Diagnostic performance of artificial intelligence tools for article screening during literature review: A systematic review

Ma. Sergia Fatima Sucaldito, Kaela Czarina Yu

Submitted to: JMIR Preprints
on: February 22, 2024

Table of Contents

Diagnostic performance of artificial intelligence tools for article screening during literature review: A systematic review

Ma. Sergia Fatima Sucaldito[1] MD; Kaela Czarina Yu[2] RPh

[1]Department of Medicine University of the Philippines Manila - Philippine General Hospital Manila PH
[2]None Manila PH

Corresponding Author:
Ma. Sergia Fatima Sucaldito MD
Department of Medicine
University of the Philippines Manila - Philippine General Hospital
Taft Avenue, Ermita
Manila
PH

Abstract

Background: The burgeoning volume of scientific literature being generated today places a great burden on evidence reviewers. On average, only 2% to 8% of articles yielded by a search strategy are ultimately included in a systematic review. Due to the burden of increasing information loads, there is a demand for methods that improve efficiency while maintaining accuracy in performing evidence reviews.

Objective: This systematic review aims to determine the accuracy and efficiency of AI-assisted abstract selection compared to manual abstract selection, as assessed by diagnostic performance and workload saved over sampling (WSS).

Methods: Two reviewers searched PubMed, Proquest, and Cochrane Library for studies evaluating the diagnostic performance and/or workload savings achieved by any AI tool, whether through full or semi-automation, in the title and abstract screening phase of literature review. Variance-weighted random effects meta-analysis was done to generate univariate measures of sensitivity, specificity, and WSS for the studies using RevMan verson 4.3 and the ‘meta’ and ‘mada’ packages on R version 4.3.1. Bivariate analysis was also performed for the measures of diagnostic accuracy and a hierarchal summary operating characteristics curve (HSROC) was generated.

Results: Twenty-two studies were included in this review, where 13 reported diagnostic performance, 14 reported WSS, and five studies reported both outcomes. In fully automated workflows, AI tools had a sensitivity of 85.6% (95% CI: 60.8%-95.8%) and a specificity of 88.7% (95% CI: 58.7%-97.7%) with considerable heterogeneity, which likely stems from the differences in the SRs and AI techniques used. In semi-automated workflows, sensitivity was 87.6% (95% CI: 77.2%-93.6%) and specificity was 94.1% (95% CI: 60.0%-99.4%) also with considerable heterogeneity. Among studies on full automation, the median workload savings for 100% recall was 50.0% (IQR: 10.2), while for studies on semi-automation, the median workload savings was 55.6% (IQR: 16.4).

Conclusions: Given the findings of this review, the diagnostic performance of AI tools appeared to be superior when used in semi-automated workflows rather than fully automated ones. This suggest that AI tools hold great potential in augmenting the accuracy and efficiency of human reviewers during study selection in literature review.

(JMIR Preprints 22/02/2024:57648)
DOI: https://doi.org/10.2196/preprints.57648

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Review

Diagnostic performance of artificial intelligence tools for article screening during literature review: A systematic review

Abstract

Background:
The burgeoning volume of scientific literature being generated today places a great burden on evidence reviewers. On average, only 2% to 8% of articles yielded by a search strategy are ultimately included in a systematic review. Due to the burden of increasing information loads, there is a demand for methods that improve efficiency while maintaining accuracy in performing evidence reviews.

Objective:
This systematic review aims to determine the accuracy and efficiency of AI-assisted abstract selection compared to manual abstract selection, as assessed by diagnostic performance and workload saved over sampling (WSS).

Methods:
Two reviewers searched PubMed, Proquest, and Cochrane Library for studies evaluating the diagnostic performance and/or workload savings achieved by any AI tool, whether through full or semi-automation, in the title and abstract screening phase of literature review. Variance-weighted random effects meta-analysis was done to generate univariate measures of sensitivity, specificity, and WSS for the studies using RevMan verson 4.3 and the ‘meta’ and ‘mada’ packages on R version 4.3.1. Bivariate analysis was also performed for the measures of diagnostic accuracy and a hierarchal summary operating characteristics curve (HSROC) was generated.

Results:
Twenty-two studies were included in this review, where 13 reported diagnostic performance, 14 reported WSS, and five studies reported both outcomes. In fully automated workflows, AI tools had a sensitivity of 85.6% (95% CI: 60.8%-95.8%) and a specificity of 88.7% (95% CI: 58.7%-97.7%) with considerable heterogeneity, which likely stems from the differences in the SRs and AI techniques used. In semi-automated workflows, sensitivity was 87.6% (95% CI: 77.2%-93.6%) and specificity was 94.1% (95% CI: 60.0%-99.4%) also with considerable heterogeneity. Among studies on full automation, the median workload savings for 100% recall was 50.0% (IQR: 10.2), while for studies on semi-automation, the median workload savings was 55.6% (IQR: 16.4).

Conclusions:
Given the findings of this review, the diagnostic performance of AI tools appeared to be superior when used in semi-automated workflows rather than fully automated ones. This suggest that AI tools hold great potential in augmenting the accuracy and efficiency of human reviewers during study selection in literature review.

Keywords: machine learning; evidence review; article screening

Introduction

The burgeoning volume of scientific literature being generated today places a great burden on evidence reviewers. In recent times, around seven million academic papers have been published

annually, leading to a staggering number of studies that need to be screened and evaluated to answer any given research question [1]. It is not unusual for a research question to generate several thousand articles for screening, especially in complex and rapidly evolving disciplines such as medicine and technology. The quality and timeliness of the knowledge synthesized from these reviews depends on the accuracy of discrimination between relevant and irrelevant studies, as well as the efficiency of the review process. On average, only 2% to 8% of articles yielded by a search strategy are ultimately included in a systematic review [2]. This relatively small collection of articles takes a substantial amount of time to select, with an observational study on meta-analysis practices discovering that only an average of 335 abstracts were screened by the average reviewer per day [3]. Furthermore, screening and selection in systematic reviews (SR) are often done by at least two reviewers in order to maximize the accuracy of the process, leading to increased total working hours. Within two years of publication, about 23% of SRs are outdated, as reviewers are not able to include all relevant articles in a timely manner [4]. Due to the burden of increasing information loads, there is a demand for methods that improve efficiency while maintaining accuracy in performing evidence reviews.

To address this problem of growing information, search engines and databases utilize various methods such as filtering and sorting to reduce screening burden and save time without reducing accuracy. A study performed by Rathbone et al., 2017 described the recall rate and reduction in screening effort of reviewers using title-only PICo-based (population, intervention, comparator) screening compared to the published yield of ten existing systematic reviews [5]. Recall, which is analogous to sensitivity, was defined as the proportion of studies in the original SR that are also retrieved using the title-only approach, while screening effort reduction described the proportion of records that no longer needed to be screened because of the title-only approach to search and selection. Results showed that in nine out of ten SRs, the recall rate was 100% while it was 67% in one study. The median screening effort reduction was 53%, ranging from 11% to 78%. While this approach reduced the effort of screening, its impact is limited in fields where vocabulary on the topic is less structured, such as in non-drug, socio-behavioral, and qualitative interventions.

Due to the need for more efficient methods for article screening and selection, new tools are also being formulated and tested to fully or partially automate the process. Artificial intelligence (AI) techniques, such as machine learning (ML) and natural language processing (NLP), have emerging applications in processing and evaluating large collections of text. NLP has been used to automate several stages of the SR process, from specifying the research question to literature search, study screening and selection, data extraction, and data synthesis.

In a review by Van Dinter et al., 2021 of articles tackling the automation of the SR process, 25 of the 41 studies (61.0%) were on the automation of citation and abstract screening, while 6 studies each were on search string building (14.6%) and maximization of recall and precision (14.6%) [6]. On application domains, 60% of the studies assessed automation of SR in the field of medicine, while 40% were on software engineering. The main NLP techniques used in the tools included latent semantic analysis, singular value decomposition, and text vectors. Most of these techniques are types of supervised machine learning (ML) wherein the algorithm is initially provided with labeled outputs during the training period. The main metrics used to evaluate the tools were precision (positive predictive value), recall (sensitivity), and F-measure. Non-ML metrics were also used, most commonly workload saved over sampling (WSS), which refers to the percentage of papers no longer screened due to the use of the tool, which was measured in nine studies. With regards to the limitations met, the most discussed was class imbalance, which refers to the large disparity between irrelevant and relevant studies in any given search yield, mentioned in 11 studies.

Though the AI-assisted methods developed show promise in studies among previously existing SR

and meta-analysis, few have assessed their performance in real-world settings. In a study by Perlmann-Arrow et al., 2022, they developed an NLP-assisted abstract screening tool and evaluated it against a living SR on SARS-CoV2 seroprevalence being updated by investigators [7]. Results showed that the time to complete screening was reduced by 45.9% with a recall of 90%, and a mean user satisfaction rating of 4.2/5.0 when the tool was compared to the two-reviewer approach. Replacing one reviewer with the tool maintained a high recall rate of 92% while screening time was reduced by 70%. This study suggested that AI tools have the potential to increase the efficiency of evidence synthesis while maintaining high recall rates and user satisfaction.

In this paper, we aimed to analyze the literature on the use of AI techniques to perform or assist in article selection through abstract screening, and to describe its pooled effect on the metrics of accuracy and workload savings.

Methods

Eligibility of Studies

This review included published quantitative studies in English, including randomized trials and observational cohort studies. We included studies that evaluated the process of article selection for SRs and meta-analyses, namely title screening, abstract screening, or both. The interventions assessed were the use of any AI tools or algorithms whose purpose was to partially or fully automate the study selection process. The comparators accepted were studies manually selected by human reviewers or studies already included in previously published SRs. Eligible studies also assessed at least one of the following outcomes: sensitivity and specificity, and/or WSS. Exclusion criteria were studies that did not specify the type of AI tool or technique used, and studies that reported neither of the desired outcomes.

Literature Search

The databases used were PubMED, ProQuest, and Cochrane Library from dates of inception until 01 January 2023. The following terms were used to search for relevant articles: ‘artificial intelligence’, ‘deep learning’, ‘machine learning’, ‘natural language processing’, ‘neural networks’, ‘machine-assisted’, ‘automated’, which were intersected with 'abstract screening' and 'title screening.'

Data collection and risk of bias assessment

Two reviewers, MS and KY assessed the titles and abstracts of the studies yielded by the search strategy. The full texts of eligible studies were retrieved and details on their objectives, population, intervention, comparator, and outcomes were extracted and entered into a table of study characteristics. Outcomes extracted include diagnostic performance and WSS. Absolute number for true positive, false positives, true negatives, and false negatives were extracted from studies that evaluated diagnostic performance.

Risk of bias for the studies was assessed using the QUADAS checklist for diagnostic accuracy. The studies included in this review are unique since the populations evaluated are essentially the pool of journal articles yielded by a search strategy while the interventions are the use of AI tools. These key differences bring about different sources of bias in addition to those found in clinical studies, such as lack of systematic method in sampling population topics (i.e. which SR topics are tested), non-standardized assessment of heterogeneity among SRs, and inherent difficulty in blinding for these studies.

Statistical Analysis

Random effects meta-analysis was done to generate univariate measures of sensitivity and specificity using RevMan version 4.3 and the ‘meta’ and ‘mada’ packages on R version 4.3.1. Bivariate analysis using R was also performed for the diagnostic performance, and a hierarchal summary receiver operating characteristics (HSROC) curve was also generated to visualize the interrelatedness between sensitivity and specificity. Heterogeneity was described using I-squared. The statistical significance for hypothesis testing was set at 0.05 for 2-tailed heterogeneity testing. To describe WSS, which was found to be non-normally distributed by Shapiro-Wilk test, median and IQR were used.

To explore sources of heterogeneity among studies that reported diagnostic performance, subgroup analysis was done according to the degree of automation, particularly if the study selection was fully automated or semi-automated, excluding the sets used for initial training. Sub-group analysis was also performed according to the tools evaluated by more than one study: Abstrackr and Rayyan.

Results

Selected studies

Two reviewers independently assessed the 64 studies yielded by the search strategy. Of these, 13 studies were found to be duplicates and were excluded. Through independent screening of titles and abstracts of the 51 remaining studies, 25 were excluded as they did not meet eligibility criteria. Further, three were excluded by virtue of only having conference abstracts, and one was excluded for not specifying the specific AI method used. There were 22 unique studies selected (Figure 1).

Figure 1. PRISMA diagram

Study characteristics

All studies found were quality improvement projects centered on determining the relative accuracy (in terms of recall/sensitivity and specificity) and workload savings achieved by AI tools when used instead of or together with human reviewers compared to the manual, two-reviewer approach to study selection traditionally used in SRs. Due to the variety of AI techniques, different tools were used in each study and even within the same study. In the sections below, we outline the general characteristics of the studies found. Additional information on individual study characteristics can be found in Appendix 1.

Risk of bias

Based on the QUADAS checklist, the majority of studies were at low risk for bias. In terms of index test, all studies had an unclear risk of bias since the investigators used existing SRs for which the relevant studies were already known. Blinding is also inherently not possible if supervised algorithms need to be trained. Hence, the overall risk of bias for the studies is low. The individual ratings for risk of bias can be found in Appendix 2.

Population

There was significant heterogeneity in the number and nature of the SRs used in the studies. A median of three SRs (IQR: 9) were tested in each study. Each SR involved a median of 17,582 primary studies (IQR: 36,219), while the prevalence of relevant primary studies included in the final review was a median of 1.67% (IQR: 4.08) of the total search yield. The SR topics were also variable, spanning various medical topics, but the majority were on therapeutics.

Machine learning methods

In some studies, several classifying schemes and algorithms were compared. For the sake of conciseness, the best-performing algorithm in terms of sensitivity among these will be reflected in the study outcomes. If sensitivity and specificity were not reported as one of the study outcomes, workload savings was used to identify the best-performing algorithm. Of the 22 studies included, five assessed the performance of the Abstrackr tool, three assessed DistillerSR, and two assessed Rayyan. The other algorithms tested included down-sample unigram, elemental similarity, General Architecture for Text Engineering (GATE), gradient-boosted Bidirectional Encoder Representations from Transformers (BERT), K-nearest neighbor, Multi-modal Missing Data aware Stacked Autoencoder (MMiDaS-AE), random forest, regular expressions (RegEx)on R, Research Screener, support vector machines (SVM), topic modelling, and word embeddings.

Comparators

Among the 22 studies, eight studies made use of fully unsupervised algorithms without additional human reviewers as the intervention group, and human inputs in existing SRs as the comparator group. The remaining 14 studies used semi-automated processes, wherein the AI tools comprised only part of the process and additional human reviewers performed additional screening for the included and/or excluded abstracts.

Evaluation Outcomes

The outcome measures include sensitivity and specificity, and workload saved over sampling (WSS). Recall, also known as sensitivity, is defined as the proportion of studies selected by the algorithm which are also included in the final SR divided by the total number of studies included in the final SR. Workload saved over sampling (WSS) refers to the percentage of papers no longer screened due to the use of the algorithm. WSS can be measured at different levels of recall, such as WSS95 (indicating saved workload at 95% recall) and WSS100 (indication saved workload at 100% recall). Of the 22 studies selected, 13 reported diagnostic performance, 14 reported WSS, and five studies reported both outcomes.

Diagnostic performance

As mentioned in the methods, the diagnostic performance was analyzed in subgroups according to the degree of automation and tool used (if two or more studies tested the same tool).

Full automation

Five studies investigated the diagnostic performance of AI tools in fully automated study selection without any assistance or cross-checking with human reviewers (Figure 2). These five studies used different AI tools or algorithms, namely word embedding, down sample unigram, support vector machine, topic modelling, and K-nearest neighbor. The sensitivity of these tools for correctly classifying relevant studies ranged from 42.0% to 98.0%. Sensitivity had a pooled value of 85.6% (95% CI: 60.8%-95.8%) with considerable heterogeneity (I2=99.3%, p<0.0001), likely stemming from the inherent differences in AI algorithms used as well as the method of training these algorithms. Specificity was pooled at 88.7% (95% CI: 58.7%-97.7%) with considerable heterogeneity (I2=99.8%, p<0.00001). With this performance, full automation could theoretically correctly classify 14 relevant articles and miss 3 articles out of every 1,000 studies screened.

Figure 2. Forest plot for studies using ML tools in a fully automated selection process

Semi-automation

There were eight studies evaluating the diagnostic performance of AI tools in conjunction with human reviewers during study selection in literature review (Figure 3). Among the studies, two used Abstrackr, two used Rayyan with differing cut-offs, and one study each used GATE, gradient-boosted BERT, random forest, and DistillerSR. The sensitivity pooled value was 87.6% (95% CI: 77.2%-93.6%) but with considerable heterogeneity (I2=95.3%, p<0.0001). Specificity had a pooled value of 94.1% (95% CI: 60.0%-99.4%) and high heterogeneity (I2=99.8%, p<0.00001), likely due to the use of different tools and training techniques. With this performance, semi-automation could correctly classify 15 relevant articles and miss two articles out of every 1,000 studies screened.

Figure 3. Forest plot for studies using ML tools in a semi-automated selection process

Abstrackr

Two studies tested the diagnostic performance of Abstrackr for study selection in a semi-automated workflow. The pooled sensitivity value was 88.5% (95% CI: 85.2%-91.1%) with substantial heterogeneity (I2=75.5%, p=0.0352), while specificity was 68.9% (95% CI: 64.5%-73.1%) with considerable heterogeneity (I2=98.6%, p<0.0001). Possible causes of heterogeneity include differences in the topics of SRs used and differences in the judgment of the human reviewers involved in the studies. Out of every 1,000 articles screened, Abstrackr could correctly select 15 relevant articles while missing 2 articles.

Figure 4. Forest plot for studies using Abstrackr

Rayyan

Two studies tested the diagnostic performance of Rayyan for study selection in a semi-automated workflow. The pooled sensitivity value is 73.2% (95% CI: 32.5%-93.9%) with considerable heterogeneity (I2=99.1%, p<0.0001), while specificity is 76.6% (95% CI: 49.9%-91.5%) also with considerable heterogeneity (I2=99.8%, p<0.0001). Sources of heterogeneity include different thresholds in Rayyan (which is a five-star rating system), differences in topics of SRs used, and differences in judgment of the human reviewers. Out of every 1,000 articles, Rayyan could correctly select 12 relevant articles while failing to include 5 relevant articles.

Figure 5. Forest plot for studies using Rayyan

HSROC

In the HSROC curve generated, the solid line depicts the summary receiver operating characteristics curve, while the gray line depicts the 95% confidence interval surrounding the summary point. The circle (○) represents the summary point for the 13 studies that reported on diagnostic performance included in this review. The sensitivity at this point is 86.6% (95% CI:75.9%-93.0%), and the specificity is 91.0% (95% CI: 72.1%-97.5%). The area-under-the-curve is 0.93.

Figure 6. HSROC curve for all studies that reported diagnostic performance

Workload savings

Fourteen studies reported workload savings with use of AI tools in fully and partially automated study selection workflows. Among studies on full automation, the median workload savings for 100% recall was 50.0% (IQR: 10.2); for 95% recall, the median saved workload was 45.3% (IQR:

7.92). Among studies on semi-automation, the median workload savings for 100% recall was 55.6% (IQR: 16.4), while for 95% recall, the median saved workload was 49.1% (IQR: 18.7).

Discussion

Given the findings of this review, the diagnostic performance of AI tools appeared to be superior when used in semi-automated workflows rather than fully automated ones. ML tools used alone in fully automated workflows were shown to have generally lower sensitivity and specificity than when used in a complementary manner to assist human reviewers. This is somewhat expected given the required baseline knowledge and nuances in judgment exercised by human reviewers during the article selection portion of literature review. Certain aspects such as value judgement cannot yet be replicated by current algorithms and remain the purview and responsibility of human evidence reviewers.

Our findings also suggest that despite the growing and ubiquitous usage of machine learning applications in medicine and industry, current algorithms and tools have insufficient accuracy to classify all relevant articles correctly. In the field of evidence review, absolute or near-absolute recall is of utmost importance and the corresponding increase in screening burden is accepted in return [8]. To address this large screening burden, AI tools may be leveraged to assist human reviewers and reduce workload, but these do not perform well enough for full automation. These findings are consistent with a prior review showing that ML has high potential for supporting initial screening of studies in literature review with additional screening done by human reviewers [9].

The findings of this review are limited by considerable heterogeneity due to the different SRs and SR topics in the study population. In the included studies, the vast majority concern the medical field, hence the performance measures described here may not be generalizable to study selection in non-medical fields. Medical research terms display greater standardization of structure and taxonomy compared to terms in other fields of research, such as qualitative, social, and behavioral research [10]. How this difference in structure may affect the training and performance of AI algorithms is not yet well-known and should be the topic of further study.

Another important problem faced by AI algorithms in article selection is the generally low prevalence of relevant studies in every systematic review. In this review, the median prevalence of relevant articles was 1.67%, leading to a problem of class imbalance and impairing the training of ML models [6]. Since the article yields were sourced largely from online databases which may have different indexing standards, the results of AI-driven SRs may also be threatened by publication bias. Finally, the different tools and algorithms used in the studies also contributed to the heterogeneity of results. Furthermore, each study generally assessed a single tool, hence inter-group comparisons between tools were not available.

Conclusion

The results of our review suggest that AI tools hold great potential in augmenting the accuracy and efficiency of human reviewers during study selection in literature review. Further studies should be done to elucidate the performance of AI tools in systematic reviews in other non-medical fields and to investigate the application of other AI methods, such as deep learning, in the literature review process.

Conflicts of Interest

None declared

Funding

None

Abbreviations

SR: systematic review
ML: machine learning

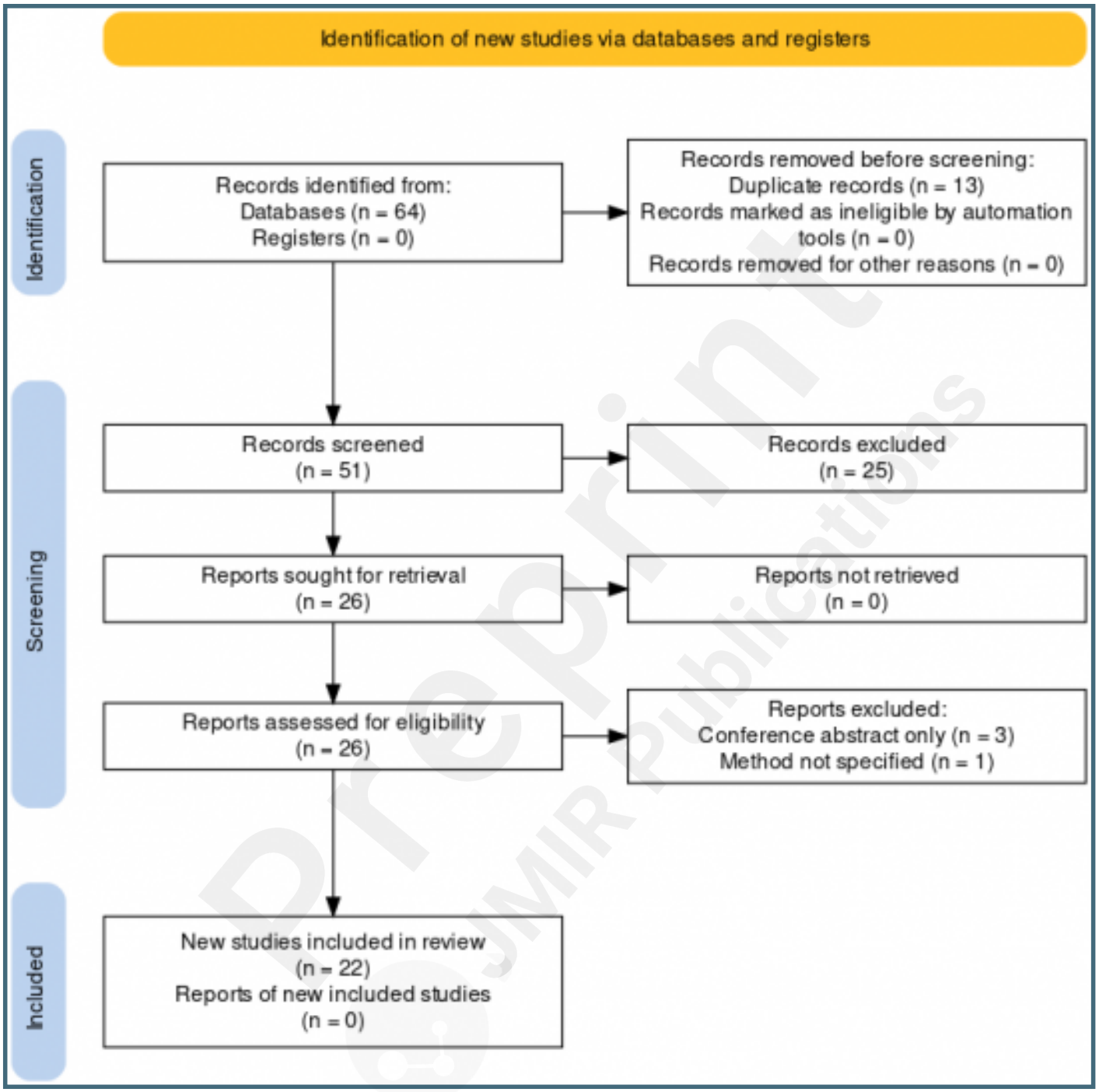
References

1. Fire M, Guestrin C. Over-optimization of academic publishing metrics: observing Goodhart's Law in action. GigaScience. 2019;8(6):giz053. doi:10.1093/gigascience/giz053
2. Wang Z, Nayfeh T, Tetzlaff J, O'Blenis P, Murad MH. Error rates of human reviewers during abstract screening in systematic reviews. PloS One. 2020;15(1):e0227742. doi:10.1371/journal.pone.0227742
3. Polanin JR, Pigott TD, Espelage DL, Grotpeter JK. Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses. Research Synthesis Methods. 2019;10(3):330-342. doi:10.1002/jrsm.1354
4. Jonnalagadda SR, Goyal P, Huffman MD. Automating data extraction in systematic reviews: a systematic review. Systematic Reviews. 2015;4:78. doi:10.1186/s13643-015-0066-7
5. Rathbone J, Albarqouni L, Bakhit M, Beller E, Byambasuren O, Hoffmann T, Scott AM, Glasziou P. Expediting citation screening using PICo-based title-only screening for identifying studies in scoping searches and rapid reviews. Systematic Reviews. 2017;6(1):233. doi:10.1186/s13643-017-0629-x
6. Van Dinter R, Tekinerdogan B, Catal C. Automation of systematic literature reviews: A systematic literature review. Information and Software Technology. 2021;136:106589. doi:10.1016/j.infsof.2021.106589
7. Perlman-Arrow S, Loo N, Bobrovitz N, Yan T, Arora RK. A real-world evaluation of the implementation of NLP technology in abstract screening of a systematic review. Research Synthesis Methods. 2023;14(4):608-621. doi:10.1002/jrsm.1636
8. Li J, Larsen K, Abbasi A, et al. TheoryOn: a design framework and system for unlocking behavioral knowledge through ontology learning. MIS Quarterly. 2020;44(4):1733-1772.
9. Wagner G, Lukyanenko R, Paré G. Artificial intelligence and the conduct of literature reviews. Journal of Information Technology. 2022;37(2):209-226. doi:10.1177/02683962211048201
10. O'Mara-Eves A, Thomas J, McNaught J, et al. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Systematic Reviews. 2015;4.

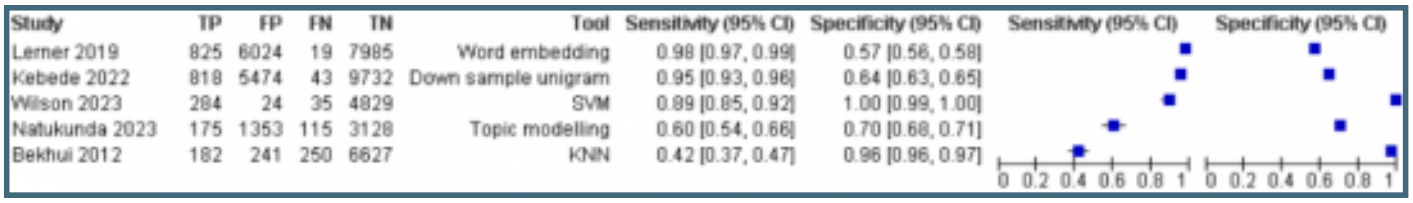
Supplementary Files

Figures

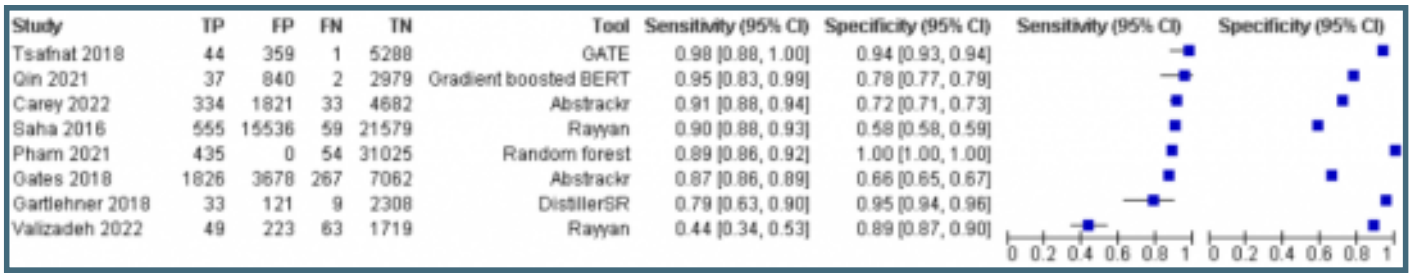
PRISMA diagram.



Forest plot for studies using ML tools in a fully automated selection process.

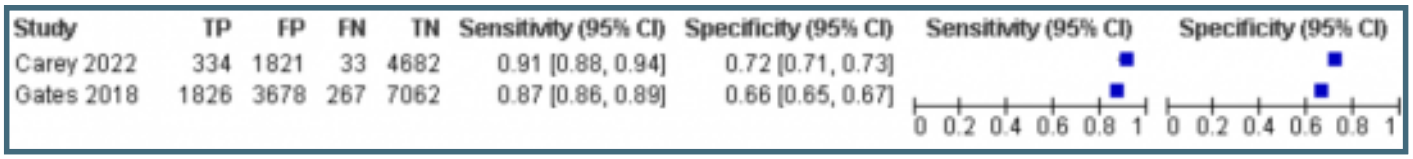


Forest plot for studies using ML tools in a semi-automated selection process.

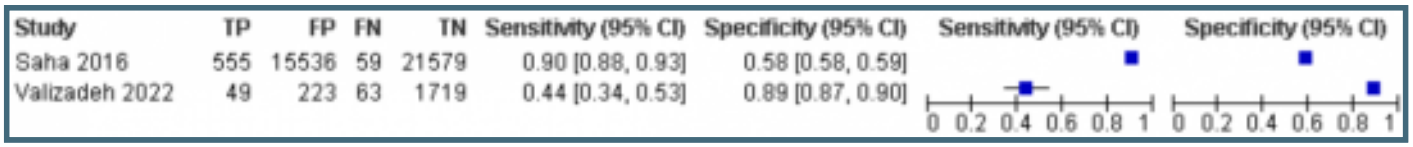


Study	TP	FP	FN	TN	Tool	Sensitivity (95% CI)	Specificity (95% CI)
Tsafnat 2018	44	359	1	5288	GATE	0.98 [0.88, 1.00]	0.94 [0.93, 0.94]
Qin 2021	37	840	2	2979	Gradient boosted BERT	0.95 [0.83, 0.99]	0.78 [0.77, 0.79]
Carey 2022	334	1821	33	4682	Abstrackr	0.91 [0.88, 0.94]	0.72 [0.71, 0.73]
Saha 2016	555	15536	59	21579	Rayyan	0.90 [0.88, 0.93]	0.58 [0.58, 0.59]
Pham 2021	435	0	54	31025	Random forest	0.89 [0.86, 0.92]	1.00 [1.00, 1.00]
Gates 2018	1826	3678	267	7062	Abstrackr	0.87 [0.86, 0.89]	0.66 [0.65, 0.67]
Gartlehner 2018	33	121	9	2308	DistillerSR	0.79 [0.63, 0.90]	0.95 [0.94, 0.96]
Valizadeh 2022	49	223	63	1719	Rayyan	0.44 [0.34, 0.53]	0.89 [0.87, 0.90]

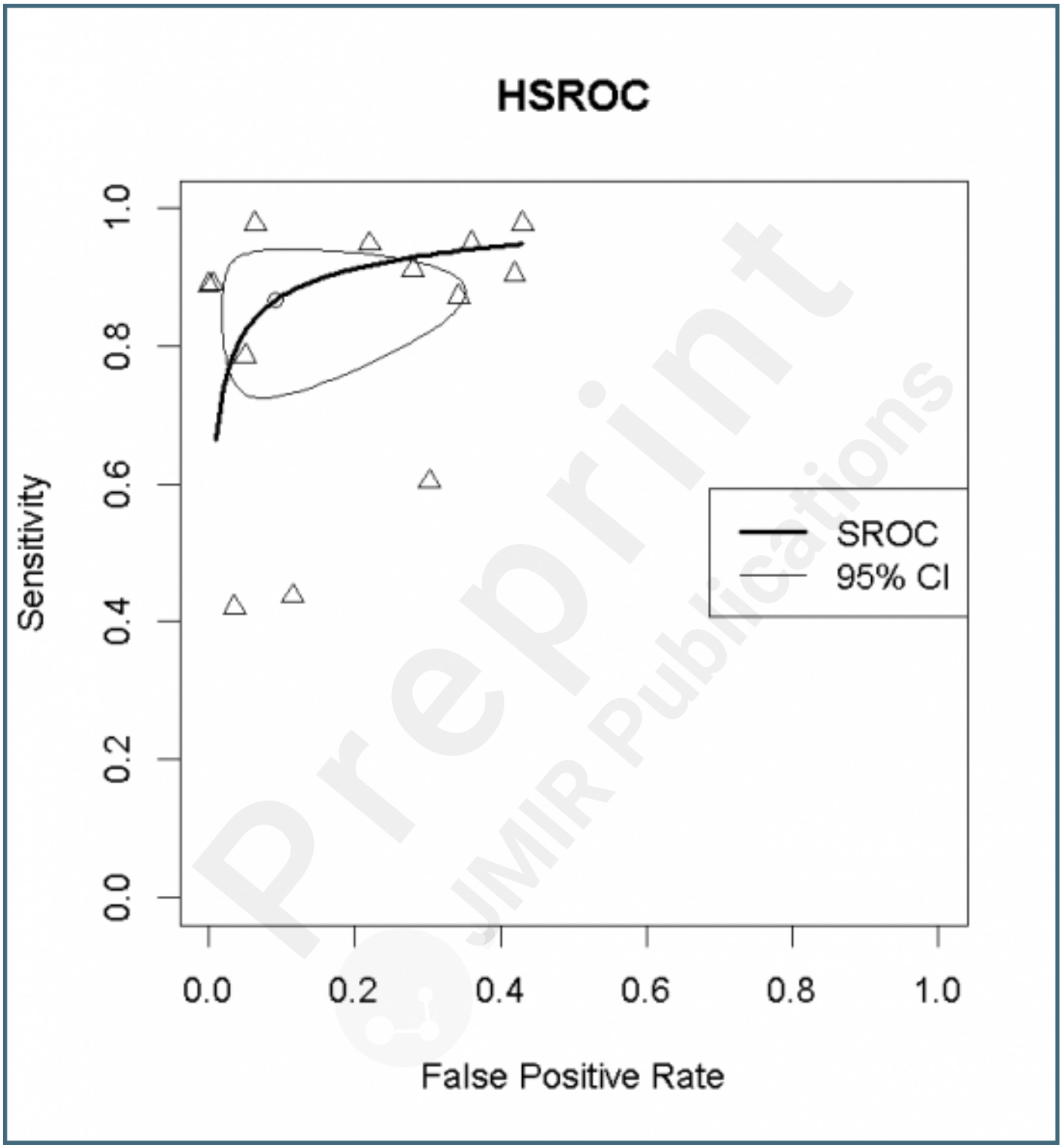
Forest plot for studies using Abstrackr.



Forest plot for studies using Rayyan.



HSROC curve for all studies that reported diagnostic performance.



Multimedia Appendixes

Table of study characteristics.
URL: http://asset.jmir.pub/assets/e8d5d8ade0ba8ab55b9157831105c244.docx

Study risk of bias assessment.
URL: http://asset.jmir.pub/assets/8baf8deece094b8eae80b99b8211a958.docx