	96.0	95.6	96.0	96.4
10	100.0	94.4	100.0	96.2	100.0	96.4
Mean (95% CI)	97.0 (94.7, 99.3)	95.3 (93.3, 97.3)	97.0 (94.7, 99.3)	96.7 (95.6, 97.8)	97.0 (94.7, 99.3)	97.2 (96.3, 98.1)
Δ mean vs. CDS (95% CI)	0 (0, 0)	0.3 (0.10, 0.5)	0 (0, 0)	1.7 (0.6, 2.3)	0 (0, 0)	2.2 (0.9, 3.4)

Abbreviations: CI, confidence interval; CDS, conventional double screen.

Table 3. Double screen of high likelihood articles and single screen of low likelihood articles, Rayyan vs. Abstrackr vs. SWIFT-Review, with comparison with conventional double screen

	Rayyan		Abstrackr		SWIFT-review	
Reviewer duo	**Sensitivity (%)**	**Specificity (%)**	**Sensitivity (%)**	**Specificity (%)**	**Sensitivity (%)**	**Specificity (%)**
1	100.0	95.8	100.0	96.2	100.0	95.8
2	98.0	92.9	98.0	92.9	98.0	96.2
3	100.0	96.0	100.0	96.2	100.0	97.8
4	96.0	90.0	96.0	93.1	96.0	94.7
5	92.0	98.2	92.0	98.0	92.0	98.2
6	96.0	98.7	96.0	98.7	96.0	98.9
7	100.0	97.1	100.0	97.6	100.0	97.8
8	92.0	96.4	92.0	96.7	92.0	98.0
9	96.0	92.2	96.0	93.8	96.0	96.4
10	100.0	94.4	100.0	96.2	100.0	96.4
Mean (95% CI)	97.0 (94.7, 99.3)	95.2 (93.2, 97.2)	97.0 (94.7, 99.3)	96.0 (94.5, 97.4)	97.0 (94.7, 99.3)	97.0 (96.1, 98.0)
Δ mean vs. CDS (95% CI)	0 (0, 0)	0.2 (−0.2, 0.4)	0 (0, 0)	1.0 (0.02, 1.8)	0 (0, 0)	2.0 (0.7, 3.3)

Abbreviations: CI, confidence interval; CDS, conventional double screen.

showed that the median proportion of missed studies across three citation datasets using Abstrackr with techniques similar to our SSTM and CDSSTM methods was 5% and 0%, respectively [10]. A 2022 study evaluated a single-citation dataset using Abstrackr in a manner akin to our SSTM method and reported sensitivity and specificity

Table 4. Single screen of high likelihood articles with no screen of low likelihood articles, Rayyan vs. Abstrackr vs. SWIFT-Review, with comparison with conventional double screen

	Rayyan		Abstrackr		SWIFT-review	
Reviewer	**Sensitivity (%)**	**Specificity (%)**	**Sensitivity (%)**	**Specificity (%)**	**Sensitivity (%)**	**Specificity (%)**
1	100.0	98.0	100.0	98.7	100.0	98.7
2	98.0	96.2	98.0	97.1	98.0	97.3
3	98.0	93.1	98.0	95.3	98.0	96.2
4	90.0	97.8	90.0	98.2	90.0	98.2
5	84.0	98.0	84.0	98.2	84.0	98.7
6	94.0	96.7	94.0	97.1	94.0	97.8
7	88.0	96.7	88.0	97.3	88.0	97.8
8	90.0	90.9	90.0	94.7	90.0	95.3
9	88.0	98.9	88.0	98.9	88.0	98.9
10	86.0	98.2	86.0	98.0	86.0	98.2
11	92.0	99.1	92.0	99.1	92.0	99.1
12	90.0	99.3	90.0	99.3	90.0	99.3
13	100.0	98.0	100.0	98.0	100.0	98.2
14	98.0	98.0	98.0	98.4	98.0	98.2
15	60.0	98.7	60.0	98.7	60.0	98.9
16	82.0	97.1	82.0	97.3	82.0	98.0
17	96.0	95.3	96.0	96.9	96.0	97.3
18	94.0	94.9	94.0	96.7	94.0	97.3
19	92.0	98.9	92.0	98.9	92.0	99.1
20	98.0	94.4	98.0	96.2	98.0	96.7
Mean (95% CI)	90.9 (86.7, 95.1)	96.9 (95.9, 97.9)	90.9 (86.7, 95.1)	97.7 (97.1, 98.2)	90.9 (86.7, 95.1)	97.9 (97.5, 98.5)
Δ mean vs. CDS (95% CI)	−6.1 (−10.7, −1.5)	1.9 (−0.3, 4.1)	−6.1 (−10.7, −1.5)	2.7 (0.5, 4.7)	−6.1 (−10.7, −1.5)	2.9 (0.9, 5.0)

Abbreviations: CI, confidence interval; CDS, conventional double screen.

Table 5. Time required to screen all citations according to screening method and TM tool

	Mean time required to review all citations (95% CI), minutes			
Screening method	**No TM**	**Rayyan**	**Abstrackr**	**SWIFT-review**
CDS	433 (327, 539)	-	-	-
SS	216 (169, 264)	-	-	-
DSTM	-	329 (248, 410)	113 (85, 140)	97 (73, 121)
CDSSTM	-	377 (288, 466)	260 (201, 319)	252 (195, 309)
SSTM	-	165 (129, 200)	56 (44, 69)	37 (29, 45)

Abbreviations: CDSSTM, combined double and single screen with text mining; CDS, conventional double screen; CI, confidence interval; DSTM, double screen with text mining; SS, single screen; SSTM, single screen with text mining; TM, text mining.

The mean time required to review all citations for each screening method was reconstructed from data from the 20 individual reviewers. These calculations assume that each citation took the same amount of time to be reviewed by one same reviewer. By multiplying the original time data by the proportion of citations identified as relevant by TM divided by the total number of citations needed to be screened without TM (i.e., 500), the theoretical time required for each screening method using TM was reconstructed. For example, to calculate the time required to complete CDSSTM with Rayyan, we added the time it took for reviewer 1 to complete their screening (with 500 citations) to the time it took for reviewer 2 to complete their screening with Rayyan (i.e., only reviewing 380 citations instead of 500). We did the same for all other reviewer duos, calculated the mean time across the 10 reviewer duos and provided a 95% confidence interval around the mean. When compared to CDS, TM methods minimally reduce screening time by 56 minutes (DSTM + Rayyan) and maximally reduce screening time by 396 minutes (SSTM + SWIFT-Review). This corresponds to a reduction in the screening time by 12.9–91.5% in comparison with CDS.

values of 91% and 72%, respectively. The reference standard used to calculate these sensitivity and specificity estimates, however, was the evaluation of another reviewer without TM. The actual proportion of relevant citations missed was 0% when considering the citations that were included in the final review [11]. Only one prior study has compared the performance and efficiency of a TM-enhanced method (SSTM) to CDS [5]. This 2016 study showed that TM decreased the time and costs of screening titles and abstracts, without compromising performance. Reported sensitivities were, however, generally very high, even when using the generally insensitive SS method. This may have occurred because the study reviewers were screening citations that they were made familiar with while carrying out an already published SR [19]. The level of experience of the reviewers was unreported, and measures of estimate precision were lacking. Each method was likely carried out by only one reviewer. Finally, the investigators reported only one approach to using TM and did not report the exact TM tool used.

Our study has much strength. We analyzed data from 20 reviewers, taking into consideration their expertise levels, and thoroughly detailing the TM methods used. We analyzed several TM tools and different ways each could be incorporated into the screening process. The reference standard used in our study was the results of an already published SR of observational studies, which had undergone peer review and was conducted by an entirely different set of reviewers. None of the reviewers in our study were aware of the results of this prior SR, and none were subject matter experts. Most recently published non-Cochrane SRs investigated observational studies [20].

Our study, however, also features certain limitations. First, the design of our study presupposed the knowledge of the final citations that would be included in an SR and used these citations as a reference standard. In practice, TM tools would learn from reviewers who are potentially making errors and the first screen would likely be overly inclusive. Second, our screening methods except for SS were retrospectively constructed using the screening results from 20 reviewers. As such, reviewers did not actually use the TM tools themselves during the screening process. Third, our test set included 500 citations. Although the generalizability of our results should be unaffected by this sample size, more precise estimates of sensitivity and specificity could have been achieved with a larger dataset. Fourth, we constructed our training set and test set to include 7.5% and 10% relevant citations, respectively. These are higher percentages than those seen with many SRs. However, in practice, training sets with a similar ratio of relevant-to-irrelevant studies may be constructed by oversampling relevant studies. For instance, before starting an SR, investigators often already have a list of pivotal studies that should be ultimately included in their SR. Including these pivotal studies in the training set (i.e., oversampling relevant studies) may help balance the training set. Further balancing the training set was not performed as it would have been unrepresentative of the type of data typically encountered in SRs. Artificially balancing the test set was not done for the same reason. Fifth, this generalizability must be interpreted in the context that we only evaluated one SR. Other SRs may necessitate larger, or smaller, training sets. Thus, although Rayyan, Abstrackr, and SWIFT-Review all performed well in this study with a training set of 200 citations, they may not perform as well with another citation dataset and with a different research question. We encourage researchers to replicate our findings using other datasets.

This study paints an optimistic picture of the use of TM in SRs, with human-supervised TM-based screening methods generating sensitivity and specificity estimates that are not different (specificity at times modestly

improved) in comparison to CDS. By evaluating the screening results of 20 reviewers, considering their level of expertise and combining their results into five screening methods, we performed a rigorous comparative study on the sensitivity and specificity of alternative methods for title and abstract screening. Our findings suggest that human-supervised TM tools may be an option for ''rapid reviews'', simpler SRs, or updates to previously published SRs and support the cautious integration of TM in SR on a study-by-study basis.

CRediT authorship contribution statement

Jimmy Li: Methodology, Formal analysis, Writing − original draft, Writing − review & editing, Visualization. **Joudy Kabouji:** Investigation, Methodology, Formal analysis, Writing − original draft, Writing − review & editing. **Sarah Bouhadoun:** Investigation, Methodology, Formal analysis, Writing − original draft, Writing − review & editing. **Sarah Tanveer:** Investigation, Methodology, Formal analysis, Writing − original draft, Writing − review & editing. **Kristian B. Filion:** Methodology, Investigation, Writing − review & editing. **Genevieve Gore:** Methodology, Investigation, Writing − review & editing. **Colin Bruce Josephson:** Investigation, Writing − review & editing. **Churl-Su Kwon:** Investigation, Writing − review & editing. **Nathalie Jette:** Investigation, Writing − review & editing. **Prisca Rachel Bauer:** Investigation, Writing − review & editing. **Gregory S. Day:** Investigation, Writing − review & editing. **Ann Subota:** Investigation, Writing − review & editing. **Jodie I. Roberts:** Investigation, Writing − review & editing. **Sara Lukmanji:** Investigation, Writing − review & editing. **Khara Sauro:** Investigation, Writing − review & editing. **Adnane Alaoui Ismaili:** Investigation, Writing − review & editing. **Feriel Rahmani:** Investigation, Writing − review & editing. **Khadidja Chelabi:** Investigation, Writing − review & editing. **Yasmine Kerdougli:** Investigation, Writing − review & editing. **Nour Meryem Seulami:** Investigation, Writing − review & editing. **Aminata Soumana:** Investigation, Writing − review & editing. **Sarah Khalil:** Investigation, Writing − review & editing. **Noémie Maynard:** Investigation, Writing − review & editing. **Mark Robert Keezer:** Conceptualization, Methodology, Writing − original draft, Writing − review & editing, Supervision, Project administration.

Declaration of competing interest

J.K., S.T., G.G., P.R.B., A Subota, S.L., K.S., A.A.I., F.R., K.C., Y.K., N.M.S., A Soumana, S.K., and N.M. report no conflicts of interest. J.L. is supported by the Fonds de Recherche Québec−Santé. K.B.F. is supported by a Senior salary support award from the Fonds de Recherche Québec−Santé and a William Dawson Scholar award from McGill University. He holds research grants from the Canadian Institutes of Health Research and has received honoraria from Quebec's Institut national d'excellence en santé et services sociaux and a stipend from the Canadian Network for Observational Drug Effect Studies. N.J. received grant funding paid to her institution for grants unrelated to this work from NINDS, NIH U24NS107201, NIH IU54NS100064, and NIH U24NS113849) during the study period and was the Bludhorn Professor of International Medicine at the Icahn School of Medicine at Mount Sinai. She receives an honorarium for her work as an Associate Editor of Epilepsia. GS Day's research is supported by NIH (K23AG064029, U01AG057195, U19AG032438), the Alzheimer's Association, and Chan Zuckerberg Initiative. He serves as a consultant for Parabon Nanolabs Inc., as a Topic Editor (Dementia) for DynaMed (EBSCO), and as the Clinical Director of the Anti-NMDA Receptor Encephalitis Foundation (Inc., Canada; uncompensated). He is the co-Project PI for a clinical trial in anti-NMDAR encephalitis, which receives support from Horizon Pharmaceuticals. He has developed educational materials for PeerView Media, Inc., and Continuing Education Inc. He owns stock in ANI pharmaceuticals. Dr. Day's institution has received support from Eli Lilly for Dr. Day's development and participation in an educational event promoting early diagnosis of symptomatic Alzheimer's disease. J.I.R. reports fellowship salary support from Canadian Network of MS Clinics and travel support from the Rebecca Hotchkiss International Scholar Exchange. M.R.K. reports unrestricted educational grants from UCB, Eisai, and Jazz Pharmaceuticals, research grants for investigator-initiated studies from UCB and Eisai as well as from government entities (Canadian Institutes of Health Research, Fonds de Recherche Québec−Santé), academic institutions (Center Hospitalier de l'Université de Montréal), and foundations (TD Bank, TSC Alliance, Savoy Foundation, Quebec Bio-Imaging Network). M.R.K.'s salary is supported by the Fonds de Recherche Québec−Santé.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2023.07.010.

References

[1] Chandler JCM, Thomas J, Higgins JPT, Deeks JJ, Clarke MJ. Chapter I: introduction. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editors. ochrane Handbook for Systematic Reviews of Interventions version 6.4 (updated August 2023). Cochrane; 2023. Available at www.training.cochrane.org/handbook. Accessed September 6, 2023.

[2] Cohen AM, Hersh WR, Peterson K, Yen P-Y. Reducing workload in systematic review preparation using automated citation classification. J Am Med Inf Assoc 2006;13:206−19.

[3] Polanin JR, Pigott TD, Espelage DL, Grotpeter JK. Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses. Res Synth Methods 2019;10:330−42.
[4] McKenzie JEBS, Ryan RE, Thomson HJ, Johnston RV, Thomas J. Chapter 3: defining the criteria for including studies and how they will be grouped for the synthesis. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions version 6.4 (updated August 2023). Cochrane; 2023. Available at www.training.cochrane.org/handbook. Accessed September 6, 2023.
[5] Shemilt I, Khan N, Park S, Thomas J. Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. Syst Rev 2016;5:140.
[6] Fan W, Wallace L, Rich S, Zhang Z. Tapping the power of text mining. Commun ACM 2006;49:76−82.
[7] European Centre for Disease Prevention and Control. Use and impact of new technologies for evidence synthesis. ECDC; 2022. Available at https://www.ecdc.europa.eu/en/publications-data/use-and-impact-new-technologies-evidence. Accessed September 6, 2023.
[8] Khalil H, Ameen D, Zarnegar A. Tools to support the automation of systematic reviews: a scoping review. J Clin Epidemiol 2022;144:22−42.
[9] Grandi SM, Filion KB, Yoon S, Ayele HT, Doyle CM, Hutcheon JA, et al. Cardiovascular disease-related morbidity and mortality in women with a history of pregnancy complications: systematic review and meta-analysis. Circulation 2019;139:1069−79.
[10] Gates A, Guitard S, Pillay J, Elliott SA, Dyson MP, Newton AS, et al. Performance and usability of machine learning for screening in systematic reviews: a comparative evaluation of three tools. Syst Rev 2019;8:278.
[11] Carey N, Harte M, Mc Cullagh L. A text-mining tool generated title-abstract screening workload savings: performance evaluation versus single-human screening. J Clin Epidemiol 2022;149:53−9.
[12] Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016;5:210.
[13] Wallace BC, Small K, Brodley CE, Lau J, Trikalinos TA. Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. Proceedings of the 2nd ACM SIGHIT international health informatics symposium; 2012;819−24. https://doi.org/10.1145/2110363.2110464.
[14] Howard BE, Phillips J, Miller K, Tandon A, Mav D, Shah MR, et al. SWIFT-review: a text-mining workbench for systematic review. Syst Rev 2016;5:87.
[15] Valizadeh A, Moassefi M, Nakhostin-Ansari A, Hosseini Asl SH, Saghab Torbati M, Aghajani R, et al. Abstract screening using the automated tool Rayyan: results of effectiveness in three diagnostic test accuracy systematic reviews. BMC Med Res Methodol 2022;22:160.
[16] R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2022.
[17] 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.
[18] O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Syst Rev 2015;4:5.
[19] Park S, Khan NF, Hampshire M, Knox R, Malpass A, Thomas J, et al. A BEME systematic review of UK undergraduate medical education in the general practice setting: BEME guide no. 32. Med Teach 2015;37:611−30.
[20] Hoffmann F, Allers K, Rombey T, Helbach J, Hoffmann A, Mathes T, et al. Nearly 80 systematic reviews were published each day: observational study on trends in epidemiology and reporting over the years 2000-2019. J Clin Epidemiol 2021;138:1−11.