Methods Research Report

Assessing the Accuracy of Machine-Assisted Abstract Screening With DistillerAI: A User Study

Agency for Healthcare Research and Quality

Methods Research Report

Assessing the Accuracy of Machine-Assisted Abstract Screening With DistillerAI: A User Study

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
5600 Fishers Lane
Rockville, MD 20857
www.ahrq.gov

Contract No. 290-2015-00011-I

Prepared by:
RTI International–University of North Carolina Evidence-based Practice Center
Research Triangle Park, NC

Investigators:
Gerald Gartlehner[1,2]
Gernot Wagner[2]
Linda Lux[1]
Lisa Affengruber[2,3]
Andreea Dobrescu[2]
Angela Kaminski-Hartenthaler[2]
Meera Viswanathan[1]

[1]RTI International–University of North Carolina Evidence-based Practice Center, Research Triangle Park, NC
[2]Danube University Krems, Department for Evidence-based Medicine and Clinical Epidemiology, Krems, Austria
[3] Department of Family Medicine, Care and Public Health Research Institute, Maastricht University, Maastricht, The Netherlands

AHRQ Publication No. 19(20)-EHC026-EF
November 2019

Key Messages

Purpose of the report

This report summarizes a methods study that evaluated the accuracy of a machine-assisted abstract screening approach that temporarily replaced a human screener with a semi-automated screening tool.

Key messages

- Results of our study rendered a mean sensitivity of 78 percent and a mean specificity of 95 percent for a machine-assisted abstract screening approach involving DistillerAI.
- Findings of our study imply that the accuracy of DistillerAI is not yet adequate to replace a human screener temporarily during abstract screening.
- The approach that we tested missed too many relevant studies and created too many conflicts between human screeners and DistillerAI.
- Rapid reviews, which do not require detecting the totality of the relevant evidence, may find semi-automation tools to have greater utility than traditional reviews.

This report is based on research conducted by the RTI International–University of North Carolina Evidence-based Practice Center (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No 290-2015-00011-I). The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

The information in this report is intended to help healthcare researchers and funders of research make well informed decisions in designing and funding research and thereby improve the quality of healthcare services. This report is not intended to be a substitute for the application of scientific judgment. Anyone who makes decisions concerning the provision of clinical care should consider this report in the same way as any medical research and in conjunction with all other pertinent information, i.e., in the context of available resources and circumstances.

AHRQ or the U.S. Department of Health and Human Services endorsement of any derivative products that may be developed from this report, such as clinical practice guidelines, other quality enhancement tools, or reimbursement or coverage policies, may not be stated or implied.

Persons using assistive technology may not be able to fully access information in this report. For assistance contact EPC@ahrq.hhs.gov.

Suggested citation: Gartlehner G, Wagner G, Lux L, Affengruber L, Dobrescu A, Kaminski-Hartenthaler A, Viswanathan M. Assessing the Accuracy of Machine-Assisted Abstract Screening With DistillerAI: A User Study. Methods Research Report. (Prepared by the RTI International–University of North Carolina Evidence-based Practice Center under Contract No. 290-2015-00011-I.) AHRQ Publication No. 19(20)-EHC026-EF. Rockville, MD: Agency for Healthcare Research and Quality; November 2019. Posted final reports are located on the Effective Health Care Program search page.
DOI: https://doi.org/10.23970/AHRQEPCMETHMACHINEDISTILLER.

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of healthcare in the United States. The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new healthcare technologies and strategies. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To improve the scientific rigor of these evidence reports, AHRQ supports empiric research by the EPCs to help understand or improve complex methodologic issues in systematic reviews. These methods research projects are intended to contribute to the research base in and be used to improve the science of systematic reviews. They are not intended to be guidance to the EPC program, although may be considered by EPCs along with other scientific research when determining EPC program methods guidance.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers as well as the healthcare system as a whole by providing important information to help improve healthcare quality. The reports undergo peer review prior to their release as a final report.

If you have comments on this Methods Research Project they may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 5600 Fishers Lane, Rockville, MD 20857, or by email to epc@ahrq.hhs.gov.

Gopal Khanna, M.B.A.
Director
Agency for Healthcare Research and Quality

Arlene Bierman, M.D., M.S.
Director
Center for Evidence and Practice Improvement
Agency for Healthcare Research and Quality

Stephanie Chang, M.D., M.P.H.
Director
Evidence-based Practice Center Program
Center for Evidence and Practice Improvement
Agency for Healthcare Research and Quality

Aysegul Gozu M.D., M.P.H.
Task Order Officer
Center for Evidence and Practice Improvement
Agency for Healthcare Research and Quality

Acknowledgments

The authors gratefully acknowledge the following individuals for their contributions to this project: Loraine Monroe for formatting and Sharon Barrell for editing.

Peer Reviewers

Prior to publication of the final report, EPCs sought input from independent Peer Reviewers without financial conflicts of interest. However, the conclusions and synthesis of the scientific literature presented in this report does not necessarily represent the views of individual reviewers.

Peer Reviewers must disclose any financial conflicts of interest greater than $10,000 and any other relevant business or professional conflicts of interest. Because of their unique clinical or content expertise, individuals with potential non-financial conflicts may be retained. The TOO and the EPC work to balance, manage, or mitigate any potential non-financial conflicts of interest identified.

The list of Peer Reviewers follows:

Alexandra Bannach-Brown, Ph.D.
Institute for Evidence-Based Practice
Bond University
Gold Coast, Queensland, Australia

Hans Lund, Ph.D.
Centre for Evidence-Based Practice
Western Norway University of Applied Sciences
Bergen, Norway

Assessing the Accuracy of Machine-Assisted Abstract Screening With DistillerAI: A User Study

Structured Abstract

Background. Web applications that employ natural language processing technologies such as text mining and text classification to support systematic reviewers during abstract screening have become more user friendly and more common. Such semi-automated screening tools can increase efficiency by reducing the number of abstracts needed to screen or by replacing one screener after adequately training the algorithm of the machine. Savings in workload between 30 percent and 70 percent might be possible with the use of such tools. The goal of our project was to conduct a case study to explore a screening approach that temporarily replaces a human screener with a semi-automated screening tool.

Methods. To address our objective, we evaluated the accuracy of a machine-assisted screening approach using an Agency for Healthcare Research and Quality comparative effectiveness review as the reference standard. We chose DistillerAI as a semi-automated screening tool for our project, applying its naïve Bayesian machine-learning option. Five teams screened the same 2,472 abstracts in parallel, using the machine-assisted approach. Each team trained DistillerAI with 300 randomly selected abstracts that the team screened dually. For the remaining 2,172 abstracts, DistillerAI replaced one human screener in each team and provided predictions about the relevance of records. We used a prediction score of 0.5 (i.e., inconclusive) or greater to classify a record as an inclusion. A single reviewer also screened all remaining abstracts. A second human screener resolved conflicts between the single reviewer and DistillerAI. We compared the decisions of the machine-assisted approach, single-reviewer screening (i.e., no machine assistance), and screening with DistillerAI alone (i.e., no human involvement after training) against the reference standard and calculated sensitivities, specificities, and the area under the receiver operating characteristics curve. In addition, we determined the interrater agreement, the proportion of included abstracts, and the number of conflicts between human screeners and DistillerAI.

Results. The mean sensitivity of the machine-assisted screening approach across the five screening teams was 78 percent (95% confidence interval [CI], 66% to 90%), and the mean specificity was 95 percent (95% CI, 92% to 97%). By comparison, the sensitivity of single-reviewer screening was also 78 percent (95% CI, 66% to 89%); the sensitivity of DistillerAI alone was 14 percent (95% CI, 0% to 31%). Specificities for single-reviewer screening and DistillerAI alone were 94 percent (95% CI, 91% to 97%) and 98 percent (95% CI, 97% to 100%), respectively. Machine-assisted screening and single-reviewer screening had similar areas under the curve (0.87 and 0.86, respectively); by contrast, the area under the curve for DistillerAI alone was just slightly better than chance (0.56). The interrater agreement between human screeners and DistillerAI with a prevalence-adjusted kappa was 0.85 (95% CI, 0.84 to 0.86).

Discussion. Findings of our study indicate that the accuracy of DistillerAI is not yet adequate to replace a human screener temporarily during abstract screening. The approach that we tested missed too many relevant studies and created too many conflicts between human screeners and

DistillerAI. Rapid reviews, which do not require detecting the totality of the relevant evidence, may find semi-automation tools to have greater utility than traditional systematic reviews.

Contents

Tables

Figures

Background

A crucial step in any systematic review is the selection of relevant abstracts. To reduce the risk of falsely excluding relevant studies, methodological guidance recommends a dual-screening process.[1, 2] Two reviewers independently determine the eligibility of each record based on a predetermined list of inclusion and exclusion criteria. In its landmark document *Finding What Works in Healthcare: Standards in Systematic Reviews*, the U.S. Institute of Medicine explicitly favors high sensitivity of literature searches and literature screening over high specificity.[3]

Screening titles and abstracts, however, is a lengthy and labor-intensive process. Systematic reviewers often need to screen thousands of irrelevant abstracts to identify a few relevant studies. A cost-effectiveness analysis estimated that screening 5,000 references takes 83 to 125 hours per reviewer for abstract and full-text review at a cost of approximately £13,000 (2013 prices; about $17,000).[4]

In recent years, Web applications that employ natural language processing technologies such as text mining and active learning to support systematic reviewers during abstract screening have become more user friendly and more common. In 2015, a systematic review by O'Mara-Eves and colleagues identified 44 studies addressing the use of text mining to reduce the screening workload in systematic reviews.[5] Commonly used tools that systematic reviewers can use without additional programming include Abstrackr,[6] DistillerAI,[7] EPPI (Evidence for Policy and Practice Information) Reviewer,[8] RobotAnalyst,[9] Rayyan,[10] and SWIFT (Sciome Workbench for Interactive computer-Facilitated Text-mining)-Review.[11] These text-mining approaches use pattern-recognition algorithms to predict the probabilities of record relevance or irrelevance. Text mining describes the process of filtering knowledge from unstructured data such as text. In the context of abstract screening, text mining is combined with text classification, which is the decision about the inclusion or exclusion of a given record.[12, 13] Applications that combine text mining with machine learning have the advantage of improving the system's performance continuously. Active learning is a special case of machine learning in which the algorithm chooses the data it learns from. Consequently, the machine continuously adapts its decision rules based on the human screeners' decisions.

Such semi-automated screening tools can increase efficiency by reducing the number of abstracts needed to screen or by replacing one screener after adequately training the algorithm of the machine.[14] Savings in workload between 30 percent and 70 percent might be possible with the use of text-mining tools in systematic reviews.[5] The downside of the use of such tools, however, is that none of these tools has perfect sensitivity and a reduction in workload might be accompanied by missing relevant studies.[5]

To date, several semi-automated screening tools have been validated.[9, 11, 15-17] Most research publications on this topic, however, have been produced by computer scientists and experts in medical informatics and artificial intelligence. Often studies have been conducted under highly controlled conditions using artificial bibliographic datasets. Furthermore, validation studies mostly used decisions about inclusion or exclusion at an abstract screening stage as a reference standard. Human decisions during abstract screening, however, vary and are an imperfect reference standard.

The goal of our project was to conduct a case study to explore a screening approach that temporarily replaces a human screener with a semi-automated screening tool. We were also interested in comparing the performance of this approach with that of single-reviewer screening and screening of abstracts by a semi-automated screening tool without human involvement after training the tool. Table 1 summarizes commonly used terms in this report.

Table 1. Definitions of commonly used terms

General Terms of Machine Learning

Active learning: A special case of machine learning in which the accuracy of predictions by the machine is constantly improved through interaction with reviewers.
Machine learning: The use of algorithms or statistical models (e.g., naïve Bayesian, support vector machines) based on sample data by computer systems to make predictions or decisions without being explicitly programmed to perform these tasks.[18]
Natural language processing technology: A semantic technology process that is used to normalize variants (e.g., different conjugations of a verb) of a single concept, and to identify complex concepts (e.g., terms made up of several words) in a text.[19, 20]
Prediction: A forecast of whether a record is relevant (include) or irrelevant (exclude) for a given systematic review.
Semi-automated screening/text mining tool: Any Web-based application that employs a combination of text mining and text classification to assist systematic reviewers to make decisions during the title and abstract screening process.
Text classification: A standard machine-learning process in which the aim is to categorize texts into groups of interest.[21]
Text mining: The process of discovering knowledge and structure from unstructured data.

Terms to characterize the performance of screening tools

Accuracy: The proportion of correctly classified records:

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

False negatives (FNs): The number of records incorrectly classified as excludes. Also referred to as “missed studies.”
False positives (FPs): The number of records incorrectly classified as includes.

Sensitivity: The ability of a screening tool to correctly classify *relevant* records as includes:

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

Specificity: The ability of a screening tool to correctly classify irrelevant records as excludes:

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

True negatives (TNs): The number of records correctly identified as excludes.
True positives (TPs): The number of records correctly identified as includes.

Objective

Our study had the following objective: To assess the accuracy of an abstract screening approach that temporarily replaces one human screener with a semi-automated screening tool.

Methods

To address our objective, we employed a diagnostic framework approach that assessed the accuracy of machine-assisted screening compared with a reference (gold) standard. Specifically, we used data from an Agency for Healthcare Research and Quality (AHRQ) comparative effectiveness review on pharmacological and nonpharmacological interventions for the treatment of depression as the reference standard.[22]

We chose DistillerAI as a semi-automated screening tool for our project. DistillerAI is a text mining and text classification tool within DistillerSR (www.evidencepartners.com/products/distillersr-systematic-review-software), a specialized, commercially available Web-based software to conduct systematic reviews. DistillerAI offers a naïve Bayesian approach or a support vector machine classifier to screen abstracts after learning from decisions of human screeners (training set). The naïve Bayesian approach provides probabilistic prediction scores regarding the inclusion or exclusion of records (0.5 is an inconclusive score). Prediction scores larger than 0.5 indicate a greater probability of a record being relevant rather than irrelevant; scores smaller than 0.5 indicate the opposite.

The support vector machine classifier offers nonprobabilistic, binary classifications (include, exclude, or can't decide). It uses data from the training set to build a model that classifies new records as relevant or irrelevant. We chose DistillerAI as a screening tool for our project because it provides optimal flexibility regarding data import and export and an efficient technical helpline.

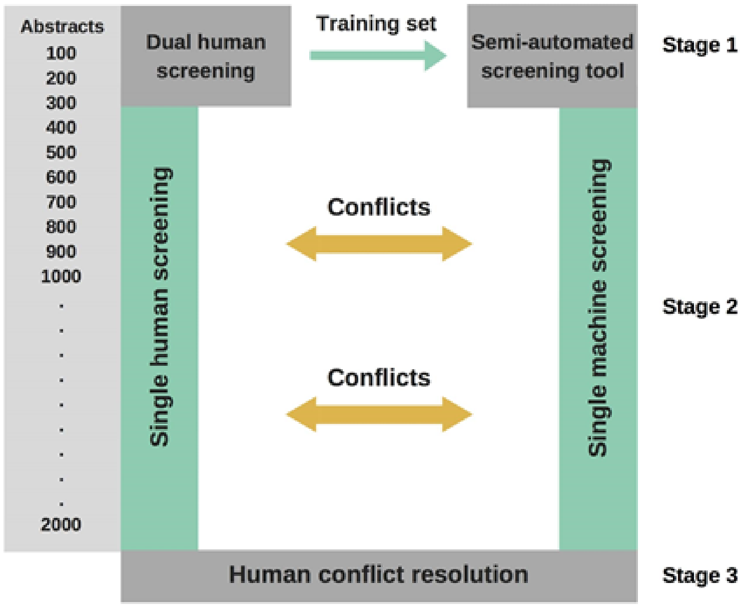
Reference Standard

As mentioned above, we used a comparative effectiveness review on pharmacological and nonpharmacological interventions as the reference standard.[22] For the purpose of this methods project, we focused on a single Key Question, which included 42 randomized controlled trials (RCTs). Because the scope was narrower than the original review, we replicated a targeted literature search with a focus on the Key Question of interest (comparative effectiveness). We searched PubMed and Embase because we knew from a bibliographic analysis that the 42 RCTs included in the report are indexed in these databases. We adapted the original search strategy of the AHRQ report and limited searches to the same period that the report had covered (1995 to 2015).

Outline of General Approach

Figure 1 depicts the screening approach in which the semi-automated screening tool temporarily replaced one human screener. Five independent teams applied this approach in parallel on the same topic. Teams consisted of professional systematic reviewers with extensive experience in literature screening and evidence syntheses.

Figure 1. Graphical presentation of the study flow



Stage 1 mimicked a regular dual-reviewer abstract screening process. After a pilot phase with 50 records to calibrate screeners, two reviewers independently screened abstracts based on predefined inclusion and exclusion criteria. They resolved conflicts by discussing the issues and reaching consensus or by involving a third, senior reviewer. In this stage, reviewers dually and independently screened 300 records that we randomly selected from our literature searches. The dually agreed upon inclusions and exclusions served as the training set for DistillerAI.

During Stage 2, DistillerAI replaced one human screener and provided prediction scores about inclusions or exclusions for all remaining records. The second human reviewer was not aware of predictions and screened the remaining abstracts. In Stage 3, a second human reviewer resolved conflicts in decisions between the human screener and DistillerAI.

Training DistillerAI

Each of the five screening teams independently trained DistillerAI. For each team, we randomly selected 300 abstracts as training sets from the database of our literature searches. Decisions about inclusions or exclusions of records in the training sets served as information for DistillerAI to build an algorithm for predictions. The manual for DistillerAI recommends 300 records as the optimal size for training sets based on internal simulation studies.

To reduce the risk of not having any included RCTs in the training set by chance, we employed weighted sampling to ensure that each training set included at least 5 of the 42 relevant studies of the AHRQ report.

After screeners had completed the training sets, we employed DistillerAI’s test function. The test function randomly selects records from the training set to determine the accuracy of predictions by comparing prediction scores to decisions about inclusion or exclusion in the training set. For each training set, we used the test function 5 times at a ratio of 80 to 20 (i.e., DistillerAI learns from 80% of the training set and predicts the randomly selected 20%). For all five training sets, DistillerAI’s naïve Bayesian approach provided better predictions than the support vector machine classifier. The mean accuracy score across the five training sets for the naive Bayesian approach was 87.9 percent compared with 47.6 percent for the support vector machine classifier. Consequently, we used DistillerAI’s naïve Bayesian approach for predictions for all five screening teams. Abstracts with prediction scores of 0.5 or greater were included; abstracts with prediction scores below 0.5 were excluded. A prediction score of 0.5 reflects a neutral prediction (i.e., DistillerAI cannot decide whether inclusion or exclusion is more likely). We chose a prediction score of 0.5 as a conservative threshold that would guarantee high sensitivity.

Outcomes

We assessed three outcomes:

- **Proportion of included abstracts.** This outcome provides information about the number of full texts that need to be retrieved and reviewed, which has a substantial impact on the subsequent workload during the full-text review stage. We used the number of unscreened records (n=2,172) after completion of the training set as a denominator for all calculations; in other words, we did not include results of the training sets in any of the calculations.
- **Proportion of conflicts and interrater agreement between human reviewers and DistillerAI.** This outcome summarizes the agreement and the number of conflicts between human reviewers and DistillerAI, which had to be resolved by a second human reviewer. The number of unscreened records (n=2,172) served as the denominator for all calculations. We also determined the interrater agreement (Prevalence-adjusted Bias-adjusted kappa) between human screeners and DistillerAI.
- **Accuracy of correctly classifying relevant and irrelevant studies.** We determined sensitivities of the machine-assisted screening approach, single-reviewer screening, and DistillerAI in identifying the 42 included studies of the reference standard as relevant. We also calculated specificities and areas under the receiver operating characteristics (ROC) curve.

Comparisons and Quantitative Analyses

We assessed the above-mentioned outcomes for three abstract screening approaches:

1. the machine-assisted screening approach (as outlined in Figure 1),
2. single-reviewer screening (i.e., no DistillerAI involvement), and
3. screening with DistillerAI alone (i.e., no human screener involvement after training DistillerAI).

For measures of accuracy, we organized results in 2x2 tables to determine true-positive, false-positive, true-negative, and false-negative decisions. We calculated sensitivities, specificities, and areas under the ROC curve with their 95% confidence intervals. For DistillerAI, we also calculated the ROC curve in an exploratory analysis using different prediction scores as thresholds. We conducted all quantitative analyses with Stata 13.1 (Stata Corporation, College Station, Texas, USA).

Results

Literatures searches rendered 2,472 references after deduplication. The 42 relevant randomized controlled trials (RCTs) of the reference standard compared second-generation antidepressants with nonpharmacological treatment options during acute-phase treatment of major depressive disorder. Nonpharmacological interventions included various psychotherapies, acupuncture, St. John's wort, omega-3-fatty acid, physical exercise, and S-adenosyl-L-methionine. Study durations ranged from 4 to 96 weeks. Trials took place in Brazil, Canada, China, Denmark, England, Finland, Germany, Iran, Italy, Romania, Sweden, the Netherlands, and the United States. Many of the available trials had serious methodological limitations. Authors of the reference report rated 16 of the 42 trials as high risk of bias and only 4 as low risk of bias.[22]

As described in the Methods section, each of the five teams dually screened 300 randomly selected records to provide training sets for DistillerAI. Subsequently, each team applied the machine-assisted screening approach on 2,172 records. The number of included studies (true positives) sampled into the training sets ranged from 10 to 16. In the following sections, we present results of the machine-assisted screening approach (as outlined in Figure 1) and contrast them with single-reviewer screening (i.e., no DistillerAI involvement) or screening with DistillerAI only (no human screener involvement after training DistillerAI).

Table 2 provides a summary of various performance measures. Denominators for calculations of performance measures in the table vary by screening team because they discount for relevant studies that had been sampled into the training sets.

Proportion of Included Abstracts

On average, the five screening teams using the machine-assisted approach included 8 percent (n=174) of screened abstracts (range 4% to 11% [n=87 to 239]). Single-reviewer screening, on average, included a similar proportion of abstracts as the machine-assisted approach (7% [n=152]; range 5% to 10% [n=109 to 217]). By comparison, DistillerAI, on average, rated only 2 percent (n=43; range 1% to 3% [n=22 to 65]) of screened abstracts as relevant for inclusion. The reference standard systematic review included 10 percent of screened abstracts.

Table 2. Different performance measures for the machine-assisted screening approach, single-reviewer screening, and screening with DistillerAI alone

Table 2a. Team 1

	Sensitivity (95% CI)	Specificity (95% CI)	Area Under the Curve (95% CI)	N of Missed Studies (Proportion)	N of Included Abstracts (Proportion)	N of Conflicts (Proportion)	N of Included Studies in Training Set
Machine-assisted screening	0.78 (0.59 to 0.90)	0.96 (0.96 to 0.97)	0.87 (0.80 to 0.95)	7/32 (22%)	97/2,172 (4%)	126/2,172 (6%)	10/300
Single-reviewer screening	0.78 (0.59 to 0.90)	0.96 (0.95 to 0.97)	0.87 (0.80 to 0.94)	7/32 (22%)	110/2,172 (5%)		
DistillerAI screening	0.03 (0.00 to 0.21)	0.99 (0.98 to 0.99)	0.51 (0.48 to 0.54)	31/32 (97%)	27/2,172 (1%)		

CI = confidence interval; N = number.

Table 2b. Team 2

	Sensitivity (95% CI)	Specificity (95% CI)	Area Under the Curve (95% CI)	N of Missed Studies (Proportion)	N of Included Abstracts (Proportion)	N of Conflicts (Proportion)	N of Included Studies in Training Set
Machine-assisted screening	0.89 (0.70 to 0.97)	0.92 (0.91 to 0.93)	0.90 (0.84 to 0.96)	3/27 (11%)	232 /2,172 (11%)	226/2,172 (10%)	15/300
Single-reviewer screening	0.89 (0.69 to 0.97)	0.91 (0.89 to 0.92)	0.90 (0.84 to 0.96)	3/27 (11%)	221/2,172 (10%)		
DistillerAI screening	0.00	0.99 (0.99 to 0.99)	0.50 (0.49 to 0.50)	27/27 (100%)	18/2,172 (1%)		

CI = confidence interval; N = number.

Table 2c. Team 3

	Sensitivity (95% CI)	Specificity (95% CI)	Area Under the Curve (95% CI)	N of Missed Studies (Proportion)	N of Included Abstracts (Proportion)	N of Conflicts (Proportion)	N of Included Studies in Training Set
Machine-assisted screening	0.65 (0.44 to 0.82)	0.96 (0.95 to 0.97)	0.81 (0.71 to 0.90)	9/26 (35%)	130/2,172 (6%)	100/2,172 (5%)	16/300
Single-reviewer screening	0.65 (0.44 to 0.82)	0.96 (0.95 to 0.97)	0.81 (0.71 to 0.90)	9/26 (35%)	104/2,172 (5%)		
DistillerAI screening	0.23 (0.10 to 0.44)	0.99 (0.98 to 0.99)	0.61 (0.53 to 0.69)	20/26 (77%)	30/2,172 (1%)		

CI = confidence interval; N = number.

Table 2d. Team 4

	Sensitivity (95% CI)	Specificity (95% CI)	Area Under the Curve (95% CI)	N of Missed Studies (Proportion)	N of Included Abstracts (Proportion)	N of Conflicts (Proportion)	N of Included Studies in Training Set
Machine-assisted screening	0.86 (0.66 to 0.95)	0.94 (0.93 to 0.95)	0.90 (0.83 to 0.96)	4/28 (14%)	199/2,172 (9%)	194/2,172 (9%)	14/300
Single-reviewer screening	0.82 (0.62 to 0.93)	0.93 (0.92 to 0.94)	0.88 (0.80 to 0.95)	5/28 (18%)	165/2,172 (8%)		
DistillerAI screening	0.32 (0.17 to 0.52)	0.97 (0.96 to 0.98)	0.65 (0.56 to 0.73)	19/28 (68%)	69/2,172 (3%)		

CI = confidence interval; N = number.

Table 2e. Team 5

	Sensitivity (95% CI)	Specificity (95% CI)	Area Under the Curve (95% CI)	N of Missed Studies (Proportion)	N of Included Abstracts (Proportion)	N of Conflicts (Proportion)	N of Included Studies in Training Set
Machine-assisted screening	0.74 (0.55 to 0.87)	0.95 (0.94 to 0.96)	0.84 (0.77 to 0.92)	8/31 (26%)	187/2,172 (9%)	181/2,172 (8%)	11/300
Single-reviewer screening	0.74 (0.55 to 0.87)	0.95 (0.94 to 0.95)	0.84 (0.77 to 0.92)	8/31 (26%)	138/2,172 (6%)		
DistillerAI screening	0.13 (0.05 to 0.31)	0.97 (0.96 to 0.98)	0.55 (0.49 to 0.61)	27/31 (87%)	65/2,172 (3%)		

CI = confidence interval; N = number.

Table 2f. Combined

	Sensitivity (95% CI)	Specificity (95% CI)	Area Under the Curve (95% CI)	N of Missed Studies (Proportion)	N of Included Abstracts (Proportion)	N of Conflicts (Proportion)	N of Included Studies in Training Set
Machine-assisted screening	0.78 (0.66 to 0.90)	0.95 (0.92 to 0.97)	0.87 (0.83 to 0.90)	6/30 (22%)	8%	165/2,172 (8%)	13/300
Single-reviewer screening	0.78 (0.66 to 0.89)	0.94 (0.91 to 0.97)	0.86 (0.82 to 0.89)	6/30 (22%)	7%		
DistillerAI screening	0.14 (0.00 to 0.31)	0.98 (0.97 to 1.00)	0.56 (0.53 to 0.59)	25/30 (86%)	2%		

CI = confidence interval; N = number.

Proportion of Conflicts and Interrater Agreement Between Human Screeners and DistillerAI

Across the five screening teams, decisions about inclusion or exclusion resulted in conflicts between the human screeners and DistillerAI in 8 percent (n=174; range 5% to 10% [n=109 to 217]) of screened abstracts. In the majority of cases, the second human reviewers who resolved these conflicts confirmed the decisions of the human screeners. The interrater agreement between human screeners and DistillerAI with a prevalence-adjusted kappa was 0.85 (95% confidence interval [CI], 0.84 to 0.86).

Accuracy of Correctly Classifying Relevant and Irrelevant Studies

The most important outcome for the assessment of the performance of the machine-assisted screening approach is the sensitivity to correctly identify the 42 included studies of the reference standard review. The combined sensitivity of the machine-assisted screening approach was 78 percent (95%CI, 66% to 90%). In other words, the machine-assisted screening approach missed, on average, 22 percent of relevant studies. Of the 42 included studies of the reference standard review, the machine-assisted screening teams collectively missed 23 studies at least once (false-negative decisions; see Table 3). Figure 2 contrasts the sensitivities of the machine-assisted screening approach with those of single-reviewer screening and DistillerAI without human involvement. Overall, sensitivities of the machine-assisted approach and single-reviewer screening were substantially higher than the sensitivity of DistillerAI (78% vs. 78% vs. 14%; Figure 2 and Table 1). On average, the machine-assisted screening approach and single-reviewer screening missed 22 percent of relevant studies compared with 86 percent of relevant studies that DistillerAI missed.

The specificity of the machine-assisted screening approach was 95 percent (95% CI, 92% to 97%). Specificities were similar between the machine-assisted approach, single-reviewer screening, and DistillerAI (95% vs. 94% vs. 98%; Figure 3).

Table 2 also presents the areas under the curve, which summarizes the discriminative abilities of the approaches to distinguish relevant from irrelevant records. Machine-assisted screening and single-reviewer screening had similar areas under the curve (0.87 and 0.86, respectively); by contrast, DistillerAI was just slightly better than chance (0.56).

Table 3. Characteristics of studies that machine-assisted screening teams missed at least once

Author and Year	Intervention	Sample Size, Risk of Bias	Falsely Excluded by:				
			Team 1	Team 2	Team 3	Team 4	Team 5
Barber et al., 2012[23]	Psychotherapy	N=106, Medium				X	
Bastos et al., 2013[24]	Psychotherapy	N=272, Medium				X	X
Blom 2007[25]	Psychotherapy	N=207, Medium		X			
Blumenthal et al., 2007[26]	Exercise	N=153, Medium			X		
Frank et al., 2011[27]	Psychotherapy	N=318, High	X		X		
Gastpar et al., 2005[28]	St. John’s wort	N=241, Medium		X			
Gertsik et al., 2012[29]	Omega-3 fatty acid augmentation of citalopram treatment	N=42, High				X	
Hegerl et al., 2010[30]	Psychotherapy	N=48, Medium					X
Huang et al., 2005[31]	Electro-scalp acupuncture	N=98, Medium	X	X	X		
Jazayeri et al., 2008[32]	Omega-3 fatty acid eicosapentaenoic acid	N=48, High					X
Kennedy et al., 2007[33]	Psychotherapy	N=31, High	X				X
Lam et al., 2013[34]	Psychotherapy	N=80, Medium		X	X		X
McGrath et al., 2013[35]	Psychotherapy	N=82, High	X			X	X
Menchetti et al., 2014[36]	Psychotherapy	N=287, Medium			X		X
Mischoulon et al., 2014[37]	Eicosapentaenoic acid	N =189, High			X		
Mynors-Wallis et al., 2000[38]	Psychotherapy	N=151, Medium			X		
Raue et al., 2009[39]	Psychotherapy	N=60, High	X	X			X
Schrader et al., 2000[40]	St. John’s wort	N=106, Medium					X
Segal et al., 2006[41]	Psychotherapy	N=301, High	X				
Song et al., 2007[42]	Electroacupuncture	N=90, High	X		X		
Zhang et al., 2009[43]	Acupuncture	N=80, Medium			X		

N = number of study participants.

Figure 2. Sensitivities of machine-assisted screening, single-reviewer screening, and DistillerAI screening

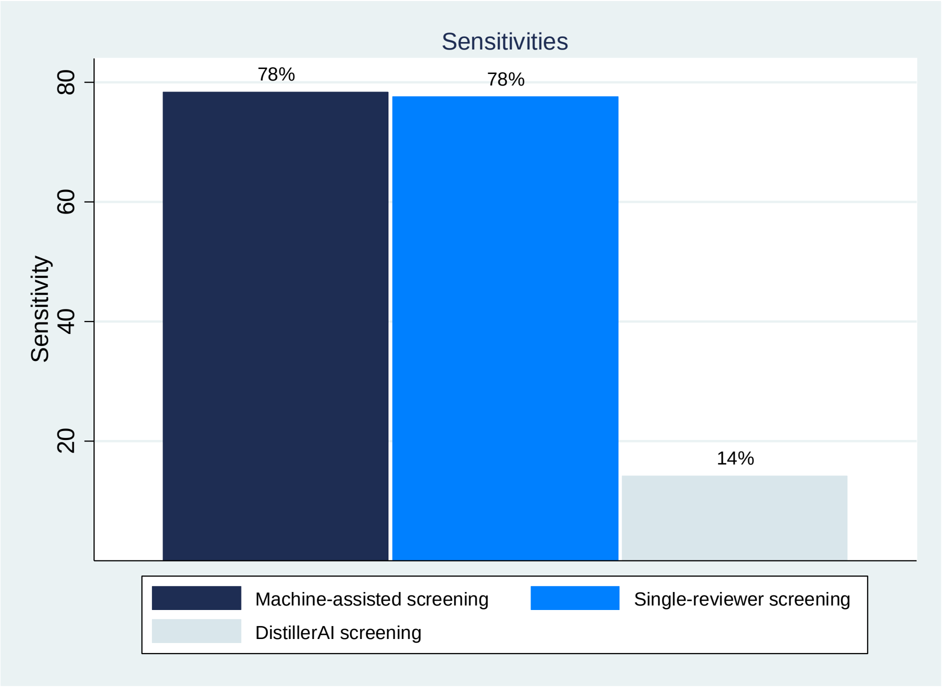
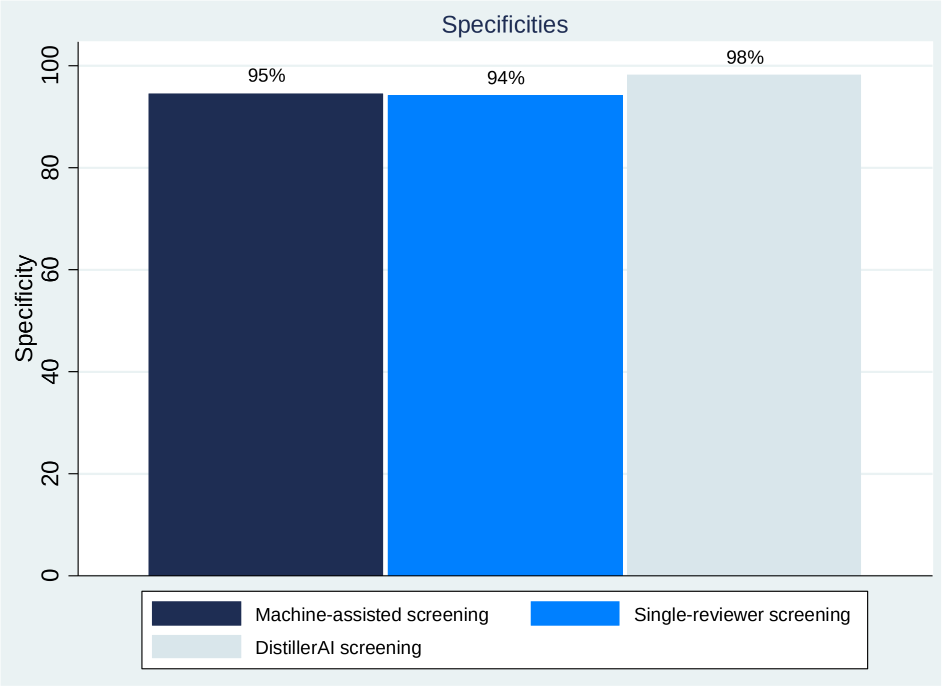


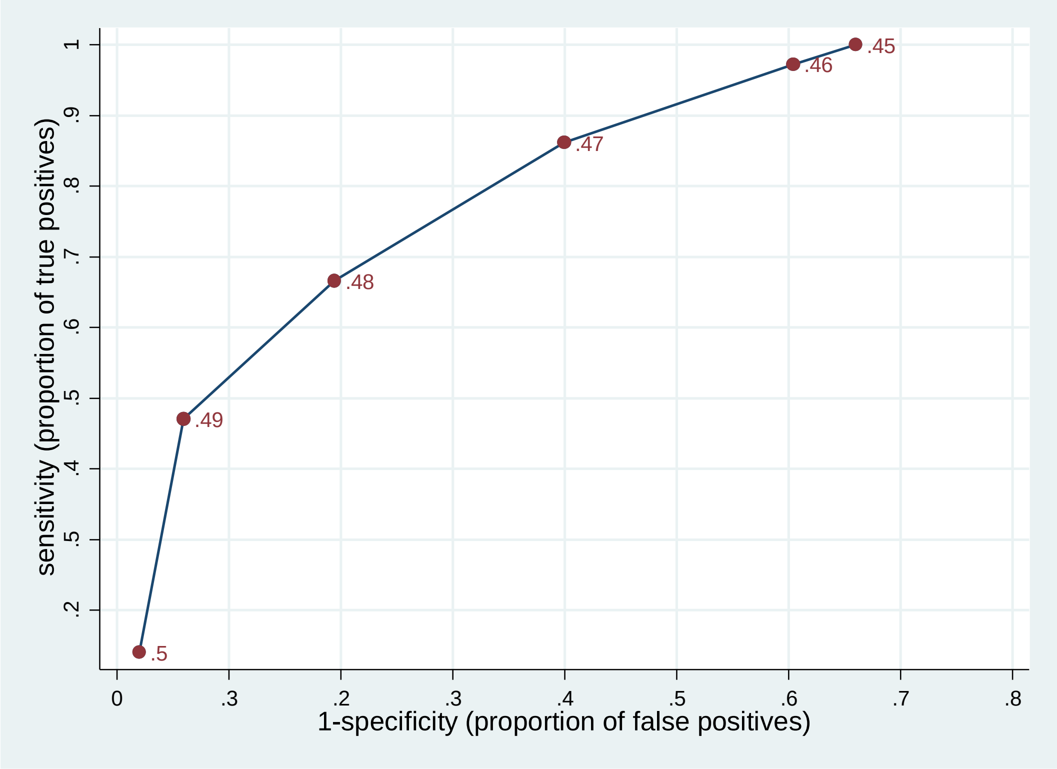
Figure 3. Specificities of machine-assisted screening, single-reviewer screening, and DistillerAI screening



Performance of DistillerAI for Different Prediction Thresholds

Because of the poor performance of DistillerAI with a threshold of 0.5, we further explored the accuracy of DistillerAI for thresholds below 0.5. Prediction scores below 0.5 indicate a greater probability that a record is irrelevant than relevant. Figure 4 presents the receiver operating characteristics curve for DistillerAI for prediction scores between 0.5 and 0.45. To achieve a sensitivity close to 100 percent, the specificity would have to be reduced to 35 percent using a prediction score of 0.45. In other words, based on our sample, DistillerAI would have to include 65 percent of all abstracts to detect all relevant studies that were included in the reference standard review.

Figure 4. Receiver operating characteristics curve for DistillerAI



Discussion

The objective of our methods study was to assess the accuracy of a machine-assisted abstract screening approach that temporarily replaces a human reviewer with a semi-automated screening tool (DistillerAI). Results of our project rendered a mean sensitivity of 78 percent and a mean specificity of 95 percent for this approach. The area under the receiver operating characteristics (ROC) curve was 0.87.

Although the area under the ROC curve indicates adequate discriminative ability, the performance of the machine-assisted abstract screening approach is less than optimal for use in systematic reviews. During abstract screening in systematic reviews, false-negative decisions (i.e., excluding relevant records) are more consequential than false-positive decisions (i.e., including irrelevant records). The subsequent full-text review will rectify false-positive decisions without consequences for the validity of a systematic review. By contrast, false-negative decisions might cause relevant records to be omitted, which could affect the validity of a systematic review. A machine-assisted screening approach that misses 22 percent of relevant studies, therefore, is not adequate for systematic reviews.

Several factors might have contributed to the poor sensitivity of the machine-assisted screening approach in our study. First, the choice of the topic probably had a substantial impact on the performance of the approach. The comparative effectiveness of pharmacological and nonpharmacological treatments comprises a wide spectrum of interventions, particularly of nonpharmacological interventions. The Cochrane Common Mental Disorders group, for example, lists more than 80 psychological interventions for the treatment of depression. A less complex topic might have led to a better performance of DistillerAI and different conclusions. Systematic reviews, however, are often multifaceted and complex. Using an unrealistically simple topic or an artificially clean dataset might have overestimated the performance under real-world conditions. Second, many of the published studies, particularly on complementary and alternative treatments, were conducted in countries where English is not the native language. Some of these abstracts were difficult to understand and interpret, which was also a contributing factor to the screening teams dually and falsely excluding five relevant studies during the screening of training sets. A partially incorrect training set is not an optimal precondition for testing the performance of machine-assisted abstract screening but might reflect real-life conditions. Nevertheless, incorrect decisions of human screeners had no apparent impact on the sensitivity of DistillerAI. For example, the team with the highest sensitivity of DistillerAI (Team 4: 0.32) falsely excluded two out of 14 relevant studies in the training set. In screening teams without false-negative decisions in the training set, sensitivities ranged from 0.03 to 0.23 (see Table 2). Third, we adhered to DistillerAI's recommendation regarding the optimal sample size for training sets (n=300). This recommendation is based on simulation studies and might have been too small to adequately train DistillerAI for our topic. The small training sets might also explain why the naïve Bayesian approach consistently provided better results than the support vector machine classifier.

Taken together, these issues might have contributed to a machine-learning phenomenon called "hasty generalizations." This term describes situations in which the training set is not fully representative of the remaining records.[5] Given the broad and complex topic, hasty generalizations might have played a role despite the attempt to ensure generalizability of the training sets with random sampling.

The performance of DistillerAI, in general, was disappointing. The average sensitivity was 0.14; in one case DistillerAI missed all relevant studies. Adding DistillerAI to single-reviewer screening did not provide additional gains in accuracy but instead created conflicts between human screeners and DistillerAI in 5 percent to 10 percent of records. These conflicts had to be resolved by a second human screener, which required effort without a gain in accuracy. In other words, DistillerAI did not improve the proportion of incorrect decisions that human screeners made when they screened abstracts. The ROC curve of DistillerAI implies that lowering the prediction threshold to 0.45 would have achieved a sensitivity close to 100 percent. With a prediction score of 0.45, however, the specificity would have decreased to 35 percent, which in turn would have caused a substantial increase in the number of conflicts between human screeners and Distiller AI because DistillerAI would have included about 65 percent of abstracts.

Our study has several strengths and weaknesses. A strength is that we used five teams who screened the same abstracts in parallel. Using five screening teams mitigated errors and subjective decisions of individual screeners, as well as the influence of screening experience and content expertise on results. Another strength of our study is that we mimicked a real-world abstract screening situation, including unintended incorrect decisions that human screeners made when they reviewed the training sets. To minimize selection bias, we randomly selected records for the training sets. Such an approach reflects real-world conditions under which machine-assisted screening would take place. We purposely did not use decisions from the reference standard dataset to train DistillerAI. The final included and excluded studies of a systematic review are the results of a process that leverages more than decisions of two screeners. The final body of evidence is also a result of feedback from the review team, review of reference lists of other systematic reviews, and comments from external peer reviewers. Finally, the choice of our reference standard is also a strength of our study. Our reference standards were the final included and excluded studies and not decisions during title and abstract screening of the reference review. Decisions during abstract screening are an insufficient reference standard because screening decisions among screening teams can vary substantially. It is conceivable that a semi-automated screening tool makes more precise screening decisions than human screeners make but would end up with inferior accuracy because of the imperfect reference standard.

A weakness of our study is that we employed a focused, stepwise literature search to recreate the evidence base for one Key Question of a systematic review. In other words, we knew from the outset which studies were relevant for the topic and tailored the searches accordingly. Our searches, therefore, presumably produced less noise than a regular systematic literature search. The spectrum and the ratio of relevant and irrelevant records were most likely different than those in a de novo systematic literature search. Furthermore, when we calculated accuracy measures, we assumed that falsely excluded studies would be missed by the review. In reality, a systematic review has subsequent processes in place that can detect incorrectly excluded records later during the review process, such as review of reference lists of other systematic reviews or external peer review. It is conceivable that some of the studies missed during abstract screening would ultimately still be included in the final systematic review. Finally, although our outcome measures provide a comprehensive picture of the accuracy and the performance of the machine-assisted screening approach, they also have limitations. We do not know whether falsely excluded studies would change the conclusions of the systematic review. This is particularly relevant for users of rapid reviews who are willing to accept that the review misses relevant studies. For them, it is more important whether conclusions would change because of missed studies. A recent international survey showed that decision makers are willing to accept up to 10 percent of incorrect conclusions in exchange for a rapid evidence product.[44]

In a recent commentary, O’Connor and colleagues explored reasons for the slow adoption of automation tools.[45] They argue that the adoption of such tools requires credible evidence that automation tools are noninferior or even superior in accuracy compared with standard practice. Our study provides evidence that noninferiority is clearly not the case yet for DistillerAI. Few other studies have assessed semi-automated screening tools under real-world conditions.[5, 15, 16] Results of these studies are consistent with our findings that semi-automated screening tools have the potential for expediting reviews but that the accuracy is still limited.[15, 16]

Future studies need to explore whether semi-automated screening tools could prove useful in identifying records that are clearly not relevant, which is a different approach than we took in our study. Future studies also need to assess the comparative accuracy of different screening tools under pragmatic, real-world screening situations. A still unanswered question is also how semi-automated screening tools perform when used with abbreviated literature searches that have a higher specificity than comprehensive systematic literature searches. Waffenschmidt et al., for example, proposed an abbreviated search strategy for randomized controlled trials.[46] This approach combines a simple-structured Boolean search in PubMed with searches using the “similar articles” function in PubMed. In a case study, this approach reduced the number of abstracts that needed to be screened by up to 90 percent without missing studies that would have changed conclusions.[20] It is conceivable that such a targeted literature search approach could improve the performance of semi-automated screening tools because they would have to deal with less noise.

Conclusions

Systematic reviews require substantial human effort for often repetitive and labor-intensive tasks. Automation to assist reviewers during systematic reviews becomes increasingly viable. Findings of our study imply that the accuracy of DistillerAI is not yet adequate to replace a human screener temporarily during abstract screening. The approach that we tested missed too many relevant studies and created too many conflicts between human screeners and DistillerAI. Rapid reviews, which do not require detecting the totality of the relevant evidence, may find semi-automation tools to have greater utility than traditional reviews.

References

1. Effective Health Care Program. Methods guide for effectiveness and comparative effectiveness reviews. Agency for Healthcare Research and Quality. AHRQ publication no. 10(14)-EHC063-EF. Rockville, MD: 2014. Chapters available at: www.effectivehealthcare.ahrq.gov. January.

2. Methods Group of the Campbell Collaboration. Methodological expectations of Campbell Collaboration intervention reviews: conduct standards. Campbell Policies and Guidelines Series No. 3. Oslo, Norway: Campbell Collaboration; 2017. https://www.campbellcollaboration.org/library/campbell-methods-conduct-standards.html. Accessed on October 1 2018.

3. Institute of Medicine of the National Academies. Finding what works in health care: standards for systematic reviews. Washington, DC: Institute of Medicine of the National Academies; 2011.

4. Shemilt I, Khan N, Park S, et al. Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. Syst Rev. 2016 Aug 17;5(1):140. doi: 10.1186/s13643-016-0315-4. PMID: 27535658.

5. 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;4:5. doi: 10.1186/2016-4053-4-5.

6. Wallace BC, Small K, Brodley CE, et al. Deploying an interactive machine learning system in an evidence-based practice center: Abstrackr. Proceedings of the ACM International Health Informatics Symposium (IHI). 2012:819-24.

7. Evidence Partners. Meet your new assistant. Ottawa, Ontario: Systematic Review and Literature Review Software by Evidence Partners; 2012. https://www.evidencepartners.com/distiller-ai/. Accessed on June 13 2019.

8. EPPI-Centre Software. EPPI-reviewer 4.0. software for research synthesis. London: EPPI-Centre Software, Social Science Research Unit, Institute of Education; 2017. https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=2947. Accessed on June 13 2019.

9. Kontonatsios G, Brockmeier AJ, Przybyla P, et al. A semi-supervised approach using label propagation to support citation screening. J Biomed Inform. 2017 Aug;72:67-76. doi: 10.1016/j.jbi.2017.06.018. PMID: 28648605.

10. Ouzzani M, Hammady H, Fedorowicz Z, et al. Rayyan--a web and mobile app for systematic reviews. Qatar: Qatar Computing Research Institute; 2016. https://rayyan.qcri.org/welcome. Accessed on June 13 2019.

11. Howard BE, Phillips J, Miller K, et al. SWIFT-Review: a text-mining workbench for systematic review. Syst Rev. 2016 May 23;5:87. doi: 10.1186/s13643-016-0263-z. PMID: 27216467.

12. Ananiadou S, McNaught J. Text mining for biology and biomedicine. Boston/London: Artech House; 2006.

13. Hearst M. Untangling text data mining. Proceedings of the 37th annual meeting of the association for computational linguistics (ACL 1999). 1999:3-10.

14. Hempel S, Shetty KD, Shekelle PG, et al. Machine learning methods in systematic reviews: Identifying quality improvement intervention evaluations. Research White Paper (Prepared by the Southern California Evidence-based Practice Center under Contract No. 290-2007-10062-I). AHRQ Publication No. 12-EHC125-EF. Rockville, MD: Agency for Healthcare Research and Quality; September 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

15. 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 Jun 15;4:80. doi: 10.1186/s13643-015-0067-6. PMID: 26073974.

16. Przybyla P, Brockmeier AJ, Kontonatsios G, et al. Prioritising references for systematic reviews with RobotAnalyst: a user study. Res Synth Methods. 2018 Sep;9(3):470-88. doi: 10.1002/jrsm.1311. PMID: 29956486.

17. Shemilt I, Simon A, Hollands GJ, et al. Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. Res Synth Methods. 2014 Mar;5(1):31-49. doi: 10.1002/jrsm.1093. PMID: 26054024.

18. Bishop CM. Pattern recognition and machine learning. Switzerland: Springer; 2006.

19. Dobrokhotov PB, Goutte C, Veuthey AL, et al. Assisting medical annotation in Swiss-Prot using statistical classifiers. International Journal of Medical Informatics. 2005;74(2-4):317-24.

20. Affengruber L, Wagner G, Waffenschmidt S, et al. Combining abbreviated searches with single-reviewer screening– three case studies of rapid reviews. BMC Med Res Methodol. Submitted for publication.

21. Thomas J, Noel-Storr A, Marshall I, et al. Living systematic reviews: 2. Combining human and machine effort. J Clin Epidemiol. 2017 Nov;91:31-7. doi: 10.1016/j.jclinepi.2017.08.011. PMID: 28912003.

22. Gartlehner G, Gaynes B, Amick H, et al. Nonpharmacological versus pharmacological treatments for adult patients with major depressive disorder. Comparative Effectiveness Review No. 161. (Prepared by the RTI-UNC Evidence-based Practice Center under Contract No. 290-2012-00008I.) AHRQ Publication No. 15(16)-EHC031-EF. Rockville, MD: Agency for Healthcare Research and Quality; December 2015. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

23. Barber JP, Barrett MS, Gallop R, et al. Short-term dynamic psychotherapy versus pharmacotherapy for major depressive disorder: a randomized, placebo-controlled trial. J Clin Psychiatry. 2012 Jan;73(1):66-73. doi: 10.4088/JCP.11m06831. PMID: 22152401.

24. Bastos AG, Guimaraes LS, Trentini CM. Neurocognitive changes in depressed patients in psychodynamic psychotherapy, therapy with fluoxetine and combination therapy. J Affect Disord. 2013 Dec;151(3):1066-75. doi: 10.1016/j.jad.2013.08.036. PMID: 24103853.

25. Blom MB, Jonker K, Dusseldorp E, et al. Combination treatment for acute depression is superior only when psychotherapy is added to medication. Psychother Psychosom. 2007;76(5):289-97. doi: 10.1159/000104705. PMID: 17700049.

26. Blumenthal JA, Babyak MA, Doraiswamy PM, et al. Exercise and pharmacotherapy in the treatment of major depressive disorder. Psychosom Med. 2007 Sep-Oct;69(7):587-96. doi: 10.1097/PSY.0b013e318148c19a. PMID: 17846259.

27. Frank E, Cassano GB, Rucci P, et al. Predictors and moderators of time to remission of major depression with interpersonal psychotherapy and SSRI pharmacotherapy. Psychol Med. 2011 Jan;41(1):151-62. doi: 10.1017/s0033291710000553. PMID: 20380782.

28. Gastpar M, Singer A, Zeller K. Efficacy and tolerability of hypericum extract STW3 in long-term treatment with a once-daily dosage in comparison with sertraline. Pharmacopsychiatry. 2005 Mar;38(2):78-86. doi: 10.1055/s-2005-837807. PMID: 15744631.

29. Gertsik L, Poland RE, Bresee C, et al. Omega-3 fatty acid augmentation of citalopram treatment for patients with major depressive disorder. J Clin Psychopharmacol. 2012 Feb;32(1):61-4. doi: 10.1097/JCP.0b013e31823f3b5f. PMID: 22198441.

30. Hegerl U, Hautzinger M, Mergl R, et al. Effects of pharmacotherapy and psychotherapy in depressed primary-care patients: a randomized, controlled trial including a patients' choice arm. Int J Neuropsychopharmacol. 2010;13(1):31-44. doi: 10.1017/S1461145709000224. PMID: 19341510.

31. Huang Y, Htut W, Li D, et al. Studies on the clinical observation and cerebral glucose metabolism in depression treated by electro-scalp acupuncture compared to fluoxetine. Int J Clin Acupunct. 2005;14(1):7-26. PMID: 0077160.

32. Jazayeri S, Tehrani-Doost M, Keshavarz SA, et al. Comparison of therapeutic effects of omega-3 fatty acid eicosapentaenoic acid and fluoxetine, separately and in combination, in major depressive disorder. Aust N Z J Psychiatry. 2008 Mar;42(3):192-8. doi: 10.1080/00048670701827275. PMID: 18247193.

33. Kennedy SH, Konarski JZ, Segal ZV, et al. Differences in brain glucose metabolism between responders to CBT and venlafaxine in a 16-week randomized controlled trial. Am J Psychiatry. 2007 May;164(5):778-88. doi: 10.1176/appi.ajp.164.5.778. PMID: 17475737.

34. Lam RW, Parikh SV, Ramasubbu R, et al. Effects of combined pharmacotherapy and psychotherapy for improving work functioning in major depressive disorder. Br J Psychiatry. 2013 Nov;203(5):358-65. doi: 10.1192/bjp.bp.112.125237. PMID: 24029535.

35. McGrath CL, Kelley ME, Holtzheimer PE, et al. Toward a neuroimaging treatment selection biomarker for major depressive disorder. JAMA Psychiatry. 2013 Aug;70(8):821-9. doi: 10.1001/jamapsychiatry.2013.143. PMID: 23760393.

36. Menchetti M, Rucci P, Bortolotti B, et al. Moderators of remission with interpersonal counselling or drug treatment in primary care patients with depression: randomised controlled trial. Br J Psychiatry. 2014 Feb;204(2):144-50. doi: 10.1192/bjp.bp.112.122663. PMID: 24311553.

37. Mischoulon D, Price LH, Carpenter LL, et al. A double-blind, randomized, placebo-controlled clinical trial of S-adenosyl-L-methionine (SAMe) versus escitalopram in major depressive disorder. J Clin Psychiatry. 2014 Dec 24. . doi: 10.4088/JCP.13m08591. PMID: 24500245.

38. Mynors-Wallis LM, Gath DH, Day A, et al. Randomised controlled trial of problem solving treatment, antidepressant medication, and combined treatment for major depression in primary care. BMJ. 2000 Jan 1;320(7226):26-30. PMID: 10617523.

39. Raue PJ, Schulberg HC, Heo M, et al. Patients' depression treatment preferences and initiation, adherence, and outcome: a randomized primary care study. Psychiatr Serv. 2009 Mar;60(3):337-43. doi: 10.1176/appi.ps.60.3.337. PMID: 19252046.

40. Schrader E. Equivalence of St John's wort extract (Ze 117) and fluoxetine: a randomized, controlled study in mild-moderate depression. Int Clin Psychopharmacol. 2000 Mar;15(2):61-8. doi: 10.1097/00004850-200015020-00001. PMID: 10759336.

41. Segal ZV, Kennedy S, Gemar M, et al. Cognitive reactivity to sad mood provocation and the prediction of depressive relapse. Arch Gen Psychiatry. 2006 Jul;63(7):749-55. doi: 10.1001/archpsyc.63.7.749. PMID: 16818864.

42. Song Y, Zhou D, Fan J, et al. Effects of electroacupuncture and fluoxetine on the density of GTP-binding-proteins in platelet membrane in patients with major depressive disorder. J Affect Disord. 2007 Mar;98(3):253-7. doi: 10.1016/j.jad.2006.07.012. PMID: 16919758.

43. Zhang WJ, Yang XB, Zhong BL. Combination of acupuncture and fluoxetine for depression: a randomized, double-blind, sham-controlled trial. J Altern Complement Med. 2009 Aug;15(8):837-44. doi: 10.1089/acm.2008.0607. PMID: 19678773.

44. Wagner G, Nussbaumer-Streit B, Greimel J, et al. Trading certainty for speed - how much uncertainty are decisionmakers and guideline developers willing to accept when using rapid reviews: an international survey. BMC Med Res Methodol. 2017 Aug 14;17(1):121. doi: 10.1186/s12874-017-0406-5. PMID: 28806999.

45. O'Connor AM, Tsafnat G, Thomas J, et al. A question of trust: can we build an evidence base to gain trust in systematic review automation technologies? Syst Rev. 2019 Jun 18;8(1):143. doi: 10.1186/s13643-019-1062-0. PMID: 31215463.

46. Waffenschmidt S, Janzen T, Hausner E, et al. Simple search techniques in PubMed are potentially suitable for evaluating the completeness of systematic reviews. J Clin Epidemiol. 2013 Jun;66(6):660-5. doi: 10.1016/j.jclinepi.2012.11.011. PMID: 23419611.

Abbreviations and Acronyms

AHRQ	Agency for Healthcare Research and Quality
CI	confidence interval
EPC	Evidence-based Practice Center
EPPI	Evidence for Policy and Practice Information
FN	false negative
FP	false positive
RCT	randomized controlled trial
ROC	receiver operating characteristics
TN	true negative
TP	true positive