Journal of the American Medical Informatics Association, 29(8), 2022, 1425–1432
https://doi.org/10.1093/jamia/ocac066
Advance Access Publication Date: 31 May 2022
Review

Review

Automated medical literature screening using artificial intelligence: a systematic review and meta-analysis

Yunying Feng[1], Siyu Liang[2], Yuelun Zhang[3,4], Shi Chen[2,4], Qing Wang[5], Tianze Huang[1], Feng Sun[6], Xiaoqing Liu[4,7], Huijuan Zhu[2,4], and Hui Pan[8]

[1]Eight-year Program of Clinical Medicine, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China, [2]Department of Endocrinology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China, [3]Medical Research Center, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China, [4]Clinical Epidemiology Unit, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China, [5]Department of Automation, Tsinghua University, Beijing, China, [6]Department of Epidemiology and Biostatistics, School of Public Health, Peking University Health Science Center, Beijing, China, [7]Department of Infectious Diseases, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China, and [8]Department of Endocrinology, State Key Laboratory of Complex Severe and Rare Diseases, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

Yunying Feng, Siyu Liang, and Yuelun Zhang contributed equally to this work.

Corresponding Author: Hui Pan, MD, Department of Endocrinology, State Key Laboratory of Complex Severe and Rare Diseases, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, 1 Shuaifuyuan, Dongcheng District, 100730 Beijing, China; panhui20111111@163.com

Received 13 December 2021; Revised 21 April 2022; Editorial Decision 22 April 2022; Accepted 16 May 2022

ABSTRACT

Objective: We aim to investigate the application and accuracy of artificial intelligence (AI) methods for automated medical literature screening for systematic reviews.

Materials and Methods: We systematically searched PubMed, Embase, and IEEE Xplore Digital Library to identify potentially relevant studies. We included studies in automated literature screening that reported study question, source of dataset, and developed algorithm models for literature screening. The literature screening results by human investigators were considered to be the reference standard. Quantitative synthesis of the accuracy was conducted using a bivariate model.

Results: Eighty-six studies were included in our systematic review and 17 studies were further included for meta-analysis. The combined recall, specificity, and precision were 0.928 [95% confidence interval (CI), 0.878–0.958], 0.647 (95% CI, 0.442–0.809), and 0.200 (95% CI, 0.135–0.287) when achieving maximized recall, but were 0.708 (95% CI, 0.570–0.816), 0.921 (95% CI, 0.824–0.967), and 0.461 (95% CI, 0.375–0.549) when achieving maximized precision in the AI models. No significant difference was found in recall among subgroup analyses including the algorithms, the number of screened literatures, and the fraction of included literatures.

Discussion and Conclusion: This systematic review and meta-analysis study showed that the recall is more important than the specificity or precision in literature screening, and a recall over 0.95 should be prioritized. We recommend to report the effectiveness indices of automatic algorithms separately. At the current stage manual literature screening is still indispensable for medical systematic reviews.

Key words: evidence-based medicine, artificial intelligence, systematic review, natural language process, diagnostic test accuracy

INTRODUCTION

In evidence-based medicine, systematic reviews are currently one of the most powerful tools for evidence collection, critical appraisal, and synthesis for certain research questions. Major steps of systematic reviews include the review question formulation, inclusion criteria development, search strategy devising and implementation, literature screening, data collection, meta-analyses where appropriate, and risk of bias assessment.[1,2] The literature screening step can be extremely time-consuming and prevent on-time completion and updates of systematic reviews.[3–5] Practitioners have had difficulty finding latest high-quality evidence for medical questions in systematic reviews, which may influence their decision-making in clinical practice.

To reduce the workload and improve efficacy of evidence synthesis, researchers are exploring artificial intelligence (AI) methods in systematic reviews, such as pattern recognition and machine learning.[6] At present, AI tools for systematic reviews, based on machine learning, text mining, and natural language processing (NLP), are on trial in the highly standardized and repetitive procedures of systematic reviews, such as literature screening, data extraction, and risk-of-bias assessment.[7–9] These automation tools have been newly mentioned in the Preferred Reporting Items for a Systematic Review and Meta-analysis (PRISMA) 2020 statement, yet little strong evidence was provided for their application.[10,11] A systematic review on automating data extraction in systematic reviews reported insufficient development of automatic methods.[12] For literature screening, automated classification systems[13] or hybrid relevance rating models[14] have been evaluated in specific datasets, requiring further extension and performance improvement. Few studies reviewing automated literature screening have been found. To address this gap in knowledge, we sought to perform a systematic review and meta-analysis on accuracy of AI methods for literature screening in medical evidence synthesis. In this review, the term *literature* refers to the literature used in assessing the diagnostic accuracy of the AI methods (similar to "participant" in traditional diagnostic accuracy studies), and *study* relates to the AI model study included in our systematic review (similar to "primary study" in traditional systematic reviews).

MATERIALS AND METHODS

This manuscript follows the Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies (PRISMA-DTA).[15] The protocol of this systematic review was registered on PROSPERO (CRD42020170815, April 28, 2020).

Eligibility criteria

We included studies that met the following criteria: (1) automatic methods were developed for literature screening for medical systematic reviews, (2) the research question and source of dataset used were reported, and (3) the literature screening results by human investigators were set as the reference standard. Editorials, commentaries, and narrative review articles were excluded.

Information source and search strategy

We developed the search strategy and conducted literature searches in 3 major public electronic databases on biomedicine and computer science: PubMed, Embase, and IEEE Xplore Digital Library. Retrieval was restricted to papers published between January 1, 2000 and December 22, 2021 (the last search date, see Supplementary Table S1 in Supplementary File S1). We chose this date range because AI algorithms prior to 2000 are unlikely to satisfy the requirements for literature screening in systematic reviews. Reference lists of initially included studies were also checked to find more relevant studies. Potentially relevant abstracts and preprints were also searched in Google Scholar. No restrictions were set on language.

Data collection and risk of bias assessment

Different from traditional systematic reviews, the "participants" in this review were original medical studies and literatures, and the index test was AI algorithms used for automatic literature screening. We defined traditional literature screening by human investigators as the reference standard. The outcomes of our meta-analysis include effectiveness of literature screening, as well as labor and time saving, which were mainly evaluated by recall (sensitivity), precision [positive predictive value (PPV)], specificity, and the work saved over sampling (WSS).[13] WSS was defined as the work saved over and above the work saved by simple random sampling for a given level of recall, and could be calculated as:

$$\mathrm{WSS} = (\mathrm{TN} + \mathrm{FN})/\mathrm{N} - (1.0 - \mathrm{R}),$$

where TN was the number of true negatives identified by the classifier, FN was the number of false negatives identified by the classifier, N was the total number of samples in the test set, and R was the recall.[13] Since semi-automation and active learning methods require manual screening or interactions during processing, studies applying these models were not considered in final meta-analysis.

Study record information including titles and abstracts from searched online databases was downloaded. Duplicate citations were removed and the records were imported into EndNote X9.3.2 software (Thomson Reuters, Toronto, Ontario, Canada) for further assessment. All citations and abstracts were independently screened by 2 reviewers (SL, YF) based on the titles and abstracts, and the full texts of potentially eligible citations were then reviewed independently by the same 2 reviewers to select the studies for final inclusion. Disagreements in both initial screening and final screening were resolved by discussion with a methodologist (YZ). The excluded studies were listed and noted according to PRISMA-DTA flowchart.

Using a designed data collection form, 2 reviewers (SL, YF) independently extracted and verified data from finally included studies. The detailed information of training sets and validation sets, AI algorithms, and effectiveness and work-saving indices were collected. Conflicts were resolved through discussion or by consulting another member of the review team (YZ).

Two reviewers (SL, YF) independently assessed risk of bias with a revised checklist based on Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2).[16] Detailed QUADAS-2 instrument used in our analysis is listed in Supplementary File S1. Disagreements were resolved by a third reviewer (YZ).

Statistical analysis

Studies that reported the diagnostic 2-by-2 table or enough relevant information to calculate the data were quantitatively combined using diagnostic meta-analysis methods. Diagnostic accuracy was expressed by indicators including recall, precision (PPV), WSS, and the summary receiver operating characteristic curve (SROC). Generally, numerous parameters can be changed and optimized in the AI models to fit different application scenarios, so it is common that 1 automation study reported multiple groups of effectiveness and labor-saving indices. In this meta-analysis, we mainly focused on 2 groups of effectiveness and labor-saving indices: (1) the precision and WSS when achieving the maximized recall and (2) the recall and WSS when achieving the maximized precision. Values of WSS were directly collected from the studies or calculated based on reported data of the studies without combination. Since WSS is dependent on the recall, it was reported at diverse recall levels in different studies. Therefore, we merely calculated them according to collected recall values, and did not combine the potential heterogeneous WSS values. A bivariate model was used to combine the diagnostic accuracy indices including the recall, specificity, and precision.[17] Predefined subgroup analyses were conducted according to AI algorithms (divided into the support vector machines (SVMs) group and other algorithms group including naïve Bayes (NB), K-nearest neighbor (k-NN), perceptron, etc), number of screened literatures for model validation (similar to the "number of participants" in traditional diagnostic accuracy study, divided by the median value of all eligible studies), and fraction of included literatures (similar to the "prevalence" in traditional diagnostic accuracy studies, divided by the median). Statistical analysis was completed in R (version 4.0.2, R Foundation for Statistical Computing, Vienna, Austria, 2020, https://www.R-project.org/) with "mada" package.[18] SROC plots were depicted using Review Manager 5 (RevMan 5).[19]

RESULTS

Search and screening

Applying the strategy shown in Supplementary Table S1 (Supplementary File S1), the electronic search was conducted on December 22, 2021 in PubMed, Embase, and IEEE Xplore Digital Library. A total of 10 102 publications were identified with 2239 duplicates (Figure 1). An additional 104 publications were found through citation searching. After screening 7967 titles and abstracts, 161 studies remained for further evaluation according to the predefined inclusion and exclusion criteria. After screening the full texts, 86 studies were included in the systematic review, and 71 studies were finally included in the meta-analysis. Detailed citations of the full-text screened publications were listed in Supplementary File S2.

Characteristics of included studies

The studies included in the systematic review were published between 2006[13] and 2021[20] and applied various AI algorithms to examine and improve automated medical literature screening (Supplementary Table S3, Supplementary File S3). SVM was the most commonly used classifier for literature screening, yet NB, k-NN, perceptron, random forest, convolutional neural networks, radial basis function kernel, and other algorithms were applied as well. Multiple studies analyzed and compared the performance of more than 1 algorithm. Most automation studies used results of literature screening from existing systematic reviews to train and evaluate their classification models, while the rest of the studies directly searched for literatures in certain databases or journals. All studies used article titles, abstracts, and metadata rather than full texts for training or validation. For studies included in the meta-analysis, the datasets were divided into training and testing sets, for AI model training and validation separately.

Risk of bias

Supplementary Table S4 (Supplementary File S3) shows the risk of bias assessment of studies included in the meta-analysis according to the revised checklist of QUADAS-2. We classified the majority of domains as unclear or low risk of bias for all of the studies included. None of the studies were rated as low risk of bias across all 4 of the categories. The "patient" (literature) selection was the major source of the risk of bias. Of the 18 studies 15 reported model performances on testing datasets from only 1 systematic review. Twelve studies searched only MEDLINE when building the datasets. Ignoring other databases such as EMBASE may lead to inappropriate exclusions. Significant risk of bias also came from the index test. Eight studies repeated random cross-validation, leading to a high distribution similarity between training and testing dataset. Five studies presented accuracy data at thresholds or hyperparameters which were not prespecified. In terms of applicability, all studies were low risk of bias given that the literature selection, AI algorithms, and reference standards conformed to this review.

Effectiveness and labor-saving indices

Among all the included studies in the meta-analysis, 15 of them were available for data synthesis with maximized recall values (Supplementary Table S5, Supplementary File S3), while 17 studies were available with maximized precision values (Supplementary Table S6, Supplementary File S3). A few studies have reported their WSS values, and we calculated the rest of the WSS based on provided information from studies. Combined estimates of recall, specificity, and precision of studies when achieving maximized recall values (Table 1) were 0.928 (95% CI, 0.878–0.958), 0.647 (95% CI, 0.442–0.809), and 0.200 (95% CI, 0.135–0.287). Combined estimates of recall, specificity, and precision of studies when achieving maximized precision values (Table 1) were 0.708 (95% CI, 0.570–0.816), 0.921 (95% CI, 0.824–0.967), and 0.461 (95% CI, 0.375–0.549).

As shown in Supplementary Table S5 (Supplementary File S3), the maximized recall of each included study in the meta-analysis ranged from 0.484 to 1.000, while precision ranged from 0.061 to 0.581. The value of WSS ranged from −0.003 to 0.897. Supplementary Table S6 (Supplementary File S3) shows the maximized precision of each included study, ranging from 0.232 to 0.800. The corresponding recall of these included studies ranged from 0.240 to 0.970, and the calculated WSS ranged from 0.095 to 0.841. According to the SROC plot of recall and specificity for all included studies when achieving the maximized recall (Figure 2A), the recall values of most studies could reach very high levels, while the specificity values had a large variation (the minimum was less than 0.1 and the maximum was over 0.9). As for the studies with the maximized precision (Figure 2B), the variation ranges of both recall values and specificity values were large, though several studies achieved high levels of recall and specificity simultaneously.

Table 2 shows the results of subgroup analyses according to different algorithms, the number of literatures, and the fraction of included literatures. In subgroup analyses for different AI algorithms, no significant differences were found between SVM and other algorithms in

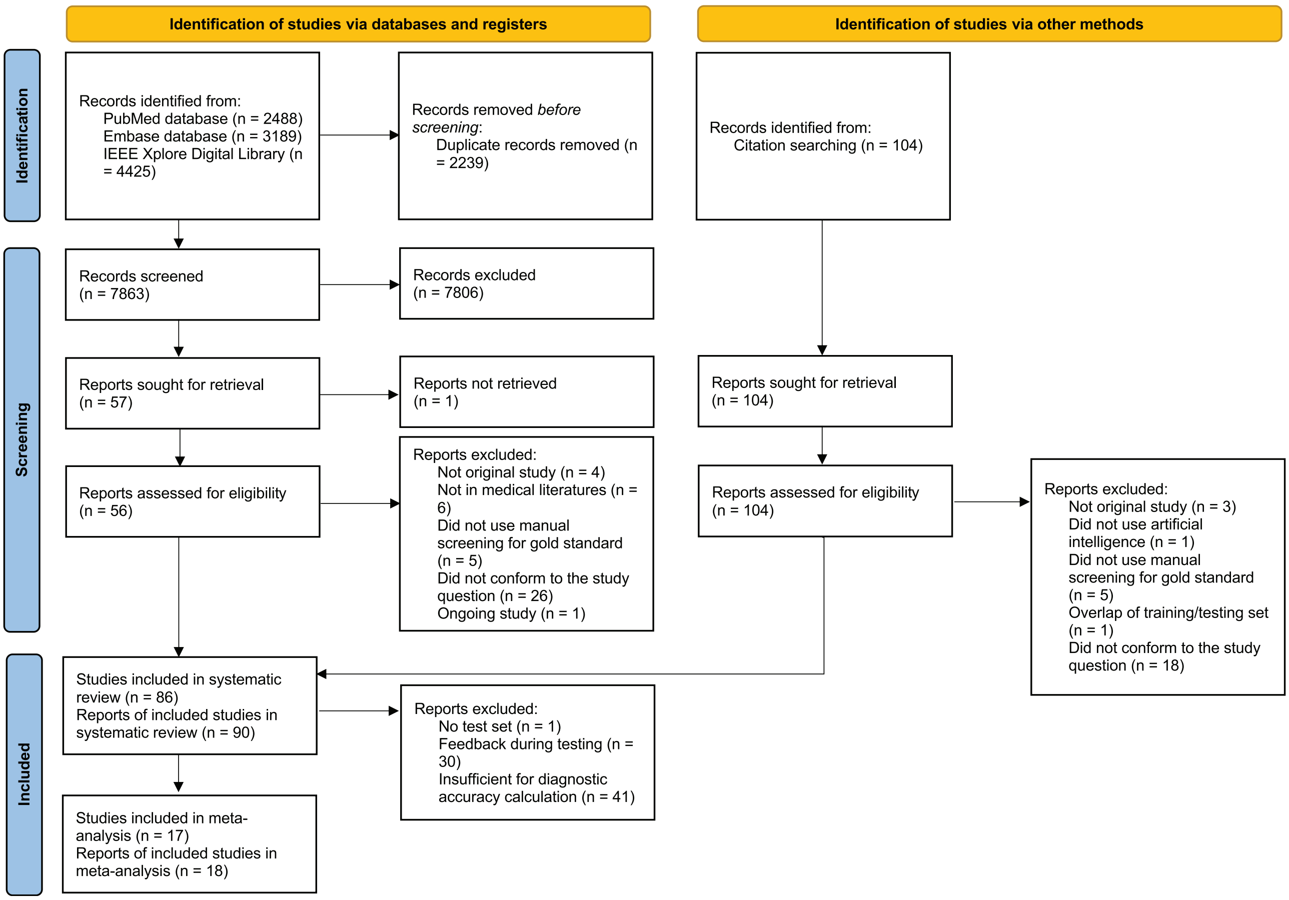
Identification of studies via databases and registers
Identification of studies via other methods
Identification
Screening
Included
Records identified from:
PubMed database (n = 2488)
Embase database (n = 3189)
IEEE Xplore Digital Library (n = 4425)
Records removed before screening:
Duplicate records removed (n = 2239)
Records identified from:
Citation searching (n = 104)
Records screened (n = 7863)
Records excluded (n = 7806)
Reports sought for retrieval (n = 57)
Reports not retrieved (n = 1)
Reports sought for retrieval (n = 104)
Reports assessed for eligibility (n = 56)
Reports excluded:
Not original study (n = 4)
Not in medical literatures (n = 6)
Did not use manual screening for gold standard (n = 5)
Did not conform to the study question (n = 26)
Ongoing study (n = 1)
Reports assessed for eligibility (n = 104)
Reports excluded:
Not original study (n = 3)
Did not use artificial intelligence (n = 1)
Did not use manual screening for gold standard (n = 5)
Overlap of training/testing set (n = 1)
Did not conform to the study question (n = 18)
Studies included in systematic review (n = 86)
Reports of included studies in systematic review (n = 90)
Reports excluded:
No test set (n = 1)
Feedback during testing (n = 30)
Insufficient for diagnostic accuracy calculation (n = 41)
Studies included in meta-analysis (n = 17)
Reports of included studies in meta-analysis (n = 18)

Figure 1. Review flow diagram.

Table 1. Combined effectiveness indices of all eligible studies in meta-analysis

Analysis	Number of studies	Recall/Se (95% CI)	Specificity (95% CI)	Precision/PPV (95% CI)
All eligible studies when achieving maximized recall	15	0.928 (0.878–0.958)	0.647 (0.442–0.809)	0.200 (0.135–0.287)
All eligible studies when achieving maximized precision	17	0.708 (0.570–0.816)	0.921 (0.824–0.967)	0.461 (0.375–0.549)

CI: confidence interval; PPV: positive predictive value; Se: sensitivity.

recall, specificity, and precision. Similarly, no significant differences were found in subgroup analysis for the number of literatures. When achieving the maximized recall, precision for larger fraction of included literatures was higher than that with smaller fraction of included literatures (precision for larger fraction was 0.296, 95% CI, 0.096–0.625, precision for smaller fraction was 0.137, 95% CI, 0.083–0.217, $P = .020$). The specificity in studies with larger fraction of included literatures was lower than that with smaller fraction of included literatures when the studies reached the maximized precision (specificity for larger fraction was 0.729, 95% CI, 0.220–0.963, specificity for smaller fraction was 0.977, 95% CI, 0.945–0.991, $P < .001$). For studies reaching the maximized recall, studies screening larger number of literatures had better overall performance than the subgroup of lower number of literatures (Figure 3B). A decline in overall performance was found in the subgroup of larger fraction of included literatures both for the maximized recall and the maximized precision (Figures 3C and 4C).

DISCUSSION

Our review identified 86 AI studies in automated medical literature screening, among them 17 studies were included in our final meta-analysis. The combined recall was 0.928 when achieving the maximized recall by optimizing the AI model. However, this value was only 0.708 when achieving the maximized precision, indicating that more literatures might be missed if the automation model focused on precision. The WSS values varied largely among studies either in the maximized recall group or in the maximized precision group.

This is the first systematic review and meta-analysis in the area of automatic literature screening aimed to quantitatively evaluate the performance of AI methods and provide recommendations based on evidence. We performed this systematic review and meta-analysis in accordance with a robust and prespecified protocol. A comprehensive literature search in major electronic databases was performed in our review, using a reproducible retrieval strategy. With 7967 screened records and 86 finally included studies, we are

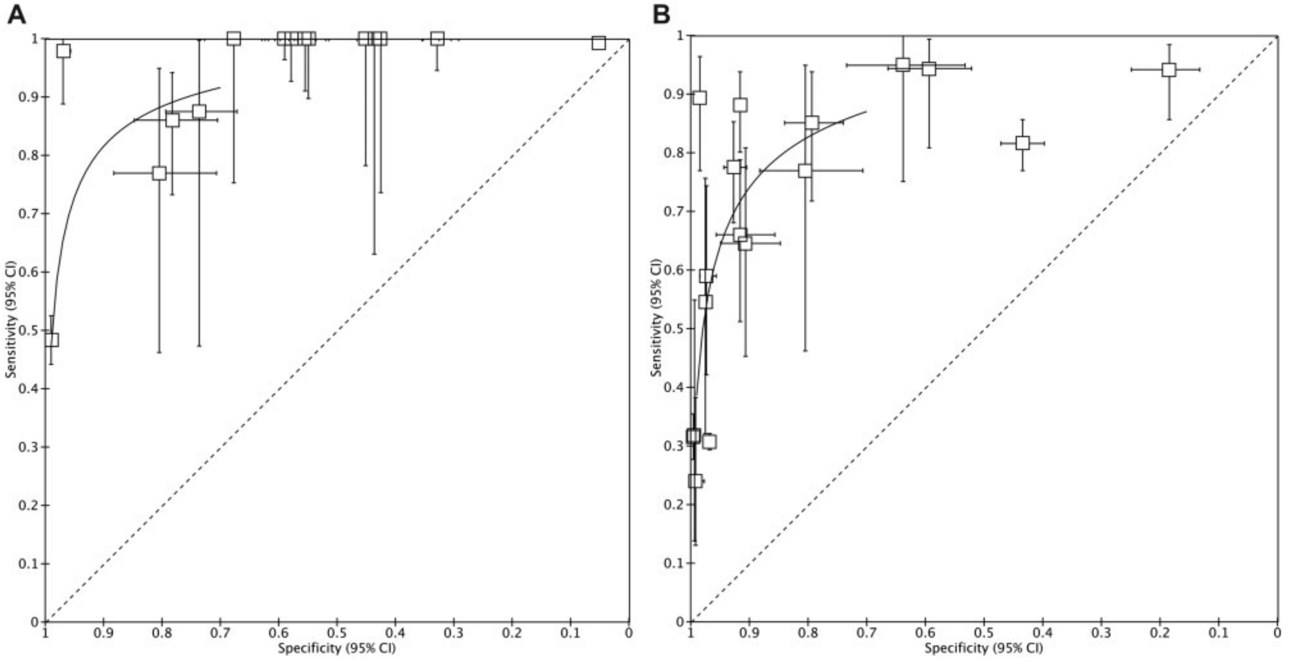

Figure 2. Summary receiver operating characteristic (SROC) plot of sensitivity and specificity of automatic algorithms for literature screening (all included studies). The hollow symbols surrounded by 95% CI region (interrupted line) represent the pairs of sensitivity and specificity from the included studies; the symbols are scaled according to sample sizes of the studies. (A) SROC plot when achieving maximized recall (sensitivity); (B) SROC plot when achieving maximized precision (positive predictive value).

confident that the included studies constitute the most representative samples of AI studies investigating the literature screening task in medical evidence synthesis. Currently, some studies failed to report necessary data for diagnostic accuracy evaluation. Besides, active learning and relevance feedback models in multiple studies introduce extra expert annotation during the screening process, which would improve the model performance accordingly. Therefore, we further excluded these studies and included a total of 17 studies in meta-analysis to evaluate diagnostic accuracy of AI methods. The relatively large number of studies allowed for pooling combined estimates of recall, specificity, and precision, and especially for comparing the combined recall when achieving the maximized recall or precision.

Literature screening is an imbalanced classification task, for the total number of screened literatures is large while fraction of included literatures is usually very low.[21] In automatic literature screening, the recall of the AI model reflects the ability to correctly identify eligible literatures.[22,23] More eligible literatures containing quality evidence would be missed by the low-recall models, introducing significant selection bias to systematic review. A low-precision model would mistakenly identify many irrelevant literatures, leading to more paper-reading load in the follow-up manual screening. For medical evidence synthesis, the introduction of bias is unacceptable. Thus, in practice, a high level of recall should be prioritized to make sure the automatic screening process includes as many eligible literatures as possible. It is meaningless to focus on precision when the level of recall failed to be optimized (theoretically should be close to 100%), for it indicates the probability of literature missing. Our study is the first one to provide solid evidence for diagnostic accuracy indices balancing in automatic literature screening. Two groups of effectiveness indices under maximized recall and maximized precision respectively in our review were combined separately. According to our analysis, current models would miss 7.2% of literatures on average when achieving maximized recall (combined recall: 0.928, 95% CI, 0.878–0.958), yet an average of 29.2% would be missed if maximizing precision (combined recall: 0.708, 95% CI, 0.570–0.816), which could lead to severe selection bias in systematic review. We therefore recommend that recall should take priority over precision and other indices.

Previous studies have not specified the acceptable recall level. Cohen et al assumed that a recall of 0.95 or greater might be required for the system to identify an adequate fraction of the relevant literatures, though no further evidence was given.[13] Our findings provide direct evidence proving that a large number of studies failed to achieve the recall of 0.95 even using a high-recall strategy in the model training. The combined recall was 0.928 (95% CI, 0.878–0.958) in our study. We therefore propose that 0.95 is still an important benchmark of recall for future screening automation to hold. The recall value over 0.95 indicates fewer eligible literature missing, though the bias could not be completely eliminated.

When a high recall is achieved, the secondary goal of training is to improve precision or specificity to decrease the false negative identification, as well as to save the work to review every literature. According to our results, the combined specificity and precision were 0.647 (95% CI, 0.442–0.809) and 0.200 (95% CI, 0.135–0.287) when achieving maximized recall. We first reveal the reference range of specificity and precision in medical automatic screening. It is useful for algorithm engineers to know the general performance of previous models when selecting algorithms, tuning hyper-parameters, and setting thresholds. It also enables medical evidence experts to have an intuitive understanding of the application of the automatic screening system. The low ranges of specificity and precision indicate that more newly adjusted algorithms are required for efficiency improvement in literature screening.

In addition, a large number of studies did not report their diagnostic accuracy indices normatively, affecting the interpretation of

Table 2. Combined effectiveness indices of subgroup analyses

Analysis	Number of studies	Recall/Se (95% CI)	*P* for subgroup difference	Specificity (95% CI)	*P* for subgroup difference	Precision/PPV (95% CI)	*P* for subgroup difference
Subgroups according to algorithms when achieving maximized recall							
Other	7	0.911 (0.819–0.959)	.614	0.720 (0.435–0. 896)	.449	0.243 (0.142–0.384)	.304
SVM	8	0.935 (0.624–0.992)		0.576 (0.073–0.959)		0.165 (0.039–0.491)	
Subgroups according to algorithms when achieving maximized precision							
Other	10	0.729 (0.554–0.854)	.657	0.917 (0.772–0. 973)	.901	0.419 (0.139–0.525)	.220
SVM	7	0.671 (0.216–0.938)		0.926 (0.374–0.996)		0.528 (0.265–0.776)	
Subgroups according to the number of literatures when achieving maximized recall							
≤338[a]	8	0.908(0.792–0.963)	.739	0.620 (0.341–0.837)	.783	0.249 (0.150–0.384)	.196
>338	7	0.925 (0.571–0.991)		0.673 (0.109–0.972)		0.155 (0.038–0.458)	
Subgroups according to the number of literatures when achieving maximized precision							
≤606[a]	9	0.771 (0.598–0.884)	.229	0.844 (0.634–0.944)	.056	0.479 (0.360–0.601)	.648
>606	8	0.623 (0.186–0.923)		0.964 (0.624–0.998)		0.439 (0.191–0.721)	
Subgroups according to the fraction of included literatures when achieving maximized recall							
≤0.064[a]	8	0.932 (0.853–0.970)	.969	0.760 (0.521–0.902)	.135	0.137 (0.083–0.217)	.020
>0.064	7	0.934 (0.620–0.992)		0.489 (0.064–0.930)		0.296 (0.096–0.625)	
Subgroups according to the fraction of included literatures when achieving maximized precision							
≤0.130[a]	9	0.616 (0.452–0.757)	.367	0.977 (0.945–0.991)	<.001	0.478 (0.355–0.604)	.804
>0.130	8	0.714 (0.329–0.927)		0.729 (0.220–0.963)		0.455 (0.196, 0.741)	

CI: confidence interval; PPV: positive predictive value; Se: sensitivity; SVM: support vector machines.

[a]The median was utilized for subgroup division.

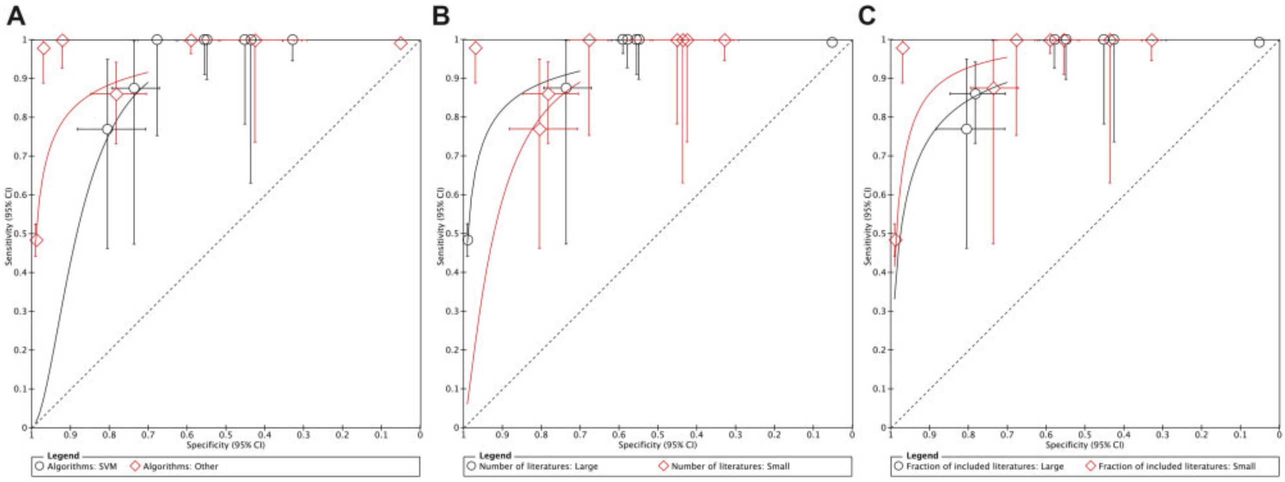

Figure 3. Summary receiver operating characteristic (SROC) plot of different subgroup analyses achieving maximized recall or sensitivity. The hollow symbols surrounded by 95% CI region (interrupted line) represent the pairs of sensitivity and specificity from the included studies; the symbols are scaled according to sample sizes of the studies. (A) Subgroup analysis based on different automatic algorithms (divided into SVM group and other algorithms group); (B) subgroup analysis based on test sets with different number of literatures (similar with the "number of participants" in traditional diagnostic accuracy study, divided by the median value of all eligible studies); (C) subgroup analysis based on test sets with different fraction of included literatures (similar with the "prevalence" in traditional diagnostic accuracy study, divided by the median).

the results. The average effectiveness indices in previous studies, such as *F*-score,[21] often give equal or fixed proportion of importance to recall, specificity, or other evaluation indicators, and thus may obscure the target indicators and severely mislead the interpretation of the results. Therefore, we recommend that the recall, precision, specificity, and other effectiveness indices should be separately reported.

Similarly, the development of AI models should cater for the real-world usage scenario. As for systematic reviews assessing the effectiveness of interventions, literature search is supposed to be conducted in multiple databases, at least including MEDLINE, Embase, and the Cochrane Library.[1] Nevertheless, the results of the automation studies included in this review have limited generalizability given that the training datasets applied by these studies were mostly MEDLINE.[13,24–29] We could find that most current popular AI algorithms or models for NLP presented superior performance in MEDLINE, yet further studies should pay more attention to the diversity of literature datasets, especially Embase and the Cochrane Library, in medical evidence syntheses. Currently, most automation studies used abstracts and metadata to train and test AI methods, it is likely that performances in recall, specificity, and precision would be improved by taking full-text screening into consideration. So far, most of the automation studies failed to integrate unsupervised learning in the training processes of AI models.[30] Although the au-

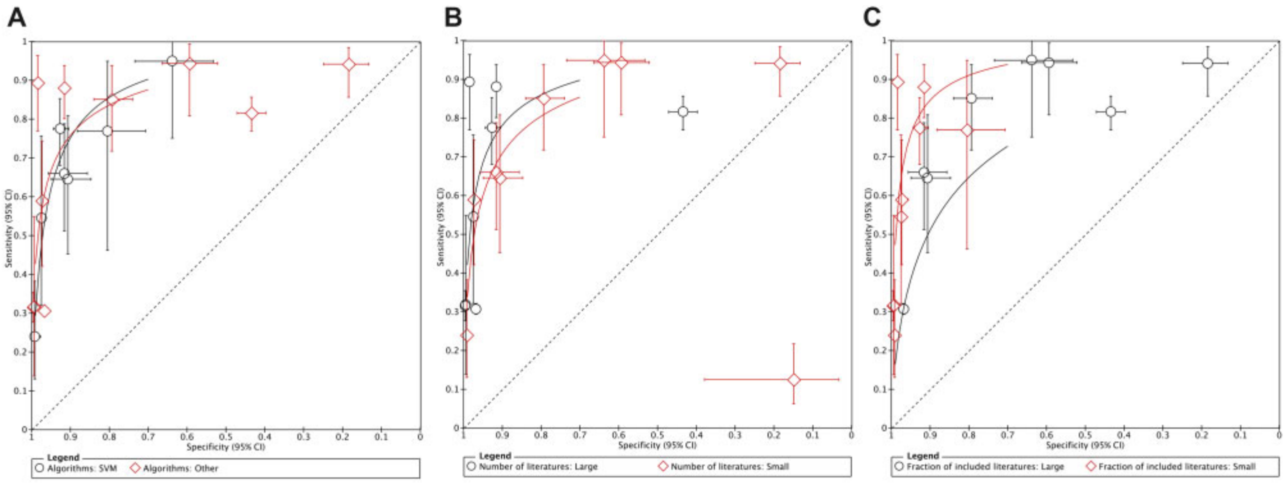

Figure 4. Summary receiver operating characteristic (SROC) plot of different subgroup analyses when achieving maximized precision or positive predictive value. The hollow symbols surrounded by 95% CI region (interrupted line) represent the pairs of sensitivity and specificity from the included studies; the symbols are scaled according to sample sizes of the studies. (A) Subgroup analysis based on different automatic algorithms (divided into SVM group and other algorithms group); (B) subgroup analysis based on test sets with different number of literatures (similar with the "number of participants" in traditional diagnostic accuracy study, divided by the median value of all eligible studies); (C) subgroup analysis based on test sets with different fraction of included literatures (similar with the "prevalence" in traditional diagnostic accuracy study, divided by the median).

tomatic methods for literature screening process would be of great help in medical evidence synthesis, the manual screening process is still indispensable at early stage of training set establishment and late stage of final exclusion.

In addition, the size of dataset is an important issue affecting model accuracy. The largest dataset in this task consisted the literatures from 30 to 50 systematic review topics.[31,32] Besides performance, the limitation on data amount may also lead to a high risk of bias. Most included studies report accuracy on only 1 systematic review. A single systematic review may focus on any specific area and could not be considered as consecutive or random samples for a task of general medical evidence synthesis. Only applying a supervised dataset in several review topics would lead to discrimination-task problems when expanding to other review topics. It may provide better reference ranges for a general screening task to report the overall recall and precision on dataset with heterogeneous sources.

We conducted subgroup analyses based on algorithm, database used in literature search, and number and fraction of included literatures. Specifically, the AI algorithms were divided into SVM and other algorithms, as current evidence showed that SVM classifiers performed well for text classification.[33–37] The algorithms and the number of screened literatures were not found to affect the accuracy of automated literature screening indicating a relatively homogeneous effectiveness. When achieving maximized recall, it is reasonable to observe higher precision (PPV) in studies with higher fraction of included literatures (similar with "prevalence" in traditional diagnostic accuracy study). As an inherent and fixed property, specificity should not be influenced by prevalence in diagnostic accuracy study. However, higher specificity was found in studies with lower fraction of included literatures with statistically significance. This finding should be interpreted with caution due to the lack of adjustment of multiple comparisons and potential confounding in the subgroup analyses, and further studies are needed to verify this subgroup difference.

Limitations

Due to diverse recall levels as well as missing reported WSS in many studies, we were unable to further analyze the work savings in automatic literature screening. There was significant heterogeneity in literature topics for investigating the screening performance of different AI algorithms, which limits the generalizability of the findings. Studies utilizing the same incompletely collected research dataset may influence the interpretation of summary outcomes due to the introduction of a higher risk of bias. Although we used traditional literature screening by human investigators as the reference standard, this reference standard is actually imperfect, since human investigators may still miss eligible literatures during screening.[38] This is also the case for some traditional diagnostic accuracy studies, for example, the diagnosis of tuberculosis.[39] Literature screening by human investigators may be the best reference standard at current stage but it is still possible that the potential misclassification weakens the reliability of our findings.

CONCLUSION

Workload reduction in automated medical literature screening has been acceptable, but the recall level of current automation studies still needs to be improved. Our findings suggest that a recall of 0.95 should be prioritized in the model training. We recommend to report recall and other indices separately rather than report average form such as *F*-score in automated medical literature screening.

FUNDING

Peking Union Medical College Hospital Research Grant for Young Scholar (pumch201912048).

AUTHOR CONTRIBUTIONS

All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by YF, SL, YZ, and TH. The first draft of the manuscript was written by YF, SL, YZ, and SC. All authors read and approved the final manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

The data used and analyzed during the current study are available from the corresponding author on reasonable request.

REFERENCES

1. Higgins J, Thomas J, Chandler J, *et al. Cochrane Handbook for Systematic Reviews of Interventions Version 6.0* (updated July 2019). Cochrane; 2019. www.training.cochrane.org/handbook. Accessed August 5, 2020.
2. Armstrong R, Hall BJ, Doyle J, *et al.* Scoping the scope' of a Cochrane review. *J Public Health* 2011; 33 (1): 147–50.
3. Sampson M, Shojania KG, Garritty C, *et al.* Systematic reviews can be produced and published faster. *J Clin Epidemiol* 2008; 61 (6): 531–6.
4. Bragge P, Clavisi O, Turner T, *et al.* The global evidence mapping initiative: scoping research in broad topic areas. *BMC Med Res Methodol* 2011; 11 (1): 92.
5. Bashir R, Surian D, Dunn AG. Time-to-update of systematic reviews relative to the availability of new evidence. *Syst Rev* 2018; 7 (1): 1–8.
6. Bishop CM. *Pattern Recognition and Machine Learning*. New York, NY: Springer; 2006.
7. Tsafnat G, Glasziou P, Choong MK, *et al.* Systematic review automation technologies. *Syst Rev* 2014; 3 (1): 74.
8. Millard LA, Flach PA, Higgins JP. Machine learning to assist risk-of-bias assessments in systematic reviews. *Int J Epidemiol* 2016; 45 (1): 266–77.
9. Marshall IJ, Kuiper J, Wallace BC. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *J Am Med Inform Assoc* 2016; 23 (1): 193–201.
10. Page MJ, McKenzie JE, Bossuyt PM, *et al.* Updating guidance for reporting systematic reviews: development of the PRISMA 2020 statement. *J Clin Epidemiol* 2021; 134: 103–12.
11. Page MJ, Moher D, Bossuyt PM, *et al.* PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ* 2021; 372: n160.
12. Jonnalagadda SR, Goyal P, Huffman MD. Automating data extraction in systematic reviews: a systematic review. *Syst Rev* 2015; 4 (1): 78.
13. Cohen AM, Hersh WR, Peterson K, *et al.* Reducing workload in systematic review preparation using automated citation classification. *J Am Med Inform Assoc* 2006; 13 (2): 206–19.
14. Enhancing academic literature review through relevance recommendation: using bibliometric and text-based features for classification. In: 2016 11th Iberian Conference on Information Systems and Technologies (CISTI); 2016; IEEE; Gran Canaria, Canary Islands, Spain.
15. McInnes MD, Moher D, Thombs BD, *et al.*; and the PRISMA-DTA Group. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. *JAMA* 2018; 319 (4): 388–96.
16. Whiting PF, Rutjes AW, Westwood ME, *et al.*; QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011; 155 (8): 529–36.
17. Reitsma JB, Glas AS, Rutjes AW, *et al.* Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 2005; 58 (10): 982–90.
18. Doebler P, Holling H. Meta-analysis of diagnostic accuracy with mada. R Packag, vol. 1, p. 15; 2015.
19. Nordic Cochrane Centre TCC. Nordic Cochrane Centre, The Cochrane Collaboration. Review Manager 5 (RevMan 5). Version 5.3; 2014.
20. Ioannidis A. An analysis of a BERT deep learning strategy on a technology assisted review task. *arXiv preprint arXiv:2104.08340*; 2021.
21. Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*; 2020.
22. Altman DG, Bland JM. Statistics notes: diagnostic tests 2: predictive values. *BMJ* 1994; 309 (6947): 102.
23. Altman DG, Bland JM. Diagnostic tests. 1: sensitivity and specificity. *BMJ* 1994; 308 (6943): 1552.
24. Cohen AM, Ambert K, McDonagh M. Studying the potential impact of automated document classification on scheduling a systematic review update. *BMC Med Inform Decis Making* 2012; 12 (1): 33.
25. Dalal SR, Shekelle PG, Hempel S, *et al.* A pilot study using machine learning and domain knowledge to facilitate comparative effectiveness review updating. *Med Decis Making* 2013; 33 (3): 343–55.
26. Leveraging advanced analytics techniques for medical systematic review update. In: 2015 48th Hawaii International Conference on System Sciences; 2015; IEEE; Kauai, HI.
27. Using semi-supervised learning for the creation of medical systematic review: an exploratory analysis. In: 2016 49th Hawaii International Conference on System Sciences (HICSS); 2016; IEEE; Kauai, HI.
28. Saha TK, Ouzzani M, Hammady HM, *et al.* A large scale study of SVM based methods for abstract screening in systematic reviews. *arXiv preprint arXiv:1610.00192*; 2016.
29. Olorisade BK, Brereton P, Andras P. The use of bibliography enriched features for automatic citation screening. *J Biomed Inform* 2019; 94: 103202.
30. An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd international conference on Machine learning; 2006; Pittsburgh, PA.
31. Ranking and Feedback-based Stopping for Recall-Centric Document Retrieval. CLEF (working notes); 2017.
32. Seed-driven document ranking for systematic reviews in evidence-based medicine; 2018.
33. Aphinyanaphongs Y, Tsamardinos I, Statnikov A, *et al.* Text categorization models for high-quality article retrieval in internal medicine. *J Am Med Inform Assoc* 2005; 12 (2): 207–16.
34. Cohen AM. Performance of support-vector-machine-based classification on 15 systematic review topics evaluated with the WSS@ 95 measure. *J Am Med Inform Assoc* 2011; 18 (1): 104.
35. Bekhuis T, Demner-Fushman D. Screening nonrandomized studies for medical systematic reviews: a comparative study of classifiers. *Artif Intell Med* 2012; 55 (3): 197–207.
36. Adeva JG, Atxa JP, Carrillo MU, *et al.* Automatic text classification to support systematic reviews in medicine. *Expert Syst Appl* 2014; 41 (4): 1498–508.
37. Timsina P, Liu J, El-Gayar O. Advanced analytics for the automation of medical systematic reviews. *Inf Syst Front* 2016; 18 (2): 237–52.
38. Edwards P, Clarke M, DiGuiseppi C, *et al.* Identification of randomized controlled trials in systematic reviews: accuracy and reliability of screening records. *Stat Med* 2002; 21 (11): 1635–40.
39. Cassidy R, Manski CF. Tuberculosis diagnosis and treatment under uncertainty. *Proc Natl Acad Sci USA* 2019; 116 (46): 22990–97.