Journal of Clinical Epidemiology 149 (2022) 53–59

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

# A text-mining tool generated title-abstract screening workload savings: performance evaluation versus single-human screening

Niamh Carey*, Marie Harte, Laura Mc Cullagh

*National Centre for Pharmacoeconomics, Old Stone Building, Trinity Centre for Health Sciences, St James's Hospital, Dublin 8, Ireland*
*Department of Pharmacology and Therapeutics, Trinity Centre for Health Sciences, St James's Hospital, Dublin 8, Ireland*

Accepted 24 May 2022; Published online 30 May 2022

## Abstract

**Background and Objectives:** Text-mining tool, Abstrackr, may potentially reduce the workload burden of title and abstract screening (Stage 1), using screening prioritization and truncation. This study aimed to evaluate the performance of Abstrackr's text-mining functions ('Abstrackr-assisted screening'; screening undertaken by a single-human screener and Abstrackr) vs. Single-human screening.

**Methods:** A systematic review of treatments for relapsed/refractory diffuse large B cell lymphoma ($n$ = 7,723) was used. Citations, uploaded to Abstrackr, were screened by a human screener until a pre-specified maximum prediction score of 0.39540 was reached. Abstrackr's predictions were compared with the judgments of a second, human screener (who screened all citations in Covidence). The performance metrics were sensitivity, specificity, precision, false negative rate, proportion of relevant citations missed, workload savings, and time savings.

**Results:** Abstrackr reduced Stage 1 workload by 67% (5.4 days), when compared with Single-human screening. Sensitivity was high (91%). The false negative rate at Stage 1 was 9%; however, none of those citations were included following full-text screening. The high proportion of false positives ($n$ = 2,001) resulted in low specificity (72%) and precision (15.5%).

**Conclusion:** Abstrackr-assisted screening provided Stage 1 workload savings that did not come at the expense of omitting relevant citations. However, Abstrackr overestimated citation relevance, which may have negative workload implications at full-text screening. 

*Keywords:* Machine learning; Text-mining; Screening; Systematic review; Methodology; Lymphoma

## 1. Introduction

### 1.1. Challenges in title and abstract screening

The screening process of a systematic literature review (SLR) is generally conducted by two or more independent human screeners [1]. Title and abstract screening (Stage 1) usually involves screening thousands of citations for relevance. Emphasis is placed on identifying all relevant citations (i.e., attaining 100% sensitivity), to minimize bias [2]. SLRs require input from highly skilled researchers and a large time commitment. Time commitments, described elsewhere, have ranged from 6 months to 2 years [3–5].

SLR complexity has increased due to a growth in the volume of research published, the use of complex methodologies such as network meta-analysis, and the increasing complexity of new interventions [4,6]. Guidance generally requires that SLRs must not be finalized until 6 to 12 months before dissemination [1]. Researchers are increasingly faced with the challenge of producing a robust SLR within the confines of time and budget.

### 1.2. Text-mining

These challenges have resulted in the recognized need to develop alternative methods [4]. Semi-automating Stage 1, using text mining, has been proposed as one solution [7]. Text mining is the process of discovering knowledge and structure from unstructured data (i.e., text) [8,9]. Relevant

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations of interest: None declared by any of the authors.

* Corresponding author. National Centre for Pharmacoeconomics, Old Stone Building, Trinity Centre for Health Sciences, St James's Hospital, Dublin 8, Ireland. Tel.: +353-1-4103427.

*E-mail address:* nicarey@tcd.ie (N. Carey).

https://doi.org/10.1016/j.jclinepi.2022.05.017

**What is new?**

**Key Findings**

- Abstrackr-assisted screening resulted in title and abstract screening workload savings when used to reduce the number of citations to be screened in a systematic literature review of treatments for relapsed/refractory diffuse large B cell lymphoma.
- These workload savings did not come at the expense of omitting relevant citations.

**What this adds to what is known?**

- The performance of Abstrackr's text-mining functions has varied in the literature. This study adds to the evidence base by using a complex systematic literature review, with a diverse range of interventions and study types. Inclusion and exclusion criteria had lexical similarity, adding a further dimension of complexity.
- Workload savings from the perspective of a national decision-making agency have been estimated.

**What is the implication, what can we change now?**

- The evidence base of text-mining tools is limited. Further research is required to assess the performance of Abstrackr's text-mining functions at higher maximum prediction scores (i.e., earlier stopping points). Generalizability of the results presented here may be limited. Thus, further research is also required to determine if Abstrackr-assisted screening can be considered in other research questions and disease areas.

information is identified as patterns, learnt from an initial set of training sample citations, which are labelled as relevant or irrelevant by a human screener [6]. The accuracy of the predictions improves through interaction with the human screener [7].

Text-mining approaches, such as screening prioritization and truncation, may reduce Stage 1 workload burden in a number of ways [7,10]. Screening prioritization identifies and presents the most relevant citations for screening first. This may allow members of the review team to begin full-text screening (Stage 2) earlier, reducing the time taken from SLR commencement to completion. Through this approach, human screeners may become more familiar with the SLR inclusion criteria earlier in the process, ultimately increasing efficiency. This may also address over-inclusiveness, whereby, human screeners tend to be cautious and include more citations early in Stage 1 [7]. Screening truncation (where citations that fall beneath a specified prediction score of relevance are excluded) may reduce the number of citations to be screened at Stage 1 [7,10]. It has been suggested that text-mining tools might supplement the work traditionally undertaken by a second, human screener [11].

### *1.3. Text-mining tool: Abstrackr*

Several text-mining tools have been developed to facilitate the semi-automation of Stage 1 [12,13]. For this analysis, Abstrackr was the text-mining tool of choice due to its widespread use in the literature [3,6,11]. It has been shown to perform favorably relative to other tools [14]. Abstrackr is a free, online machine learning tool that incorporates both screening prioritization and truncation functions [6,15]. The first step in the screening process, using Abstrackr (herein 'Abstrackr-assisted screening'), involves uploading the relevant citations. Then, the human screener labels (i.e., relevant or irrelevant) the training sample of citations. The ability of Abstrackr to accurately predict citation relevance depends on the correct labelling of the training sample [16]. Abstrackr's prediction algorithm updates once per day; Abstrackr processes information gained through the labelled training sample. Once processed, Abstrackr generates both 'hard' predictions (i.e., include or exclude), and a prediction score (between 0 and 1), for each remaining citation. A maximum prediction score of the remaining citations is also presented. The human screener may then choose to continue screening in Abstrackr and thus, improve Abstrackr's learning capacity to generate an updated list of predictions. The existing literature suggests that once the maximum prediction score of the remaining citations falls below 0.40, zero citations are generally predicted to be relevant by Abstrackr [17,18]. At this point, cessation of the human screening of titles and abstracts may be considered.

The performance of Abstrackr's text-mining functions, as measured by a variety of metrics (Table 1), has varied [3,6,10]. It has been recommended that further research is required to assess performance on a diverse range of screening tasks [3,10].

## 2. Aim of study

This study aims to evaluate the performance of Abstrackr's text-mining functions, when compared to Single-human screening, in an SLR of treatments for relapsed/refractory diffuse large B cell lymphoma (R/R DLBCL). The SLR research question was 'what is the efficacy of CD19 CAR T cell therapies (tisagenlecleucel and axicabtagene ciloleucel) vs. salvage chemotherapy in patients with R/R DLBCL after two or more lines of systemic therapy?'. We aim to investigate the reliability of Abstrackr's predictions once a maximum prediction score of 0.39540 (base case) is reached. The performance metrics described by Gates et al. [3] and Rathbone et al. [6] will be investigated here.

**Table 1.** Definition of performance metrics used to assess the performance of text-mining functions in Abstrackr, adapted from Gates et al. [3] and Rathbone et al. [6]

| Performance metric | Definition |
|---|---|
| Sensitivity (true positive rate) | The proportion of citations correctly predicted as relevant by Abstrackr out of the total deemed relevant by the human screener [19] |
| Specificity (true negative rate) | The proportion of citations correctly predicted as irrelevant by Abstrackr out of the total deemed irrelevant by the human screener [19] |
| Precision | The proportion of citations correctly predicted as relevant by Abstrackr amongst all citations predicted as relevant by Abstrackr (both correct and incorrect) |
| False negative rate | The proportion of citations that were incorrectly predicted as irrelevant by Abstrackr, out of the total number of citations deemed relevant by the human screener |
| Proportion missed | The proportion of citations incorrectly predicted as irrelevant by Abstrackr that were included in the final evidence base, out of the total number of citations included in the final evidence base[a] |
| Workload savings | The proportion of citations predicted as irrelevant by Abstrackr out of the total number of citations to be screened (i.e., the proportion of citations that would not need to be screened manually) |
| Time savings | Time saved based on the citations that would not need to be screened (i.e., those predicted as irrelevant by Abstrackr); estimated based on a screening rate of 0.5 min per citation and an 8-hour work day |

[a] Definition adapted by the authors.

## 3. Methods

### 3.1. Choice of data set

An SLR of treatments for R/R DLBCL was chosen due to the large number of citations identified through database searching. Text-mining tools have been reported to perform better on large (≥2,500) screening samples [10]. The large number of treatments specified in the inclusion criteria was also a consideration. The SLR inclusion and exclusion criteria are presented in Appendix A.

### 3.2. Search methods

Electronic databases EMBASE, MEDLINE (via EBSCO), and CENTRAL (via the Cochrane Library) were searched from 01 January 2001 to 25 October 2019.

### 3.3. Citation management

Identified citations were imported to Endnote®. Duplicates were systematically searched for using software in Endnote® and identified manually. Following exclusion of duplicates, 7,723 citations were included in Stage 1. Screening was conducted by two human screeners, both experienced in producing SLRs. Screener 1 undertook the process using Abstrackr ('Abstrackr-assisted screening'), whilst Screener 2 undertook the process using Covidence ('Single-human screening').

### 3.4. Abstrackr-assisted screening

Screener 1 uploaded citations to Abstrackr. Screener 1 then screened an initial training sample of 200 randomly selected citations. This is in line with previously reported training sample sizes [11,14]. The algorithm was then allowed to process the information (from the training sample) overnight.

Once this information was processed and the initial predictions were generated by Abstrackr, Screener 1 set the screening settings to 'single-screen mode'. The order of citations was set to 'most likely to be relevant,' so that the most relevant citations, as predicted by Abstrackr, were presented to the human screener in priority order. Screener 1 screened titles and abstracts for relevance. Screener 1 continued to screen in Abstrackr until the algorithm indicated that a maximum prediction score of 0.39540 (base case) was reached [17]. Of note, stopping once a maximum prediction score of less than 0.40 was reached was prespecified; however, due to the time required for Abstrackr's algorithm to update (i.e., overnight), the maximum prediction score could not be measured in real time. This resulted in screening until a maximum prediction score of 0.39540 was reached. At this point, Screener 1 assumed that any remaining unscreened citations were irrelevant and did not conduct any further screening in Abstrackr. This inherently assumes that any unscreened citations at this point have been 'screened' and deemed irrelevant by Abstrackr (i.e., Abstrackr is acting as the second, human screener for these citations). Citations that were deemed 'relevant' or 'maybe' by Screener 1, during Stage 1, were brought forward for Stage 2. Citations, and their associated labels, were exported from Abstrackr to Microsoft Excel®. It was assumed that Abstrackr deemed all citations with a score greater than 0.39540 as relevant, despite the label provided by Screener 1.

#### 3.4.1. Sensitivity analysis

Sensitivity analysis was conducted, whereby Abstrackr-assisted screening continued until pre-specified maximum predication scores of 0.35 and 0.30 (corresponding to realised scores of 0.34458 and 0.29021, respectively) were reached. The aim here was to determine if the trade-off between workload saving and accuracy of Abstrackr could be improved at alternative prediction scores.

### 3.5. Single-human screening: Covidence

Screener 2 uploaded citations to Covidence and screened all citations (title and abstract). Citations deemed 'relevant' or 'maybe' by Screener 2 were brought forward for Stage 2. Citations, and their associated labels, were exported from Covidence to Microsoft Excel®.

### 3.6. Data analysis to assess the performance of Abstrackr's text-mining functions

Data from 2 × 2 cross-tabulations, based on the number of citations predicted relevant or irrelevant by Abstrackr-assisted screening vs. the number deemed relevant or irrelevant by Single-human screening, were used to calculate performance metrics. The formulas used to calculate these metrics and associated 2 × 2 cross-tabulations are presented in Appendix B. Here, it is assumed that Single-human screening identified all relevant citations.

## 4. Results

Of the 7,723 citations, 2,568 (33%) citations (titles and abstracts, including training sample) were screened in Abstrackr before a maximum prediction score of 0.39540 (base case) was reached. In line with previous studies [17,18], zero remaining citations were predicted to be relevant by Abstrackr. Of these 2,568 citations, 451 were brought forward for Stage 2.

Single-human screening (by Screener 2) of all citations on Covidence resulted in 424 citations being brought forward for Stage 2.

Citations that were deemed relevant by one human screener (i.e., Screener 1 using Abstrackr-assisted screening) but deemed irrelevant by the other human screener (i.e., Screener 2 using Covidence) were brought forward for Stage 2 [14]. Six citations were included in the final evidence base.

### 4.1. Performance of Abstrackr's text-mining functions

Data from the 2 × 2 cross-tabulations, used to calculate the performance metrics of Abstrackr-assisted screening (vs. Single-human screening), are presented in Table 2.

The performance metrics, based on these data (base case), are presented in Table 3. All metrics presented relate to Stage 1. Estimated workload and time savings (Stage 1) do not account for the subsequent (Stage 2) workload burden associated with the high number of false positives ($n$ = 2,001). Further discussion is provided in 5.1 .

#### 4.1.1. Sensitivity analysis

The performance metrics, based on sensitivity analysis, are presented in Table 3. An additional 584 and 1,284 citations required screening (compared to the base case) before reaching maximum prediction scores of 0.34458 and 0.29021, respectively. The 2 × 2 cross-tabulations used to calculate these metrics are presented in Appendix B.

An additional sensitivity analysis was conducted to explore the impact, on time savings, of assuming a higher screening rate of 1 minute per citation [20]. Under this assumption, the Stage 1 time savings were 10.7 days (0.39540 prediction score, base case), 9.5 days (0.34458 prediction score), and 8.1 days (0.29021 prediction score).

## 5. Discussion

### 5.1. Main findings

In the research question specified here, Abstrackr-assisted screening, conducted until a maximum prediction score of 0.39540 (base case) was reached, identified all relevant citations and reduced Stage 1 workload by 67% (5.4 days) when compared with Single-human screening (using Covidence). Abstrackr demonstrated high sensitivity (91%). Although the false negative rate was 9%, the actual proportion of relevant citations missed was 0%. No citations that were indicated as irrelevant by Abstrackr, but relevant by Screener 2 (using Covidence), were included in the final evidence base. Specificity (72%) and precision (15.5%) were low; Abstrackr overestimated citation relevance.

Sensitivity analysis, conducted at maximum prediction scores of 0.34458 and 0.29021, resulted in higher sensitivity and lower false negative rates (compared to the base case). However, these improvements came at the expense of decreased specificity and precision, and reduced workload savings. At a maximum prediction score of 0.29021, just one citation was predicted to be irrelevant by Abstrackr that was judged to be relevant by Screener 2. Those producing SLRs may be willing to make the trade-off between increased sensitivity and reduced workload saving. Notably, the proportion missed was zero in both the base case and sensitivity analysis. The results of this sensitivity analysis may stimulate further discussion on what the most appropriate stopping point should be and provides an insight into the trade-offs required to improve sensitivity.

Of importance, in this study, only citations deemed relevant by Screener 1 (using Abstrackr, $n$ = 451) were brought forward for Stage 2. Thus, mitigating against the negative impact of the high number of predicted false positives ($n$ = 2,001), and generating time savings. Performance

**Table 2.** 2 × 2 cross tabulation of Abstrackr predictions, at a maximum prediction score of 0.39540, vs. human-screener (Screener 2) judgments

| | | Human screener (Screener 2) judgments | | |
|---|---|---|---|---|
| | | Excl. | Incl. | Total |
| Abstrackr Predictions | Excl. | 5,118[a] (True Negative) | 37[b] (False Negative) | 5,155 |
| | Incl. | 2,001[c] (False Positive) | 367[d] (True Positive) | 2,368 |
| | Total | 7,119 | 404 | 7,523[e] |

[a] Abstrackr and Screener 2 excluded the same 5,118 citations; the number of true negatives predicted by Abstrackr.
[b] Abstrackr excluded 37 citations that Screener 2 included; the number of false negatives predicted by Abstrackr.
[c] Abstrackr included 2,001 citations that Screener 2 excluded; the number of false positives predicted by Abstrackr.
[d] Abstrackr and Screener 2 included the same 367 citations; the number of true positives predicted by Abstrackr.
[e] The total number of citations included in the analysis, excluding the 200 citation training sample.

metrics in this analysis are based solely on Abstrackr's predictions (thus, ignoring the labels provided by Screener 1). Relying solely on Abstrackr's predictions, the high number of false positives would add to workload burden at Stage 2. Assuming it takes 4 minutes to retrieve a full text, and 5 minutes for full-text screening, full-text screening of these false positives would require 37.5 days [20]. This outweighs workload savings generated at Stage 1. This analysis assumed that all citations screened before reaching the predefined maximum prediction score were deemed relevant by Abstrackr. This may overestimate the number of citations predicted to be relevant by Abstrackr and may partly contribute to the high number of predicted false positives.

This study contributes to the limited evidence base on the performance of Abstrackr's text-mining functions. In contrast to other studies, whereby screening in Abstrackr was conducted until the first set of predictions were available [3,6], this study continued screening until a predefined maximum prediction score was reached. The advantage here is that Abstrackr's learning capacity is expected to improve, based on the increased data (and therefore 'learning') provided by the human screener.

The SLR inclusion criteria contained a number of treatments with lexical similarity, which may partly contribute to the low precision. Also, there was a high level of imbalance between relevant and irrelevant citations; just 17.5% of screened citations (equivalent to 6% of all citations) in the base case were included for Stage 2. In such instances, the predictions are biased towards the majority irrelevant citations [21], which produces falsely weighted predictions (i.e., irrelevant citations) [6]. These issues reflect those which have been encountered elsewhere [6].

Previous studies found that SLRs that have more complex PICOS (population; intervention; comparator; outcome; study design) criteria tend to achieve less magnitude of workload savings (as defined in this study) [6]. Also, it has been

**Table 3.** Performance metrics of Abstrackr-assisted screening, when compared to Single-human screening at a maximum prediction score of 0.39540 (base case), 0.34458 and 0.29021 (sensitivity analyses) in the systematic literature review of treatments for R/R DLBCL ($n$ = 7,523)[h,i]

| Performance metric | Sensitivity[a] (%) | Specificity[b] (%) | Precision[c] (%) | False negative rate[d] (%) | Proportion missed[e] (%) | Workload savings[f] (%) | Time savings[g] (d) |
|---|---|---|---|---|---|---|---|
| Result (prediction score 0.39540) | 91 | 72 | 15.5 | 9 | 0 | 67 | 5.4 |
| Sensitivity Analysis | | | | | | | |
| Result (prediction score 0.34458) | 97 | 64 | 13 | 3 | 0 | 59 | 4.8 |
| Result (prediction score 0.29021) | 100 | 54 | 11 | 0 | 0 | 50 | 4.0 |

DLBCL: Diffuse large B-cell lymphoma; R/R: Relapsed/refractory.
[a] Proportion of citations correctly predicted as relevant by Abstrackr out of the total deemed relevant by Screener 2.
[b] Proportion of citations correctly predicted as irrelevant by Abstrackr out of the total deemed irrelevant by Screener 2.
[c] Proportion of citations correctly predicted as relevant by Abstrackr amongst all citations predicted as relevant by Abstrackr (both correct and incorrect).
[d] Proportion of citations incorrectly predicted as irrelevant by Abstrackr out of the total deemed relevant by Screener 2.
[e] Proportion of citations incorrectly predicted as irrelevant by Abstrackr that were included in the final evidence base, out of the total number of citations included in the final evidence base.
[f] Proportion of citations predicted as irrelevant by Abstrackr out of the total number of citations to be screened, including the training set (i.e., the proportion of citations that would not need to be screened manually).
[g] Time saved based on the citations that would not need to be screened (i.e., those predicted as irrelevant by Abstrackr); based on a screening rate of 0.5 min per citation and an 8-hour work day.
[h] Calculations presented in Appendix B.
[i] Excludes 200 citations included in the training sample.

suggested that text-mining tools may perform better for SLRs that only include randomized controlled trials [14]. In this study, despite the complexity of comparators (due to lexical similarity) and the inclusion of single-arm and observational studies, the workload savings were notable.

In this study, a single research question was presented. It has been previously noted that Abstrackr's predictions were more reliable when fewer research questions were defined [14]. In instances where a greater number of research questions are defined, the algorithm may find it more challenging to discern patterns during the training phase. To enhance pattern learning, a larger training sample may be required. However, this may be impractical and may negatively impact workload savings.

Although, as standard, emphasis is placed on attaining 100% sensitivity in SLRs, it seems unlikely that a single-human screener would consistently attain this. An analysis of 280 single-human screeners observed a sensitivity of 86.6% in this cohort, based on 24,942 screening decisions and 2,000 abstracts [22]. Dual-human screening in the same analysis attained 97.5% sensitivity. Specificity was 79.2% and 68.7% for single-human screening and dual-human screening, respectively [22].

### 5.2. Limitations

The topic-dependent nature of these results is highlighted; findings are based on a single SLR in one disease area. Results are also likely impacted by the screening sample size, experience, and topic expertise of the human screeners. This limits the generalizability of results. Further research is warranted to investigate if results can be replicated for other research questions and disease areas.

This study assumes that Single-human screening predicts all relevant citations with 100% accuracy. However, as described in 5.1, this may not be the case. In this study, both screeners were highly experienced. However, to limit any uncertainties associated with this assumption, Screener 2 (Single-human screening) was the more experienced individual. Ideally, however, performance would have been compared to a validated, pre-screened database of citations. Pre-screening of this database would be conducted by two human screeners, in line with the gold standard approach.

The performance of Abstrackr's text-mining functions was determined based on the behaviour of Screener 1 and how well Abstrackr agreed with the judgments of Screener 2. However, the judgments of Screener 1 and Screener 2 were not perfectly aligned. The performance of Abstrackr's text-mining functions (given Screener 1's behaviour) and inter-rater reliability, in this study, are conflated. As such, the positive performance of Abstrackr's text-mining functions may be underestimated. Consideration may have been given to training and screening in Abstrackr by both Screener 1 and Screener 2. Under this approach, Abstrackr could gain insight from both screeners.

The performance of Abstrackr's text-mining functions, based on higher maximum prediction scores, was not evaluated here. It is plausible that higher maximum prediction scores (i.e., earlier stopping points) may result in further workload savings without missing relevant citations. It is important that further research investigates this, by downloading predictions at pre-specified thresholds (i.e., every time Abstrackr updates). This, in combination with the sensitivity analysis conducted as part of this research, would provide a comprehensive overview of the impact of using alternative stopping points. This approach would also give a more realistic indication of the number of false positives predicted by Abstrackr.

The number of times Abstrackr updated, and produced an updated list of predictions, was not recorded during this research. Thus, the number of times Abstrackr had an opportunity to retrain cannot be determined. Downloading Abstrackr's predictions at pre-defined thresholds, and recording the number of times Abstrackr updated, would facilitate an analysis of the pattern of the performance of Abstrackr's text-mining functions throughout screening.

## 6. Conclusion

Using Abstrackr's text-mining functions, to cease Stage 1 screening at a maximum prediction score of 0.39540, generated workload savings that did not come at the expense of omitting relevant citations. However, the importance of conducting further research to investigate performance at stopping points defined by higher maximum prediction scores is emphasized. Sensitivity analysis at maximum prediction scores of 0.34458 and 0.29021 produced improved sensitivity but came at the expense of workload savings. Although Stage 1 workload and time savings were notable, the proportion of false positives was high. The associated workload burden of these false positives may have negative workload implications at Stage 2, if relying solely on Abstrackr's predictions. Given that best practice requires two human screeners, a second screener might consider use of Abstrackr to exclude citations below 0.39540, but rely on their own judgments before this threshold. However, further research is warranted before generalizing these results to different research questions.

## CRediT authorship contribution statement

**Niamh Carey:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing − original draft, Visualization, Project administration. **Marie Harte:** Conceptualization, Methodology, Software, Investigation, Writing − review & editing. **Laura Mc Cullagh:** Writing − review & editing, Supervision.

## Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2022.05.017.

## References


[1] Cochrane Handbook for Systematic Review of Interventions version 6.2. Chichester (UK): John Wiley & Sons; 2021.

[2] Tricco AC, Brehaut J, Chen MH, Moher D. Following 411 Cochrane protocols to completion: a retrospective cohort study. PLoS One 2008;3:e3684.

[3] Gates A, Johnson C, Hartling L. Technology-assisted title and abstract screening for systematic reviews: a retrospective evaluation of the Abstrackr machine learning tool. Syst Rev 2018;7(1):45.

[4] Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open 2017; 7(2):e012545.

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

[6] Rathbone J, Hoffmann T, Glasziou P. Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers. Syst Rev 2015;4:80.

[7] 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(1):5.

[8] Ananiadou S, Mc Naught J. Text mining for biology and biomedicine. Boston/London: Artech House; 2006.

[9] Hearst M. Untangling text data mining. 37th Annual Meeting of the Association for Computational Linguistics; 1999. https://dl.acm.org/doi/10.3115/1034678.1034679. Accessed June 10, 2022.

[10] Tsou AY, Treadwell JR, Erinoff E, Schoelles K. Machine learning for screening prioritization in systematic reviews: comparative performance of Abstrackr and EPPI-Reviewer. Syst Rev 2020;9(1):73.

[11] Gates A, Gates M, Sebastianski M, Guitard S, Elliott SA, Hartling L. The semi-automation of title and abstract screening: a retrospective exploration of ways to leverage Abstrackr's relevance predictions in systematic and rapid reviews. BMC Med Res Methodol 2020; 20:139.

[12] Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Syst Rev 2016;5(1):210.

[13] EPPI-Reviewer. Machine leaning functionality in EPPI-Reviewer 2021. Available at https://eppi.ioe.ac.uk/CMS/Portals/35/machine_learning_in_eppi-reviewer_v_7_web_version.pdf. Accessed July 12, 2021.

[14] 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(1):278.

[15] Wallace B, Small K, Brodley C, Lau J, Trikalinos T. Deploying an interactive machine learning system in an evience-based practice center: abstrackr. 2nd ACM SIGHIT International Health Informatics Symposium. New York: Association for Computing Machinery; 2012.

[16] Gates A, Gates M, DaRosa D, Elliott SA, Pillay J, Rahman S, et al. Decoding semi-automated title-abstract screening: findings from a convenience sample of reviews. Syst Rev 2020;9(1):272.

[17] Balke E. Abstrackr: online, open-access, free software for citation screening. KTDRR and Campbell collaboration research evidence training session. Austin, TX: Centre on Knowledge Translation for Disability & Rehabilitation Research; 2019.

[18] Reddy SM, Patel S, Weyrich M, Fenton J, Viswanathan M. Comparison of a traditional systematic review approach with review-of-reviews and semi-automation as strategies to update the evidence. Syst Rev 2020; 9(1):243.

[19] Altman D, Bland J. Diagnostic tests. 1: Sensitivity and specificity. British Medical Journal 1994;308(1552).

[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] Wu G, Chang E. KBA: Kernel boundary alignment considering imbalanced data distribution. IEEE Trans Knowl Data Eng 2005; 17:786−95.

[22] Gartlehner G, Affengruber L, Titscher V, Noel-Storr A, Dooley G, Ballarini N, et al. Single-reviewer abstract screening missed 13 percent of relevant studies: a crowd-based, randomized controlled trial. J Clin Epidemiol 2020;121:20−8.