Methods Research Report

Performance and Usability of Machine Learning for Screening in Systematic Reviews: A Comparative Evaluation of Three Tools

AHRQ
Agency for Healthcare Research and Quality

Methods Research Report

Performance and Usability of Machine Learning for Screening in Systematic Reviews: A Comparative Evaluation of Three Tools

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. 290-2015-00001-I

Prepared by:
University of Alberta Evidence-based Practice Center
Edmonton, Alberta, Canada

Investigators:
Allison Gates, Ph.D.
Samantha Guitard, M.Sc.
Jennifer Pillay, M.Sc.
Sarah A. Elliott, Ph.D.
Michele P. Dyson, Ph.D.
Amanda S. Newton, Ph.D., R.N.
Lisa Hartling, Ph.D.

AHRQ Publication No. 19(20)-EHC027-EF
November 2019

Key Messages

Purpose of project

For title and abstract screening, we explored the reliability of three machine learning tools when used to automatically eliminate irrelevant records or complement the work of a single reviewer. We evaluated the usability of each tool.

Key messages

- The reliability of the tools to automatically eliminate irrelevant records was highly variable; a median (range) 70% (0-100%) of relevant records were missed compared to dual independent screening.
- Abstrackr and RobotAnalyst improved upon single reviewer screening by identifying studies that the single reviewer missed, but performance was not reliable. DistillerSR provided no advantage over single reviewer screening.
- The tools' usability relied on multiple properties: user friendliness; qualities of the user interface; features and functions; trustworthiness; ease and speed of obtaining the predictions; and practicality of the export files.
- Standards for conducting and reporting evaluations of machine learning tools for screening will facilitate their replication.

This report is based on research conducted by the University of Alberta Evidence-based Practice Center under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290-2015-00001-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 decision makers—patients and clinicians, health system leaders, and policymakers, among others—make well-informed decisions and thereby improve the quality of healthcare services. This report is not intended to be a substitute for the application of clinical judgment. Anyone who makes decisions concerning the provision of clinical care should consider this report in the same way as any medical reference and in conjunction with all other pertinent information, i.e., in the context of available resources and circumstances presented by individual patients.

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

Suggested citation: Gates A, Guitard S, Pillay J, Elliott SA, Dyson MP, Newton AS, Hartling L. Performance and Usability of Machine Learning for Screening in Systematic Reviews: A Comparative Evaluation of Three Tools. (Prepared by the University of Alberta Evidence-based Practice Center under Contract No. 290-2015-00001-I) AHRQ Publication No. 19(20)-EHC027-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/AHRQEPCMETHMACHINEPERFORMANCE

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

Stephanie Chang, M.D., M.P.H.
TOO and Director
Evidence-based Practice Center Program
Center for Evidence and Practice Improvement
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

Acknowledgments

We thank Dr. Michelle Gates for piloting the usability survey and screening exercise, and for providing revisions on the draft report. We thank Dr. Meghan Sebastianski for reviewing the qualitative analysis. We also thank our colleagues who took the time to test the machine learning tools and provide feedback.

Peer Reviewers

Prior to publication of the final report, EPCs sought input from independent Peer Reviewers without financial conflicts of interest. However, the 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.

James Thomas, Ph.D.
Associate Director, EPPI-Centre, University College London
London, United Kingdom

Steve McDonald, M.A., Grad Dip
Co-Director, Cochrane Australia
Melbourne, Australia

Performance and Usability of Machine Learning for Screening in Systematic Reviews: a Comparative Evaluation of Three Tools

Structured Abstract

Background. Machine learning tools can expedite systematic review (SR) completion by reducing manual screening workloads, yet their adoption has been slow. Evidence of their reliability and usability may improve their acceptance within the SR community. We explored the performance of three tools when used to: (a) eliminate irrelevant records (Automated Simulation) and (b) complement the work of a single reviewer (Semi-automated Simulation). We evaluated the usability of each tool.

Methods. We subjected three SRs to two retrospective screening simulations. In each tool (Abstrackr, DistillerSR, and RobotAnalyst), we screened a 200-record training set and downloaded the predicted relevance of the remaining records. We calculated the proportion missed and the workload and time savings compared to dual independent screening. To test usability, eight research staff undertook a screening exercise in each tool and completed a survey, including the System Usability Scale (SUS).

Results. Using Abstrackr, DistillerSR, and RobotAnalyst respectively, the median (range) proportion missed was 5 (0 to 28) percent, 97 (96 to 100) percent, and 70 (23 to 100) percent in the Automated Simulation and 1 (0 to 2) percent, 2 (0 to 7) percent, and 2 (0 to 4) percent in the Semi-automated Simulation. The median (range) workload savings was 90 (82 to 93) percent, 99 (98 to 99) percent, and 85 (85 to 88) percent for the Automated Simulation and 40 (32 to 43) percent, 49 (48 to 49 percent), and 35 (34 to 38 percent) for the Semi-automated Simulation. The median (range) time savings was 154 (91 to 183), 185 (95 to 201), and 157 (86 to 172) hours for the Automated Simulation and 61 (42 to 82), 92 (46 to 100), and 64 (37 to 71) hours for the Semi-automated Simulation. Abstrackr identified 33-90% of records erroneously excluded by a single reviewer, while RobotAnalyst performed less well and DistillerSR provided no relative advantage. Based on reported SUS scores, Abstrackr fell in the usable, DistillerSR the marginal, and RobotAnalyst the unacceptable usability range. Usability depended on six interdependent properties: user friendliness, qualities of the user interface, features and functions, trustworthiness, ease and speed of obtaining predictions, and practicality of the export file(s).

Conclusions. The workload and time savings afforded in the Automated Simulation came with increased risk of erroneously excluding relevant records. Supplementing a single reviewer's decisions with relevance predictions (Semi-automated Simulation) improved upon the proportion missed in some cases, but performance varied by tool and SR. Designing tools based on reviewers' self-identified preferences may improve their compatibility with present workflows.

Contents

Tables

Figures

Appendixes

Introduction

There is growing recognition that expedited systematic review (SR) processes need to be developed, tested, and implemented to assist reviewers in keeping pace with the rapid publication of primary studies.[1] To ensure that the conclusions of SRs are valid, reviewers use rigorous approaches to locate all relevant evidence related to their research question.[2] This typically entails exhaustive, highly sensitive search methods.[3] To select studies for inclusion, often two reviewers independently filter through a large quantity of records in two stages (first by title and abstract, then by full text) to identify the few (three percent on average)[1] that are relevant. Screening thus represents a time consuming step in the evidence synthesis process. Single-reviewer screening reduces the total workload; however, the risk of erroneously excluding relevant records and biasing the SR's findings also increases.[4] As title and abstract screening is a review step that may be particularly amenable to automation or semi-automation,[5-7] there is increasing interest in ways that review teams can leverage machine learning tools to expedite screening while maintaining SR validity.[8]

There is an abundant and active body of research investigating ways that machine learning tools might be used to reduce screening workloads,[9] much of which exists in the computer science literature. One way that machine learning tools expedite screening is by prioritizing relevant records; that is, by presenting them to reviewers in decreasing order of predicted relevance. This allows reviewers to identify relevant studies sooner[5] so that other members of the review team can move forward with subsequent SR steps (e.g., full text screening, data extraction, quality appraisal) earlier.[9] Moreover, many machine learning tools can predict the relevance of remaining records after the reviewers screen a “training set”. Despite the abundance of literature in the area, what remains unclear to systematic reviewers is how and when review teams can reliably leverage these relevance predictions to semi-automate screening. Moreover, with so many tools available (many freely), reviewers would benefit from understanding the similarities and differences in their reliability, usability, learnability, and associated costs.

A review of published studies on applications of machine learning for screening found that these tools could be used safely to prioritize relevant records and cautiously to replace the work of one of the human reviewers.[9] The evidence for using machine learning tools to automatically eliminate irrelevant records was less certain, and the approach is not currently recommended.[9] Despite their promise, the adoption of machine learning tools among systematic reviewers has been slow.[9-11] O'Connor et al. summarized the potential barriers to the adoption of machine learning tools among systematic reviewers. Fundamental concerns included distrust in machine learning approaches by review teams and end users; set-up challenges and incompatibility with current SR production processes; doubts as to whether machines can reliably perform SR tasks; and poor awareness of available tools.[12] In light of known barriers,[12-15] we designed an explorative study to investigate the relative advantages (workload and estimated time savings) and risks (erroneously excluding relevant studies) of different approaches to leveraging machine learning tools to automate or semi-automate title and abstract screening. We also aimed to explore their usability among experienced systematic reviewers. The findings of this study address two facilitators to the adoption of new technologies: being perceived as providing greater relative advantages, and compatibility with current SR workflows.[12]

Objectives

For a sample of three SRs, we aimed to retrospectively explore and compare how three machine learning tools would perform for title and abstract screening when used (a) in the context of single reviewer screening to eliminate irrelevant records, and (b) in the context of dual independent screening to complement the work of one of the human reviewers. We based performance on three metrics: the proportion of studies missed, the workload savings, and the estimated time savings compared to dual independent screening. We also aimed to compare user experiences across the three tools among experienced reviewers at our evidence synthesis center.

Methods

Conduct

We followed an a priori protocol, available upon request. Within the methods, we describe changes made to the protocol that occurred while undertaking the study.

Machine Learning Tools

Abstrackr (http://abstrackr.cebm.brown.edu), DistillerSR (the machine learning tool being DistillerAI) (http://www.evidencepartners.com), and RobotAnalyst (http://www.nactem.ac.uk/robotanalyst/) are online machine learning tools that aim to enhance the efficiency of SR production by semi-automating title and abstract screening. From a user's perspective, the three tools function similarly. After uploading all citations retrieved via the electronic searches to the user interface, titles and abstracts appear on-screen and the reviewers are prompted to label each as include, exclude, or unsure. The machine learning algorithms use the reviewers' relevance labels and other data (e.g., relevance terms tagged by the reviewers, text mining for MeSH terms and keywords) to predict records that may be safely excluded and those that require further screening.

Although many machine learning tools exist to expedite screening,[16] we chose Abstrackr, DistillerSR, and RobotAnalyst for the following reasons. First, the development of these tools has been well documented.[17-19] At least for Abstrackr and RobotAnalyst, research teams aside from the tools' developers have evaluated their real-world performance in SRs,[20-22] facilitating comparisons of our findings to previous research. We also chose the tools for practical reasons. All three allow the user to download the relevance predictions after screening a small training set, a function that is not available in all tools. Moreover, both Abstrackr and RobotAnalyst are free to use. Although DistillerSR is a pay-for-use software, our center maintains a user account, so it was logical to include it in this study.

We also selected tools that offered a heterogeneous array of features. Key differences between the tools played an important role in the user experience testing by exposing participants to diverse user interfaces. It also provided them the opportunity to identify features that enhanced or detracted from the user experience across tools.

Performance Testing

Sample of Systematic Reviews

We selected a convenience sample of three SRs completed or underway at our center: the Alberta Research Centre for Health Evidence (ARCHE) and University of Alberta Evidence-based Practice Center (UAEPC), University of Alberta, Edmonton, Alberta, Canada. All three of the SRs investigated healthcare interventions, as follows: first- and second-generation antipsychotics for children and young adults;[23] treatments for bronchiolitis in infants in acute care (International prospective register of systematic reviews (PROSPERO) number: CRD42016048625); and screening for impaired visual acuity and vision-related functional limitations in older adults.[24] For brevity, we hereafter refer to these SRs as Antipsychotics, Bronchiolitis, and Visual Acuity, respectively. Table 1 shows the PICOS (population, intervention, comparator, outcome, and study design) criteria for each SR.

Table 1. Population, intervention, comparator, outcome, and study design (PICOS) criteria for the systematic reviews

Criteria	Antipsychotics	Bronchiolitis	Visual Acuity
Population	Children and young adults aged ≤24 years experiencing a psychiatric disorder or behavioral issues outside the context of a disorder	Infants and young children aged <24 months experiencing their first episode of wheeze, or diagnosed with bronchiolitis or RSV	Community-dwelling adults aged ≥65 years with unrecognized impaired visual acuity or vision-related functional limitations
Intervention	Any Food and Drug Administration-approved first- or second-generation antipsychotic	Any bronchodilator, any corticosteroid, hypertonic saline, oxygen therapy, antibiotics, heliox	Vision screening tests (alone or within multicomponent screening/assessment) performed by primary healthcare professionals
Comparators	Placebo, no treatment, any other antipsychotic, the same antipsychotic in a different dose	Placebo, usual care, no treatment, normal saline, or another intervention of interest	No screening, delayed screening, attention control, screening involving all components of intervention except vision component, usual care
Outcomes	Intermediate and effectiveness outcomes, adverse effects and major adverse effects, adverse effects limiting treatment, specific adverse events, persistence and reversibility of adverse effects	Outpatient admissions, inpatient length of stay, change in clinical score, oxygen saturation, respiratory rate, heart rate, pulmonary function, adverse events, escalation of care, length of illness, duration of oxygen therapy	Benefits (e.g., mortality, adverse consequences of poor vision), harms (e.g., serious adverse events), implementation factors (e.g., uptake of referrals)
Study designs	RCTs and nRCTs, controlled cohort studies, controlled before-after studies	RCTs	RCTs, controlled experimental and observational studies

nRCT = non-randomized controlled trial; RCT = randomized controlled trial; RSV = respiratory syncytial virus

Screening Procedure

For each SR, we uploaded all records retrieved by the searches to each tool via RIS (Research Information Systems) files downloaded from EndNote (v. X9, Clarivate Analytics, Philadelphia, Pennsylvania). We set up the SRs in each tool for single-reviewer screening and with the records presented in random order. Although we had originally intended to use the "most likely to be relevant" prioritization, we were not successful in applying this setting in all tools (i.e., due to server errors or glitches in two of the tools, Abstrackr and RobotAnalyst).

When using machine learning tools for screening, inaccurate labels in the training set (i.e., as applied by the human reviewer(s)) will result in inaccurate predictions. Thus, for a training set of 200 records we retrospectively replicated the senior reviewer's (i.e., the reviewer with the most content expertise or SR experience) screening decisions in each tool. We decided on a 200-record training set because in a previous evaluation,[21] we found that this number was sufficient to bring about predictions. Moreover, the developers of DistillerAI recommend a minimum training set size of 40 excluded and 10 included records, and a maximum size of 300 records (after which learning diminishes).[25] Because the records appeared in random order, the training set differed across the tools for each review. Although this could affect the predictions, in a previous evaluation we found little difference in Abstrackr’s predictions over three independent trials.[18]

At our center, it is standard practice that any record marked as “include” (i.e., relevant) or “unsure” by either reviewer during title and abstract screening is eligible for scrutiny by full text. In other words, the “include” and “unsure” decisions are equivalent. For this reason, our screening files typically include one of two screening decisions for each record: include/unsure or exclude. Because we were unable to retrospectively ascertain whether the decision for individual records was “include” or “unsure”, we entered all “include/unsure” decisions as “relevant” in each tool.

After screening the training set for each SR in each tool, we downloaded the relevance predictions for the remaining records. In DistillerSR and RobotAnalyst the predictions were available immediately. In Abstrackr, they were typically available the following day. In instances where the predictions did not become available within 48 hours, we continued to screen in batches of 100 records until they did. The format of the downloaded predictions varied by tool. Abstrackr produced “hard screening predictions” (true, i.e. include or false, i.e. exclude) and relevance probabilities for each remaining record. We used the hard screening predictions rather than applying custom thresholds based on the probabilities. Both DistillerSR and RobotAnalyst provided binary predictions (include or exclude) for all remaining records. Although customization was possible in DistillerSR, we used the “simple review” function to automatically classify the remaining records.

Retrospective Simulations

Based on existing reviews,[5, 9 ,14] we postulated that the predictions downloaded from the machine learning tools could be leveraged in two ways: (a) to automatically exclude irrelevant records, or (b) to complement the work of one of the human reviewers. We thus devised two hypothetical approaches and ran retrospective simulations to test our hypothesis. In the first approach (Automated Simulation, the automatic exclusion of records), after screening a training set of 200 records, the senior reviewer would download the predictions and exclude all records predicted to be irrelevant. To reduce the full-text screening workload, the reviewer would continue to screen the records predicted to be relevant. Of these, the records that the reviewer agreed were relevant would move forward to full text screening. In the second approach (Semi-automated Simulation, complementing the work of one human reviewer), we aimed to determine whether the predictions could be leveraged to improve upon the work of the a single reviewer (as naturally, a single reviewer can be expected to erroneously exclude relevant records).[4] In this simulation, the senior reviewer would follow the same approach as in the Automated Simulation, and the second reviewer would screen all of the records as per usual. Any record marked as relevant by the second reviewer or the senior reviewer/tool’s predictions would move forward to full text screening.

To test the performance of each approach, we created a workbook in Excel (v. 2016, Microsoft Corporation, Redmond, Washington) for each SR. The workbooks included a row for each record retrieved via the searches and a column for each of: the record identification number, the title and abstract screening decisions for the senior and second reviewers, the full text consensus decisions (i.e., the records included in the final reports), and the relevance predictions from each tool. We then determined the title and abstract consensus decisions that would have resulted from each simulation. As per standard practice at our center, we considered any record marked as “include” by either of the reviewers to be relevant for scrutiny by full text.

User Experience Testing

We approached eleven research staff at our center via e-mail to participate in the user experience testing. These staff members were experienced in producing SRs (e.g., research assistants, project coordinators, research associates), but had no or very little experience with machine learning tools for screening. From the time of the first e-mail contact, we allowed invited participants one month to undertake the study, which entailed completing a screening exercise in each tool and a user experience survey. We sent two reminder e-mails prompting potential participants to partake in the study. Participation was voluntary and completion of the survey implied consent. We received ethical approval to complete the user experience testing from the University of Alberta Research Ethics Board on 24 January 2019 (Pro00087862).

We designed a screening exercise that aligned with typical screening practices at our center (Appendix A). The aim of the exercise was to guide participants through the steps involved in setting up a SR, uploading a set of records, screening a training set, and downloading the predictions in each tool. We provided minimal guidance but instructed participants to use the "Help" function in each tool if needed. We also encouraged participants to browse the available functions and to keep track of features that they liked or disliked.

We selected a SR currently underway at our center for the screening exercise (digital technology distractions for pain in children, PROSPERO CRD42017077622). We selected this SR because the eligibility criteria were relatively straightforward. We wanted participants to focus on their experience in each tool and did not want complex screening criteria to be a distraction. To reduce the risk of response bias, we used the random numbers generator in Excel to randomize the order in which each participant tested the three tools.

The survey (Appendix B), hosted in REDCap (Research Electronic Data Capture),[26] asked participants to complete the System Usability Scale (SUS)[27] for each tool. The SUS is a 10-item questionnaire that assesses subjective usability using a Likert-like scale.[27] The survey also asked participants to: elaborate on their experiences with each tool, via free-text responses; rank the three tools in order of preference for screening; and describe the features that supported or detracted from the usability of the tools.

Before beginning the user experience testing, the screening exercise and survey were pilot tested by two researchers at our center. We made minor changes to both the screening exercise (reduced the suggested number of citations to screen in each tool to minimize participant burden) and survey (edited for typos) following the pilot; however, because the changes were minimal, we retained the data from the two researchers who completed the pilot testing for analysis, with permission.

Analysis

Performance

We exported the simulation data in Excel to SPSS Statistics (v. 25, IBM Corporation, Armonk, New York) for analysis. We used data from 2 x 2 cross-tabulations to calculate standard[9] performance metrics for each simulation, as follows:

- **Proportion of records missed (i.e., error)**: of the studies included in the final report, the proportion that would have been excluded during title and abstract screening.

We made informal comparisons of the proportion missed for each simulation and tool to single reviewer screening to estimate the acceptability of its performance (i.e., whether it would improve upon single reviewer screening).

- **Workload savings (i.e., absolute screening reduction)**: of the records that need to be screened at the title and abstract stage, the proportion that would not need to be screened manually.
- **Estimated time savings**: the time saved by not screening records manually. We assumed a screening rate of 0.5 minutes per record[28] and an 8-hour work day.

Appendix C shows sample calculations for the Antipsychotics SR using Abstrackr's predictions.

User Experiences

We exported the quantitative survey data from REDCap to an Excel workbook for analysis, and the qualitative survey data to Word (v. 2016, Microsoft Corporation, Redmond, Washington). For each participant, we calculated the overall value of usability for each tool following the recommendations of Brooke (1996):[27] the sum of the score contributions from each item, where items 1, 3, 5, 7, and 9 contribute the scale position minus 1, and items 2, 4, 6, 8, and 10 contribute 5 minus the scale position; the sum is multiplied by 2.5 to obtain an overall value out of 100. We calculated the median and interquartile range (IQR) of scores for each tool, and categorized their usability following the recommendations of Bangor et al. (2008):[29] not acceptable (0 to 50), marginal (50 to 70), and acceptable (70 to 100). For the ranking of tools by preference, we calculated counts and percentages.

We analyzed the qualitative data following standard, systematic approaches to thematic analysis.[30] Because the tools that we trialed are constantly evolving, their functions and user interfaces are likely to change with time. Thus, we synthesized the qualitative data for all of the tools collectively to elucidate which qualities make a tool more or less appealing and usable to systematic reviewers. By synthesizing the comments collectively, we aimed to identify some of these qualities without focusing on the strengths and weaknesses of individual tools (which are subject to change). One researcher (AG) initially read the text and applied one or more codes to each line. Next, the researcher identified the most significant and frequent codes, combined similar codes, and renamed the categories of codes. The researcher developed memos for each theme, using examples from participants' experiences with each tool to illustrate more and less desirable features. To reduce the risk of interpretive bias, a second researcher external to the study team (and who did not partake in the user experience testing) reviewed the coding and themes for differences in interpretation. All disagreements were resolved via discussion.

Results

Performance

Table 2 shows the screening characteristics for each SR. The screening workload was relatively large for all SRs, ranging from 5,861 to 12,156 records following the removal of duplicates. Across SRs, two to 10 percent of records were retained for scrutiny by full text after the initial dual independent screening. Two percent or less of all records retrieved were included in the final SRs. The Visual Acuity review was unique in that only one record from the 11,229 screened was included in the final report. By contrast, the final reports for the Antipsychotics and Bronchiolitis reviews included 127 of 12,156 and 137 of 5,861 records, respectively.

Predictions were available after screening 200 records for all SRs in all tools with the exception of Visual Acuity in Abstrackr. As planned, we screened an additional 100 records, and the predictions became available. For two of the SRs RobotAnalyst did not upload the full list of records from the RIS file. Because all of our troubleshooting attempts (at least six attempts and contact with the tool's developers) failed we assumed that the additional 170 records for Bronchiolitis and 183 records for Visual Acuity would need to be screened manually. Because we could not obtain predictions for these records, we used the human reviewers' original decisions (include or exclude) when applying the Simulations.

In Abstrackr, DistillerSR, and RobotAnalyst, the training sets included a median (range) of 12 (4, 15), 14 (2, 14), and 15 (3, 20) includes respectively, with the balance being excludes. After screening the training sets, Abstrackr, DistillerSR, and RobotAnalyst predicted that a median (range) 18 (12, 33) percent, 0.1 (0, 1) percent, and 29 (20, 29) percent of the remaining records were relevant, respectively. Cross-tabulations showing records included in the final report relative to those deemed relevant via each Simulation are in Appendix D.

Table 2. Characteristics of the reviews and screening predictions for each tool

Characteristic	Antipsychotics, N records (%)	Bronchiolitis, N records (%)	Visual Acuity, N records (%)
Screening workload[a]	12156	5861	11229
Included by title/abstract[b]	1178 (10)	518 (9)	224 (2)
Included in the review[b]	127 (1)	137 (2)	1 (<1)
Includes/excludes in training set	Abstrackr: 15/185 DistillerSR: 14/186 RobotAnalyst: 20/180	Abstrackr: 12/188 DistillerSR: 14/186 RobotAnalyst: 15/185	Abstrackr:[c] 4/296 DistillerSR: 2/198 RobotAnalyst: 3/197
Screened by tool[d]	11956 (98)	5661 (97)	11029 (98)
Predicted relevant by Abstrackr	2117 (18)	656 (12)	3639 (33)
Predicted relevant by DistillerSR	7 (<1)	83 (1)	0 (0)
Predicted relevant by RobotAnalyst	3488 (29)	1082 (19)	3221 (29)

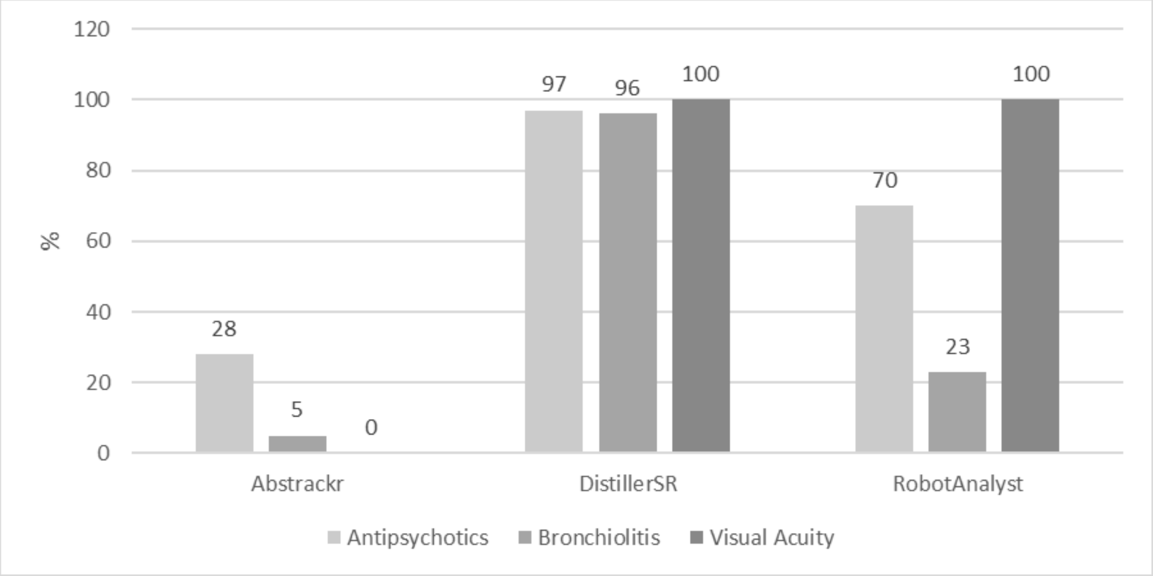
[a]Total number of records retrieved via the electronic searches. Each record was screened by two reviewers.
[b]Included following the initial screening by two independent reviewers (retrospective).
[c]All training sets were 200 records, with the exception of the Visual Acuity review which required a 300-record training set in Abstrackr before predictions were produced.
[d]After a 200-record training set.

Automated Simulation

Proportion Missed

Records "missed" are those that would not have moved forward to full text screening, but were included in the final reports. The median (range) proportion missed was 5 (0, 28) percent, 97 (96, 100) percent, and 70 (23, 100) percent using Abstrackr, DistillerSR, and RobotAnalyst, respectively (Figure 1).

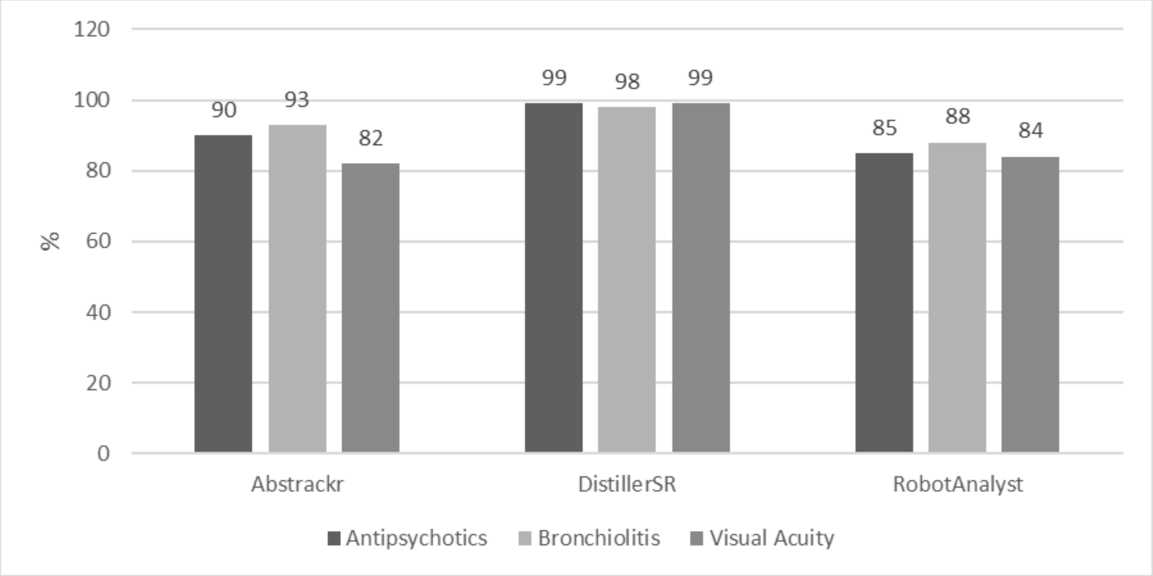
Figure 1. Proportion missed (percent) by tool and systematic review, Automated Simulation.



Workload Savings

The median (range) workload savings was 90 (82, 93) percent, 99 (98, 99) percent, 85 (84, 88) percent for Abstrackr, DistillerSR, and RobotAnalyst, respectively (Figure 2).

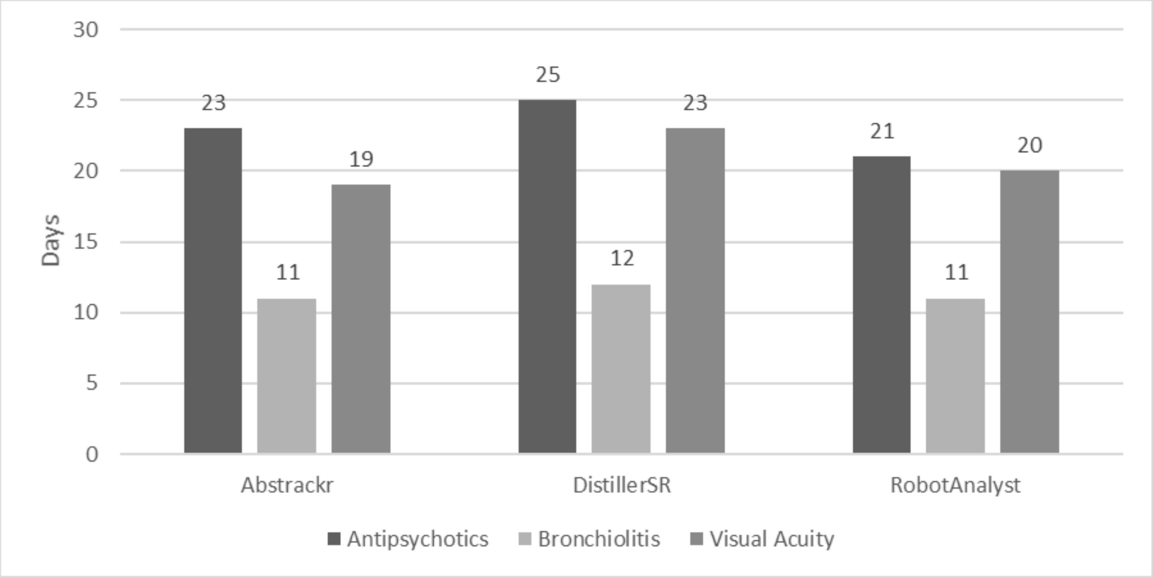
Figure 2. Workload savings (percent) by tool and systematic review, Automated Simulation.



Estimated Time Savings

The median (range) time savings was 154 (91, 183), 185 (95, 201), and 157 (86, 172) hours for Abstrackr, DistillerSR, and RobotAnalyst, respectively (i.e., a respective 19 (11, 23), 23 (12, 25), and 20 (11, 21) days) (Figure 3).

Figure 3. Estimated time savings (days) by tool and systematic review, Automated Simulation.

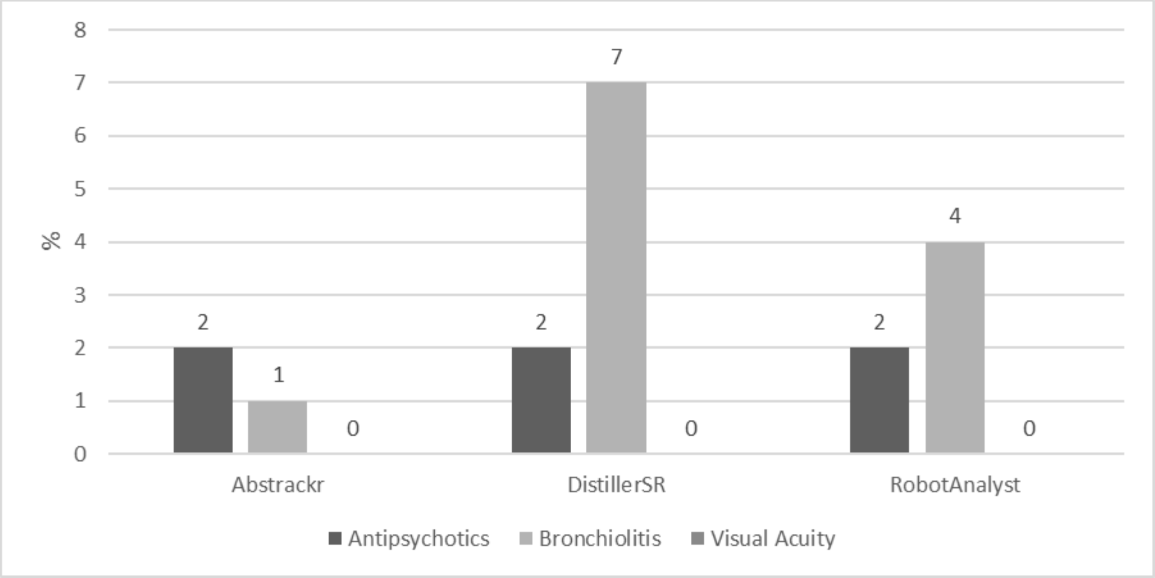


Semi-automated Simulation

Proportion Missed

The median (range) proportion missed was 1 (0, 2) percent, 2 (0, 7) percent, and 2 (0, 4) percent, respectively (Figure 4). Important to the performance of the semi-automated simulation is the contribution of each tool's predictions to the overall screening accuracy. Had the second reviewer independently screened the records for Antipsychotics, Bronchiolitis, and Visual Acuity independently, a respective 3 (2%), 10 (7%), and 0 records would have been missed compared to dual independent screening. Abstrackr correctly predicted the relevance of 1 (33%) and 9 (90%) records erroneously excluded by the second reviewer in the Antipsychotics and Bronchiolitis reviews, respectively. DistillerSR did not correctly predict the relevance of any of the records erroneously excluded by the second reviewer in either review, thus providing no advantage over single reviewer screening. RobotAnalyst correctly predicted the relevance of 4 (40%) records erroneously excluded by the second reviewer in Bronchiolitis, but none of those erroneously excluded in Antipsychotics.

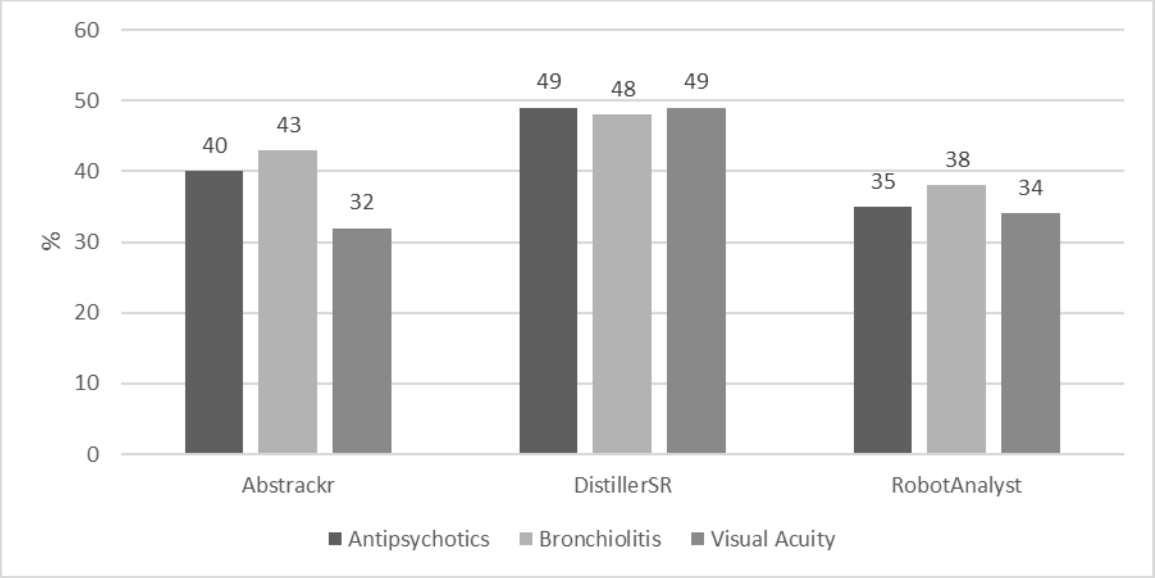
Figure 4. Proportion missed (percent) by tool and systematic review, Semi-automated Simulation.



Workload Savings

The median (range) workload savings was 40 (32, 43) percent, 49 (48, 49) percent, and 35 (34, 38) percent for Abstrackr, DistillerSR, and RobotAnalyst, respectively (Figure 5).

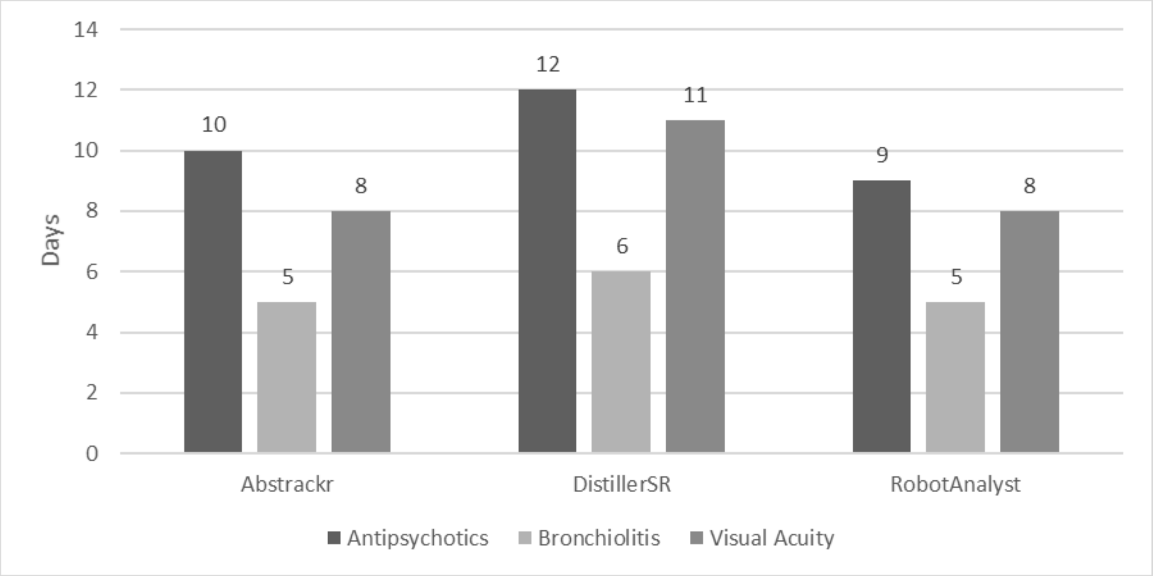
Figure 5. Workload savings (percent) by tool and systematic review, Semi-automated Simulation.



Estimated Time Savings

The median (range) time savings was 61 (42, 82), 92 (46, 100), and 64 (37, 71) hours for Abstrackr, DistillerSR, and RobotAnalyst, respectively (i.e., a respective 8 (5, 10), 11 (6, 12), and 8 (5, 9) days) (Figure 6).

Figure 6. Estimated time savings (days) by tool and systematic review, Semi-automated Simulation.



Post-Hoc Analysis

On the recommendation of another Evidence-based Practice Center following our initial testing we repeated the same procedures for a 500-record training set. We undertook the simulations for the larger training set only in Abstrackr, accounting for time and resource limitations. For the Automated Simulation, the proportion missed increased to 41 percent for Antipsychotics (compared with 28 percent) and 9 percent for Bronchiolitis (compared with 5 percent). There was no change in the proportion missed for Visual Acuity. The workload savings increased marginally for each SR: from 90 percent to 95 percent for Antipsychotics, 93 percent to 94 percent for Bronchiolitis, and 82 percent to 83 percent for Visual Acuity. Similarly, there were marginal increases in the estimated time savings: from 183 to 193 hours for Antipsychotics, 91 to 92 hours for Bronchiolitis, and 154 to 156 hours for Visual Acuity.

For the Semi-automated Simulation, one additional record was missed for Antipsychotics; however, the proportion missed did not change. There was no change in the proportion missed for the other two SRs. The workload savings increased marginally: from 40 percent to 45 percent for Antipsychotics, 43 percent to 44 percent for Bronchiolitis, and 32 to 33 percent for Visual Acuity. Similarly, the estimated time savings increased marginally: from 82 to 92 hours for Antipsychotics, 42 to 43 hours for Bronchiolitis, and 61 to 62 hours for Visual Acuity.

User Experiences

Eight research staff participated in the user experience testing (73 percent response rate). The median (interquartile range) overall system usability score was 79 (23), 64 (31), and 31 (8) for Abstrackr, DistillerSR, and RobotAnalyst, respectively. Based on these scores, Abstrackr fell in the usable, DistillerSR the marginal, and RobotAnalyst the unacceptable usability range.[29] Table 3 includes details of the scores for each item of the SUS. In terms of preference, 62 percent of participants chose Abstrackr as their first choice and 38 percent as their second choice. Thirty-eight percent of participants chose DistillerSR as their first choice, 50 percent as their second choice, and 13 percent as their last choice. Thirteen percent of participants chose RobotAnalyst as their second choice and the remaining 88 percent as their last choice.

Table 3. System Usability Scale responses for each item, per tool

Item	Abstrackr	DistillerSR	RobotAnalyst
I think that I would like to use the tool frequently	3.5 (1)	4 (0.5)	1 (1)
I found the tool to be unnecessarily complex	2 (1)	3.5 (1.25)	3 (0.5)
I thought the tool was easy to use	4 (1.25)	2.5 (2)	2 (1.5)
I think that I would need the support of a technical person to be able to use the tool	1 (1)	2.5 (1.25)	4 (1.25)
I found the various function in the tool were well integrated	4 (1.25)	3.5 (2.25)	3 (1.25)
I thought there was too much inconsistency in the tool	2 (0.25)	1 (1.25)	4 (1.25)
I would imagine that most people would learn to use the tool very quickly	4.5 (1)	3 (1.25)	3 (0.25)
I found the tool very cumbersome to use	2 (0.5)	3 (1.25)	5 (0)
I felt very confident using the tool	4 (1)	3.5 (1.25)	2 (2.25)
I needed to learn a lot of things before I could get going with the tool	2 (0.25)	3 (0.5)	2.5 (1)
Overall score (/100)	79 (23)	64 (31)	31 (8)

[a]Likert-like scale: 1 = strongly disagree, 3 = neutral, and 5 = strongly agree. Values represent the median (interquartile range) of responses.

The qualitative analysis of participants' comments revealed that usability was contingent on six interdependent properties: user friendliness, qualities of the user interface, features and functions, trustworthiness, ease and speed of obtaining the predictions, and practicality of the export files. Appendix E includes focused codes and participant quotes for each property. In the following paragraphs, we describe each briefly.

User Friendliness

Some participants found Abstrackr to be easy to use, had little trouble finding and making use of the available functions, and described the screening as enjoyable. Others found that working through the program was not very intuitive. Particular areas of difficulty included figuring out how to upload records, change the review settings, export the records, and return to the main page. Participants also had discordant views of the user friendliness of DistillerSR. Although some participants believed it was user friendly and easy to navigate, there was overwhelming feedback that it was unnecessarily complex and required more skill to work through. For example, two participants reported needing to watch a tutorial to set up the screening process, while another noted that training might be required to use the program more efficiently. Comments regarding the user friendliness of RobotAnalyst were mostly negative. Although one participant found it intuitive and easy to navigate, the majority described a frustrating experience. Major issues included a slow server speed, constant pop-ups and error messages, loss of data (i.e., previous screening decisions disappearing/being deleted), and a generally cumbersome screening process.

Qualities of the User Interface

For the most part, participants seemed to like Abstrackr's user interface. Although some described it as rudimentary, many also praised the look and layout for its simplicity and lack of distractions. Participants also seemed to like DistillerSR's user interface, describing it as "clean", "bright", "clear", and "consistent". One participant specifically noted the use of white space around the abstract as a positive feature. Others, however, found there to be too much "going on" on the screen, and reported feeling overwhelmed and having difficulty narrowing in on the functions that they needed. There were few comments regarding RobotAnalysts's user interface, but most of these were positive. Some participants found it "pretty" and liked the colors, buttons, and layout. The only negative comment was from one participant who found the user interface to be "a bit busy".

Features and Functions

Participants liked that they could add notes, tag records, and easily change their screening decisions in Abstrackr. They also liked having the ability to modify review settings, for example the order of the records and the number of reviewers. Other participants described a lack of clarity about how some of the functions worked (particularly the most likely to be relevant prioritization and the tagging). One participant commented that the user guide was not very helpful. There was a strong sense that most participants felt DistillerSR had too many features, making it feel sophisticated but overly complex. Other specific issues included difficulty fixing screening mistakes, that two clicks were required to make a decision, and that there was redundancy in the screening form. Nevertheless, more than one participant commented that the tutorial videos were very helpful and the help function useful. One participant noted that the ability to drag and drop the upload file was a great feature, while another liked the ability to track the number of records screened and number remaining. Most of the comments regarding RobotAnalyst's features were negative, with the exception of one participant noting that the on-screen availability of the predictions was "nice." Participants took particular issue with the fact that they had to open each record to read the abstract, that the system did not automatically advance to the next record, and that there were frequent and unnecessary pop-up messages.

Trustworthiness

Most participants thought Abstrackr was trustworthy and that the program functioned smoothly and quickly. One participant, however, noted that the program was "unpredictable", could sometimes be slow or crash (especially when uploading records), and the error messages were difficult to understand. All comments regarding the trustworthiness of DistillerSR were positive. Participants reported that the program seemed professional (more so than Abstrackr), that the server was fast and responsive, and that the program appeared reliable. Comments regarding the trustworthiness of RobotAnalyst were mostly negative. Multiple participants called the program "glitchy", so much so that two participants were unable to complete the screening task. Other major complaints included a slow server speed, the spontaneous loss of screening data, and "a constant stream" of error messages. Multiple participants called the program unreliable or untrustworthy and reported that they would not use it for screening.

Ease and Speed of Obtaining Predictions

Most comments related to obtaining the predictions in Abstrackr pertained to the delay (multiple hours) in being able to access them. In both DistillerSR and RobotAnalyst the predictions become available and can be applied in real time (usually after far fewer than 200 records are screened). Because Abstrackr's predictions only appear once the server updates (seemingly once daily), reviewers must screen a training set and then wait until the following day to see if the predictions are ready to download. This was described as an "annoyance"; however, two participants also noted that this was not a serious issue given the potential for time savings. Participants seemed to appreciate that predictions were available in real time in DistillerSR. While some noted that running the artificial intelligence function was easy, others could not find it or figure out how to use it. Similarly, some noted that it would be helpful to have a tutorial on how to use the artificial intelligence function and what inclusion threshold to apply. Participants also appreciated that predictions were available in real time in RobotAnalyst, but noted that applying them to the review was slow.

Practicality of the Export Files

Participants expressed opposing views related to the practicality of Abstrackr's export files. While some participants believed that the export files were usable, practical, and would easily be imported into other programs (original export is an Excel file), others found them to be impractical and not user friendly. Particular issues included missing information (e.g., year of publication, reference identification number from EndNote) and ambiguous labels (i.e., 1 and -1 instead of include and exclude). Some noted that the files would need a lot of editing before they would be useable to reviewers, which could be tedious and prone to error. Thoughts about DistillerSR's export files were similar. Some reviewers appreciated the variety of output formats available, the quick download speed, and the overall simplicity, practicality, and user friendliness of the files. Others found that the process to download the files was unnecessarily complicated. As DistillerSR offers multiple output formats, some participants noted that on their first export attempt important information was missing from the files. One participant was not able to figure out how to download the predictions. A positive comment about RobotAnalyst's export files was that they were not difficult to download. Otherwise, participants found the export to be impractical, calling it "poorly organized" and "difficult to decipher."

Discussion, Limitations, and Conclusion

Before deciding whether to leverage machine learning tools' predictions in real-world SRs, review teams will need to balance the benefits (i.e., workload and estimated time savings) and risks (i.e., potential to miss relevant records) of their use. In this explorative study, adding Abstrackr's predictions to a single reviewer's decisions (Semi-automated Simulation) reduced the proportion of records missed compared with screening by the single reviewer alone, but performance varied by SR. Balanced with the potential for time savings, this approach could provide an acceptable alternative to dual independent screening, at least in some SRs. Conversely, RobotAnalyst performed less well than Abstrackr and DistillerSR provided no advantage over single-reviewer screening. We can only speculate that differences between tools may be a function of the relevance thresholds applied (we used the standard settings), or differences in the machine learning algorithms. Replication on heterogeneous samples of reviews will help inform when replacing one reviewer with the predictions of a machine may be worth the associated risk. Although the workload and time savings were superior when the tools were used to exclude irrelevant records (Automated Simulation), far more studies were erroneously excluded in most cases. Given the magnitude of risk involved, it is unlikely than any review team would adopt this approach.

An important consideration in this study is the size and nature of the training sets. Empirical data from the computer science literature show that learning increases quickly at the beginning of active learning (i.e., first few hundred records) and more slowly thereafter.[31] Thus, to obtain reliable predictions large training sets can be required, i.e., 60 percent of all records[17] or 2500 to 3000 citations.[32] It is unsurprising, then, that as a means to eliminate irrelevant records, the 200-record training produced unreliable predictions. Unfortunately, larger training sets may be impractical in real-world applications of the tools. The 200-record training set was sufficient, in many cases, when paired with a single reviewer to capture 95 percent or more of the relevant studies; however, this was not always reflective of an improvement over single reviewer screening. At present, the ideal training set size is unknown and likely review-specific.[8] For instance, in the present evaluation Abstrackr's predictions were most reliable for Bronchiolitis, which compared to Antipsychotics had fewer research questions and included only randomized controlled trials. We can speculate that machine learning performs better for simpler reviews, or reviews that include only randomized trials; however, our small sample precludes definitive conclusions. Studies that test the real-world advantages and risks of different training set sizes on SRs with varying characteristics (e.g., proportion of included studies, included study designs, number of review questions, complexity of the interventions) would help inform how the tools might be optimally applied.

As reported by O'Connor et al., even if machine learning-supported approaches to screening were ready to deploy, many review teams would be hesitant to adopt them pending widespread acceptance by credible methods groups (e.g., Cochrane, the Campbell Collaboration), peer reviewers, grant panels, and journal editors.[12] Moving toward this ideal, there is a need for a standardized approach to evaluating the performance and usability of the tools, and reporting on these evaluations.[10, 12, 33, 34] Consistently conducted and reported evaluations will facilitate their replication across tools and SRs,[33, 34] which will be imperative to the development of evidence-based guidance for their practical application in real-world screening tasks.[12] Continued efforts to develop and validate ways that machine learning tools' predictions may be leveraged by review teams to optimize time savings while maintaining acceptable reliability will require testing and replication on large, diverse samples of reviews. The development of a set of core outcome metrics, based on a consensus process including end users (systematic reviewers) and tool developers, may help improve upon the value of these studies to both groups. For example, whether missed studies would impact a review's conclusions is important to systematic reviewers but less frequently a consideration for tool developers. Designing tools that allow reviewers to customize the level of risk (i.e., by setting their own cut-points for inclusion based on probability estimates, as are reported by Abstrackr alongside the hard screening predictions) may also contribute to garnering trust.

Another important contributor to the adoption of machine learning tools for screening will be their usability and fit with standard SR workflows.[12] The usability of the three tools varied considerably and relied upon multiple properties: being user friendly (i.e., easy to navigate, intuitive); having a simple, uncluttered user interface; having sufficient, easy-to-use features to facilitate screening and project management (but not superfluous ones); being trustworthy (i.e., functioning smoothly, not losing data) and glitch-free; developing predictions in real time; and the availability of practical and useful records of the screening and predictions. Many of participants' comments could be generalized to the use of the tools for screening, regardless of whether the machine learning functions were employed. Despite being the only pay-for-use software among the three, DistillerSR was often not the favorite among reviewers, mainly because the multiple available features were overwhelming; however, advantages compared with the free tools included a more professional look and feel, greater trustworthiness, and the availability of user support.

To our knowledge, few studies have evaluated the usability of machine learning tools for screening in SRs.[13, 14] In addition to studies evaluating the performance of available tools, those appraising usability are needed to develop a better understanding of their compatibility with SR workflows.[11] Standard methods for evaluating usability (e.g., the SUS) will facilitate comparisons between tools (both within and across individual studies). Atlena et al. used the SUS to evaluate the usability of common software packages to support SRs, including Rayyan, EPPI-Reviewer, and Abstrackr, all of which have machine learning capabilities. The authors found no significant differences in usability scores between the tools and all scored near or within the acceptable usability range (they did not investigate DistillerSR or RobotAnalyst).[13] By contrast, in our study only Abstrackr fell within that range. Usability evaluations of a broader range of tools and among various review groups will contribute to the identification of the most promising tools and inform continued improvement based on the self-identified needs of reviewers.

Strengths and Limitations

Our study is one of few to compare the performance and user experiences across multiple machine learning tools for screening in SRs. Further, our study responds to a recent call from the International Collaboration for Automation of Systematic Reviews to trial and validate available tools[10] and addresses reported barriers to the adoption of machine learning tools for systematic reviewers.[12] The findings should be interpreted in light of a number of limitations. First, we used a random selection of records because we had difficulty successfully deploying the relevance prioritization within some of the tools, recognizing that a randomly selected training set would be less efficient and thus the predictions less accurate. Nevertheless, the random ordering reduced the risk of basing the predictions on biased training sets.

We used the standard settings in each tool to obtain predictions (i.e., the "hard screening predictions" in Abstrackr and the "simple review" function in DistillerSR). In the absence of empirical guidance for customizing the tools' settings (e.g., setting review-specific inclusion and exclusion thresholds), using the standard functions likely best approximated real-world use of the tools. Nevertheless, had we customized the prediction settings in each tool, it is likely that the findings would have differed.

Because the records were presented in random order, the training sets differed for each review across the tools. Although this could have affected the findings, it is also reflective of how the tools would be applied in real-world evidence synthesis projects. In a previous evaluation, we found that Abstrackr's predictions did not differ substantially across three trials when employing this approach.[21] The only way to maintain consistency across the training sets would have been to present them in order by record identification number. Such an approach would introduce bias because all records from a single database and by individual authors would be screened sequentially.

We used a 200-record training set and a small sample of three SRs. As demonstrated herein, the size of the dataset will affect the resulting predictions. This study needs to be replicated on a larger sample of SRs and across a greater number of tools to identify which are the most promising and determine which screening tasks might be most amenable to semi-automation. Our findings should not be generalized to other tools, SRs, or semi-automated screening approaches. We did not investigate the impact of the missed studies on the results of the SRs, an important consideration for systematic reviewers. Future studies should plan for the time and resources to undertake these analyses in their protocols.

Given the retrospective nature of our study, time savings was estimated based on the reduced screening workload and a standard screening rate. This estimate did not account for time spent troubleshooting usability issues. It also did not account for variability in the time spent screening records as reviewers progress through the screening task, or for obviously excluded compared to records of uncertain relevance.[31] Prospective evaluations (i.e., alongside traditional SR processes) are needed to determine the true time savings that may be gained with semi-automated approaches.

Conclusions

Using Abstrackr's predictions to complement the work of a single screener reduced the number of studies that were erroneously excluded by up to 90%, although performance varied by review. RobotAnalyst provided a lesser advantage compared to Abstrackr, and Distiller provided no advantage over single-reviewer screening. In light of the workload and time savings, using Abstrackr to complement the work of a single screener may be acceptable in some cases; however, additional evaluations on larger samples of reviews are needed before this approach could be recommended. Although using any tool to automatically exclude irrelevant records could save substantial amounts of time, the potential for erroneously excluding large numbers of relevant records made the approach far more risky. The usability of the tools was highly variable. Further research is needed to inform how machine learning might be best applied to reduce screening workloads, and to identify the types of screening tasks that are most suitable to semi-automation. Designing (or refining existing) tools based on reviewers' self-identified preferences may improve their usability and enhance adoption.

References

1. Borah R, Brown AW, Capers PL, et al. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open. 2017 Feb;7(2):e012545. doi:10.1136/bmjopen-2016-012545. PMID: 28242767.

2. Chandler J HJ, Deeks JJ, Davenport C, Clarke MJ. Chapter 1: Introduction. In: Higgins JPT, Churchill R, Chandler J, et al. (editors), Cochrane Handbook for Systematic Reviews of Interventions. Vol. 5.2.0. Cochrane; 2017.

3. Lefebvre C, Manheimer E, Glanville J. Chapter 6: Searching for studies. In: Higgins JPT, Green S (editors), Cochrane Handbook for Systematic Reviews of Interventions. Vol. 5.1.0. Cochrane; 2011.

4. Waffenschmidt S, Knelangen M, Sieben W, Bühn S, Pieper D. Single screening versus conventional double screening for study selection in systematic reviews: a methodological systematic review. BMC Med Res Methodol. 2019 Jun 28;19(1):132. doi: 10.1186/s12874-019-0782-0. PMID: 31253092.

5. Thomas J, McNaught J, Ananiadou S. Applications of text mining within systematic reviews. Res Synth Methods. 2011 Mar;2(1):1-14. doi:10.1002/jrsm.27. PMID: 26061596.

6. Tsafnat G, Glasziou P, Choong MK, et al. Systematic review automation technologies. Syst Rev. 2014 Jul;3(1):74. doi:10.1186/2046-4053-3-74. PMID: 25005128.

7. Beller E, Clark J, Tsafnat G, et al. Making progress with the automation of systematic reviews: principles of the International Collaboration for the Automation of Systematic Reviews (ICASR). Syst Rev. 2018 May;7(1):77. doi: 10.1186/s13643-018-0740-7. PMID: 29778096.

8. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst Rev. 2019 Jul 11;8(1):163. doi: 10.1186/s13643-019-1074-9. PMID: 31296265.

9. 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. Syst Rev. 2015 Jan;4(1):5. doi:10.1186/2046-4053-4-5. PMID: 25588314.

10. O'Connor AM, Tsafnat G, Gilbert SB, et al. Moving toward the automation of the systematic review process: a summary of discussions at the second meeting of International Collaboration for the Automation of Systematic Reviews (ICASR). Syst Rev. 2018 Jan 9;7(1):3. doi: 10.1186/s13643-017-0667-4. PMID: 29316980.

11. Thomas J. Diffusion of innovation in systematic review methodology: why is study selection not yet assisted by automation? OA Evidence-Based Medicine. 2013;1(2):12.

12. 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.

13. van Altena AJ, Spijker R, Olabarriaga SD. Usage of automation tools in systematic reviews. Res Synth Methods. 2019 Mar;10(1):72-82. doi:10.1002/jrsm.1335. PMID: 30561081.

14. Paynter R, Bañez LL, Berliner E, et al. EPC Methods: An Exploration of the Use of Text-Mining Software in Systematic Reviews. Report No.: 16-EHC023-EF. Rockville, MD: Agency for Healthcare Research and Quality; 2016. PMID: 27195359.

15. O'Connor AM, Tsafnat G, Gilbert SB, et al. Still moving toward automation of the systematic review process: a summary of discussions at the third meeting of the International Collaboration for Automation of Systematic Reviews (ICASR). Syst Rev. 2019 Feb 20;8(1):57. doi: 10.1186/s13643-019-0975-y. PMID: 30786933.

16. Marshall C. Systematic Review Toolbox. 2019. http://systematicreviewtools.com/index.php. Accessed April 5, 2019.

17. Wallace BC, Small K, Brodley CE, et al. Deploying An Interactive Machine Learning System In An Evidence-Based Practice Center: Abstrackr. Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium; 2012 Jan 28-30; New York, New York: Association for Computing Machinery; 2012. doi:10.1145/2110363.2110464.

18. The National Centre for Text Mining. RobotAnalyst. 2019. http://www.nactem.ac.uk/robotanalyst/. Accessed April 5, 2019.

19. Evidence Partners. Publications. 2019. https://www.evidencepartners.com/about/publications/. Accessed April 5, 2019.

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

21. Gates A, Johnson C, Hartling LJSr. Technology-assisted title and abstract screening for systematic reviews: a retrospective evaluation of the Abstrackr machine learning tool. Syst Rev. 2018 Mar;7(1):45. Doi:10.1186/s13643-0.18-0707-8. PMID: 29530097.

22. 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;4(1):80. doi:26073974. PMID: 26073974.

23. Pillay J, Boylan K, Carrey N, et al. First-and Second-Generation Antipsychotics In Children And Young Adults: Systematic Review Update. Report No.: 17-EHC001-EF. Rockville, MD: Agency for Healthcare Research and Quality; 2017 Mar. PMID: 28749632.

24. Pillay J, Freeman EE, Hodge W, et al. Screening For Impaired Visual Acuity And Vision-Related Functional Limitations In Adults 65 Years And Older In Primary Health Care: Systematic Review. Edmonton, AB: Evidence Review Synthesis Centre; 2017 Nov. (http://canadiantaskforce.ca/ctfphc-guidelines/overview/).

25. Evidence Partners. DistillerAI FAQs. 2019. https://www.evidencepartners.com/resources/distillerai-faqs/. Accessed April 5, 2019.

26. Harris PA, Taylor R, Thielke R, et al. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009 Apr;42(2):377-81. doi:10.1016/j.jbi.2008.08.010. PMID: 18929686.

27. Brooke J. SUS-A quick and dirty usability scale. Usability Evaluation in Industry. 1996;189(194):4-7.

28. Wallace BC, Trikalinos TA, Lau J, et al. Semi-automated screening of biomedical citations for systematic reviews. Bioinformatics. 2010 Jan;11(1):55. doi:10.1186/1471-2105-11-55. PMID: 20102628.

29. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. International Journal of Human-Computer Interaction. 2008;24(6):574-94. doi:10.1080/10447310802205776.

30. Vaismoradi M, Turunen H, Bondas TJN, et al. Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study. Nurs Health Sci. 2013 Sep;15(3):398-405. doi:10.1111/nhs.12048. PMID: 23480423.

31. Wallace BC, Small K, Brodley CE, et al. Modeling Annotation Time to Reduce Workload in Comparative Effectiveness Reviews. Proceedings of the 1st ACM International Health Informatics Symposium; 2010 Nov 11-12. New York, New York: Association for Computing Machinery; 2010. doi:1-/1145/1882992.1882999.

32. Chen Y. Developing Stopping Rules for a Machine Learning System in Citation Screening [dissertation]. Providence, RI: Brown University; 2019.

33. Olorisade BK, Brereton P, Andras P. Reproducibility of studies on text mining for citation screening in systematic reviews: evaluation and checklist. J Biomed Inform. 2017 Sep;73:1-13. doi: 10.1016/j.jbi.2017.07.010. PMID: 28711679.

34. Olorisade BK, Quincey E, Brereton P, et al. A Critical Analysis of Studies that Address the Use of Text Mining for Citation Screening in Systematic Reviews. Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering; 2016 Jun 1-3. Limerick, Ireland: Association for Computing Machinery; 2016. doi:10.1145/2915970.2915982.

Abbreviations and Acronyms

AHRQ	Agency for Healthcare Research and Quality
ARCHE	Alberta Research Centre for Health Evidence
EPC	Evidence-based Practice Center
IQR	Interquartile range
nRCT	Non-randomized controlled trial
PICOS	Population, Intervention, Comparator, Outcome, Study design
PROSPERO	International prospective register of systematic reviews
RCT	Randomized controlled trial
RIS	Research Information Systems
RSV	Respiratory syncytial virus
SR	Systematic review
UAEPC	University of Alberta Evidence-based Practice Center

Appendix A. Screening Exercise for the User Experiences Testing

Screening Exercise

Before completing the survey, we ask that you undertake a brief screening exercise in each of the tools. The survey will ask you about your experiences using each tool, so keep note of anything that you like or dislike, or any difficulties that you encounter.

You will be screening records for a review on digital technologies for pain. The search found 2662 records. The RIS file, which contains all 2662 records, was attached to the invitation e-mail. The eligibility criteria for the review were also attached to the invitation e-mail.

Instructions for Each Tool

The goal of this exercise is to, in each tool: (a) create a project, (b) upload the studies to be screened, (c) screen a small training set, (d) download a record of the screening that you completed, and (e) download the predicted relevance of the remaining studies.

Feel free to investigate the various functions in each tool, and to use the "help" function if you are having trouble. Don't worry if you cannot complete the task. We are interested in your experiences, both positive and negative.

The following includes instructions for each tool. Complete the exercise in each tool in the random order that was assigned to you.

Abstrackr

1. Navigate to the Abstrackr website: http://abstrackr.cebm.brown.edu/account/login
2. Register to create an account and log in.
3. Create a new project where one reviewer will screen the studies ("Single-screen" mode) and the studies will be presented in "most likely to be relevant" order.
4. Screen 100-200 studies.
5. Download a record of the decisions for the studies that you screened. Open the file to see what the output looks like. Consider whether this is practical format or not.
6. Check for the availability of predictions of the relevance of the remaining studies.
7. If no predictions are available, check back within 24 hours to see if they become available (the server updates overnight).
8. If predictions are available, download a record of the predicted relevance of the remaining studies. Open the file to see what the output looks like. Consider whether this is practical format or not.

DistillerSR

1. Navigate to the DistillerSR website: https://v2dis-prod.evidencepartners.com/Login/Login.php
2. Log in using the following username and password: [username] [password]
3. Contact Allison [email address] who will assign you to a project.

4. Upload the references from the RIS file to your project.
5. Screen 50 studies.
6. Download a record of the decisions for the studies that you screened from the “Datarama” as an Excel spreadsheet. Open the file to see what the output looks like. Consider whether this is practical format or not.
7. Check for the availability of predictions (“DistillerAI”). If they are not available, keep screening until they are ready.
8. Run DistillerAI using the “simple review” setting to generate predictions for the remaining studies.
9. Download a record of the predicted relevance of the remaining studies from “Datarama” as an Excel spreadsheet. Open the file to see what the output looks like. Consider whether this is practical format or not.

RobotAnalyst

1. Navigate to the RobotAnalyst website: http://nactem.ac.uk/RA/login.html
2. Log in using the following username and password: [username] [password]
3. Note: if you have trouble logging in, try reloading the page. If this does not work, try clearing your browser history.
4. Create a new collection where the default label is “undecided” and the studies are screened in random order.
5. Screen 50 studies.
6. Download a record of the decisions for the studies that you screened. Open the file to see what the output looks like. Consider whether this is practical format or not.
7. Update the predictions. If the predictions are not yet available, keep screening until they are ready.
8. Download a record of the predicted relevance of the remaining studies. Open the file to see what the output looks like. Consider whether this is a practical format or not.

You’re done! We would like to know what you thought about each of these tools. Please don’t forget to fill out the user experiences survey: [link to survey]

Appendix B. User Experiences Survey

Thank you for taking part in this study to help us understand user experiences with machine learning tools for screening in systematic reviews (Pro00087862). As a reminder, your participation in this study is voluntary. Once you have completed the survey, it will be impossible to withdraw from the study.

By completing this survey, you agree that: you have read the information and recruitment letter and the study has been explained to you, you have been given the opportunity to ask questions and your questions have been answered, you have been told who to contact if you have further questions, and you agree to participate in the study as described in the recruitment and information letter. Completion of the survey will imply consent.

Thank you!

[Attachment: recruitment and information letter]

1. **Reflecting on your experiences screening in Abstrackr, to what extent do you agree with the following statements?** [scale of 1 (strongly disagree) to 5 (strongly agree)]

   a. I think that I would like to use Abstrackr frequently.
   b. I found Abstrackr to be unnecessarily complex.
   c. I thought Abstrackr was easy to use.
   d. I think that I would need the support of a technical person to be able to use Abstrackr.
   e. I found the various functions in Abstrackr were well integrated.
   f. I thought there was too much inconsistency in Abstrackr.
   g. I would imagine that most people would learn to use Abstrackr very quickly.
   h. I found Abstrackr very cumbersome to use.
   i. I felt very confident using Abstrackr.
   j. I needed to learn a lot of things before I could get going with Abstrackr.

2. **Please provide any positive or negative comments related to your experiences screening in Abstrackr.** [free-text responses]

3. **Reflecting on your experiences screening in DistillerSR, to what extent do you agree with the following statements?** [scale of 1 (strongly disagree) to 5 (strongly agree)]

   a. I think that I would like to use DistillerSR frequently.
   b. I found DistillerSR to be unnecessarily complex.
   c. I thought DistillerSR was easy to use.
   d. I think that I would need the support of a technical person to be able to use DistillerSR.
   e. I found the various functions in DistillerSR were well integrated.
   f. I thought there was too much inconsistency in DistillerSR.
   g. I would imagine that most people would learn to use DistillerSR very quickly.
   h. I found DistillerSR very cumbersome to use.
   i. I felt very confident using DistillerSR.
   j. I needed to learn a lot of things before I could get going with DistillerSR.

4. **Please provide any positive or negative comments related to your experiences screening in DistillerSR.** [free-text responses]

5. **Reflecting on your experiences screening in RobotAnalyst, to what extent do you agree with the following statements?** [scale of 1 (strongly disagree) to 5 (strongly agree)]

   a. I think that I would like to use RobotAnalyst frequently.
   b. I found RobotAnalyst to be unnecessarily complex.
   c. I thought RobotAnalyst was easy to use.
   d. I think that I would need the support of a technical person to be able to use RobotAnalyst.
   e. I found the various functions in RobotAnalyst were well integrated.
   f. I thought there was too much inconsistency in RobotAnalyst.
   g. I would imagine that most people would learn to use RobotAnalyst very quickly.
   h. I found RobotAnalyst very cumbersome to use.
   i. I felt very confident using RobotAnalyst.
   j. I needed to learn a lot of things before I could get going with RobotAnalyst.

6. **Please provide any positive or negative comments related to your experiences screening in RobotAnalyst.** [free-text responses]

7. **Considering your experiences with the three tools, which would you prefer to use for screening in a systematic review?**

   a. Abstrackr [first choice, second choice, or third choice]
   b. DistillerSR [first choice, second choice, or third choice]
   c. RobotAnalyst [first choice, second choice, or third choice]

8. **Which features of any of the tools support their usability and appeal? Please explain.** [free-text responses]

9. **Which features of any of the tools hinder their usability and appeal? Please explain.** [free-text responses]

10. **If you any additional comments related to your experiences with the three tools, please include them here.** [free-text responses]

Appendix C. 2x2 Tables and Calculations for the Performance Metrics (Example from the Antipsychotics Review in Abstrackr)

2x2 Cross-tabulations

Automated Simulation:

	Excluded from final report	Included in final report	Row total
Excluded by Simulation	11450	36	11486
Included by Simulation	579	91	670
Column total	12029	127	12156

Semi-automated Simulation:

	Excluded from final report	Included in final report	Row total
Excluded by Simulation	11101	2	11103
Included by Simulation	928	125	1053
Column total	12029	127	12156

Predictions were available after screening 200 records. Abstrackr predicted that 2117 of the remaining records were relevant and 9839 were irrelevant.

Sample Calculations

Proportion Missed (i.e., error)

Of the studies included in the final report, the proportion that would have been excluded at title and abstract screening for each simulation.

Automated Simulation: proportion missed = 36 / 127 = 0.28 or 28%
Semi-automated Simulation: proportion missed = 2 / 127 = 0.016 or 1.6%

Workload savings (i.e., absolute screening reduction)

Of the number of records that would need to be screened at the title and abstract stage (assuming dual independent screening), the proportion that would not need to be screened manually.

Automated Simulation: workload savings = (9839 + 12156) / (12156 x 2) = 0.90 or 90%
Semi-automated Simulation: workload savings = 9839 / (12156 x 2) = 0.40 or 40%

Time Savings

The time saved by not screening records manually, assuming a screening rate of 0.5 minutes per record and an 8-hour work day.

Automated Simulation: time savings = [(9839 +12156) x 0.5 min/record] x 1 hour/60 min x 1 day/8 hours = 23 days

Semi-automated Simulation: time savings = (9839 x 0.5 min/record) x 1 hour/60 min x 1 day/8 hours = 10 days

Appendix D. 2 x 2 Cross-tabulations for Each Review in Each Tool

2x2 Tables for the Automated Simulation

Antipsychotics, Abstrackr

	Excluded from final report	Included in final report	Row total
Excluded by Simulation	11450	36	11486
Included by Simulation	579	91	670
Column total	12029	127	12156

Antipsychotics, DistillerSR

	Excluded from final report	Included in final report	Row total
Excluded by Simulation	12011	123	12134
Included by Simulation	18	4	22
Column total	12029	127	12156

Antipsychotics, RobotAnalyst

	Excluded from final report	Included in final report	Row total
Excluded by Simulation	8558	89	8647
Included by Simulation	3471	38	3509
Column total	12029	127	12156

Bronchiolitis, Abstrackr

	Excluded from final report	Included in final report	Row total
Excluded by Simulation	5559	7	5566
Included by Simulation	165	130	295
Column total	5724	137	5861

Bronchiolitis, DistillerSR

	Excluded from final report	Included in final report	Row total
Excluded by Simulation	5685	131	5816
Included by Simulation	39	6	45
Column total	5724	137	5861

Bronchiolitis, RobotAnalyst

	Excluded from final report	Included in final report	Row total
Excluded by Simulation	5548	31	5579
Included by Simulation	176	106	282
Column total	5724	137	5861

Visual Acuity, Abstrackr

	Excluded from final report	Included in final report	Row total
Excluded by Simulation	11109	0	11109
Included by Simulation	119	1	120
Column total	11228	1	11229

Visual Acuity, DistillerSR

	Excluded from final report	Included in final report	Row total
Excluded by Simulation	11226	1	11227
Included by Simulation	2	0	2
Column total	11228	1	11229

Visual Acuity, RobotAnalyst

	Excluded from final report	Included in final report	Row total
Excluded by Simulation	11148	1	11149
Included by Simulation	80	0	80
Column total	11228	1	11229

2x2 Tables for Semi-automated Simulation

Antipsychotics, Abstrackr

	Excluded from final report	Included in final report	Row total
Excluded by Simulation	11101	2	11103
Included by Simulation	928	125	1053
Column total	12029	127	12156

Antipsychotics, DistillerSR

	Excluded from final report	Included in final report	Row total
Excluded by Simulation	11165	2	11167
Included by Simulation	864	125	989
Column total	12029	127	12156

Antipsychotics, RobotAnalyst

	Excluded from final report	Included in final report	Row total
Excluded by Simulation	7980	3	7983
Included by Simulation	4049	124	4173
Column total	12029	127	12156

Bronchiolitis, Abstrackr

	Excluded from final report	Included in final report	Row total
Excluded by Simulation	5357	1	5358
Included by Simulation	367	136	503
Column total	5724	137	5861

Bronchiolitis, DistillerSR

	Excluded from final report	Included in final report	Row total
Excluded by Simulation	5394	10	5404
Included by Simulation	330	127	457
Column total	5724	137	5861

Bronchiolitis, RobotAnalyst

	Excluded from final report	Included in final report	Row total
Excluded by Simulation	5364	6	5370
Included by Simulation	360	131	491
Column total	5724	137	5861

Visual Acuity, Abstrackr

	Excluded from final report	Included in final report	Row total
Excluded by Simulation	11075	0	11075
Included by Simulation	153	1	154
Column total	11228	1	11229

Visual Acuity, DistillerSR

	Excluded from final report	Included in final report	Row total
Excluded by Simulation	11074	0	11074
Included by Simulation	154	1	155
Column total	11228	1	11229

Visual Acuity, RobotAnalyst

	Excluded from final report	Included in final report	Row total
Excluded by Simulation	11042	0	11042
Included by Simulation	186	1	187
Column total	11228	1	11229

Appendix E. Focused Codes and Supporting Quotes for the Properties of Each Tool

Comments Related to Abstrackr

Properties and focused codes	Supporting quotes
User friendliness *Positives:* user friendly/easy to screen records; easy to navigate; easy to make a new review; relatively simple program; makes screening more enjoyable *Negatives:* difficult to figure out how to upload records; changing the review settings is arduous; not very intuitive; not the most user friendly; interface could be improved to streamline processes for project management	"The user interface is relatively appealing, but not distracting. This is a relatively simple program, and when it works, it makes screening a bit more enjoyable." "Very rudimentary graphics, display and options, but this also made Abstrackr very easy to use; probably the easiest to use as not many functionalities and very easy to find […]" "Easy to use for screening but not very intuitive when trying to export records or go back to main page […]" "Instructions said to create project where studies would be presented in "most likely to be relevant" order - I couldn't figure out how to select that option."
Qualities of the user interface *Positives:* appealing user interface; not distracting; nice look and layout *Negatives:* not the most pretty looking; very rudimentary graphics and display	"I liked the look; lay-out for screening; easy to use and advanced to next record easily and quickly." "I wanted to say also that this tool is not the most pretty looking; user friendly, but does at least appear trustworthy, so I would use it again."
Features and functions *Positives:* can add notes and tags to the records; can select single or dual screen mode; can change the order of the records; can change decision in the case of mistakes *Negatives:* user guide is not that helpful; unclear how to change the order of the records; record IDs not shown; unclear how or if the most likely to be relevant prioritization works; unclear if decisions file will include tags; unclear how the tagging works	"For first timers, it is difficult to figure out how to upload the records from EndNote into Abstrackr, and the user help guide was not very helpful." "I do like that you can tag studies (e.g., SRs) but not sure how it all works or if the decisions file actually displays these." "I also like that you can change your decisions if you make a mistake, but the process to do so is a bit cumbersome (have to go through a few pages)."
Trustworthiness *Positives:* appears trustworthy; would use again; advanced to next record easily and quickly *Negatives:* sometimes slow or crashes; unclear error messages; unpredictable user interface; server seems slow; delays in uploading records	"I wanted to say also that this tool is not the most pretty looking; user friendly, but does at least appear trustworthy, so I would use it again." "When uploading the records, the program is sometimes very slow or can crash. You get a bright orange screen with little indication as to what you may have done wrong. Spontaneously (or so it seems), the program will start working again."

Properties and focused codes	Supporting quotes
Ease and speed of obtaining the predictions *Positives:* waiting for the predictions is not a deal breaker/not a big issue given the time savings *Negatives:* slow to develop predictions; requires a larger training set than other programs; predictions cannot be updated manually; must wait overnight (or hours) for predictions, which is long compared to other programs; no way to know how many records will need to be screened before you get predictions	"One last thing is that it is a bit annoying to have to wait a day for the predictions to appear. This is especially the case since there is no way to know how many records will need to be screened before you get predictions." "Compared to the other programs, it seems like Abstrackr requires a larger training set before providing predictions. Also, the predictions cannot be updated manually by the user; instead, one must wait overnight for the predictions to be produced. This is a little bit inconvenient, but probably not a deal breaker for me." "Waiting time on predictions is a bit lengthy compared to other machine learning programs but considering the time saved in screening overall it isn't a big issue."
Practicality of the export file(s) *Positives:* consensus column; contains information that is helpful when reviewing the screening decisions (e.g., title, authors); seems usable and practical; could easily import into a different format or program *Negatives:* year published is missing; does not always include the original reference IDs; format is not user friendly; would need a lot of work before they would be usable; hard to use/might be tedious; use might be prone to error; unclear why -1 and 1 labels are used	"Output format seems practical as it contains the authors, titles and abstracts which will be helpful in reviewing screening decisions" "The format for downloading the predictions and screening record is not very user friendly. Sometimes, the file does not include the original reference IDs […]" "Downloaded record of decisions for screened studies - into excel using CSV format; format seemed usable and could likely import this easily into other formats; programs." "[…] quite a bit of work is required to reformat the files before they would be usable for a reviewer." "I did not find the coding in the Excel spreadsheet explained anywhere (0, 1, -1); found the export to be practical"

Comments Related to DistillerSR

Properties and focused codes	Supporting quotes
User friendliness *Positives:* user friendly; easy to use; easy to navigate; could find right options after spending time looking at menus; easy to track progress (record IDs follow flow of records) *Negatives:* additional features make screening cumbersome; may require training to use it more efficiently; requires more skill to set up; unnecessarily complex; sometimes difficult to navigate; initially did not know where to find needed functions; had to watch a tutorial to set up the screening process	"Once everything is set up the screening is very easy and the process quite user-friendly" "Distiller has a very appealing user interface, and once the project is set up, it is pretty easy to use. That said, it requires much more skill to actually set up a project in Distiller compared to the other programs. Since Distiller has so many more functions […] for screening it is more cumbersome than the other two programs." "The format of screening is nice, clear and consistent (e.g., abstract is organized with white space). During screening, the order of references followed the order of screening so it was easy to track progress."

Properties and focused codes	Supporting quotes
Qualities of the user interface *Positives:* very appealing user interface; clean and bright; easy interface for screening; clear; consistent; abstract is organized with white space; very nice to look at *Negatives:* a lot “going on”/a lot of information on each screen and on drop-down menus; a bit overwhelming at first	“My favourite part about this program is probably its clean and bright user interface […]” “[…] the interface is very nice to look at.” “[…] a lot of information on each screen and dropdown menu so initially did not know where to find the needed functions” “[…] I liked the interface and colours [..]”
Features and functions *Positives:* tutorial videos are quite helpful; ”drop files” function is a great feature; help function is very useful *Negatives:* too many features; not easy to go back and fix mistakes; requires two clicks to make a decision; unclear why there are two “submit form” buttons; tutorial videos went through things too slowly; having choices listed consecutively (i.e., vertically compared to horizontally) could lead to errors	“Distiller seemed like a very sophisticated tool, with lots of options etc. to choose from, which made it a bit overwhelming at first.” “[…] the software has too many features that makes it look like unnecessarily complex.” “[…] and I liked the little ? symbols which provided tips and explanations.” “Distiller seemed complex, there was a lot "going on" and sometimes I found it hard to navigate through the different pages, project data sheets, etc” “The "drop files" function for uploading records was a great feature.”
Trustworthiness *Positives:* more professional looking than Abstrackr; server is very fast and responsive; reliable/trustworthy; advances well	“[…] it was more professional looking than Abstrackr.” “My favourite part about this program is probably its clean and bright user interface, quick server, and reliability.” “Of the tools I think that this one is the one that I would trust the most with my records. It seemed to have a professional backing to it.”
Ease and speed of obtaining the predictions *Positives:* predictions become available quickly; predictions can be applied in a matter of seconds; running DistillerAI was easy; faster than Abstrackr *Negatives:* seems ahead of its time; unclear if predictions can be removed from the review; couldn’t figure out how to get predictions; took me forever to find Distiller AI; need a tutorial on best settings to use for DistillerAI; unclear how to know best threshold for setting the predictions	“Predictions are much faster than Abstrackr, and are available for all studies in less than ten minutes.” “[…] AI feature became available after 106 screens” “Running the AI was relatively easy, though if I were actually to go through with it I feel like I could use a tutorial on the best settings to use. They seem a bit ahead of the times - you can choose the 'accuracy' of the prediction but how do I know what would work best?” “I don't know if you can 'undo' the AI once it's done but that would be nice.”

Properties and focused codes	Supporting quotes
Practicality of the export file(s) *Positives:* great variety of output formats; almost instantaneous output to Excel; download output is very practical; download output is very useful; highly user friendly; would not require much formatting to be usable; not overly complicated *Negatives:* initial output didn't have the information I needed/needed to select correct display options; unsure what "coding terms" meant; lacking author, title, abstract; couldn't figure out how to export the predictions; unnecessarily complicated; missing important information that would make it more usable	"I liked how Distiller had a great variety of output formats for the screening decisions and predictions […]." "Output to Excel file was almost instantaneous" "[…] the Excel file that can be exported is highly user friendly and would not require much formatting to be in a usable format for a review." "Downloading the records was quite easy and the output looks practical to use - basically the REF IDs and the decisions in 2 columns." "Output is less helpful for reviewing decisions as the author, title and abstract are not included when exported to Excel." "I seem to have generated predictions but couldn't figure out how they would be exported."

Comments Related to RobotAnalyst

Properties and focused codes	Supporting quote(s)
User friendliness *Positives:* looks easy to navigate; it was intuitive; it was easy to use; screening form is easy to use and understand; uploading records is simple *Negatives:* meaning of pop-ups and error messages was unclear; pop-ups and error messages were very distracting; screening decisions would disappear and had to re-screen; very difficult to work with; logging in was difficult; screening is cumbersome; uploading the records was a hassle due to slow server speed; did not like having to scroll down to see records; difficult to attempt to track records for which decisions have been recorded; cumbersome	"After screening records, they change colour; however, then they would change back and it would appear like they had not yet been screened. I was not really sure what to make of it. Uploading the records is a hassle, because the server is so slow." "Firstly, the web page is extremely slow, even when performing minor tasks like simply logging in. Uploading the records is simple but ridiculously slow considering it was a relatively small set of records. The fact that you have to actually click on each record to see the abstract is very cumbersome and does not make any sense if the purpose of the program is to facilitate screening. Once you can see the whole record, the drop-down menu to choose include; exclude is cumbersome - would be better to have a radio box (not sure what that's called). The program also took several seconds to register a decision. Then, out of nowhere, it would randomly lose all the previous decisions on the page."
Qualities of the user interface *Positives:* the user interface is okay; landing page looks very nice; liked the use of colours and buttons; very pretty; I liked the layout *Negatives:* a bit busy	"Slow to load but very pretty! I liked the use of colours and buttons […]" "The user interface is okay, but a little busy."

Properties and focused codes	Supporting quote(s)
Features and functions *Positives:* predictions were available on-screen during the screening process *Negatives:* had to open each record to read the abstract; constant pop-up to update the predictions is annoying; does not automatically advance to the next record; lack of logical numbers for records; drop down menu to choose include or exclude is cumbersome; impossible to tell status of upload	"A major downfall of this program is that you need to click on the title to see the abstract, unlike other tools where the abstract and title automatically appear on-screen. This is especially cumbersome given the slow server speed." "Once the predictions are ready, there is a constant pop-up every few records reminding you to update them. It is a little annoying." "[…] lack of logical order of record numbers makes it additionally difficult to attempt to track records for which decisions had been recorded […]"
Trustworthiness *Negatives:* unable to complete the screening unsure what to make of the screening process; seemed glitchy; unreliable; sometimes had to click twice for the screening decisions; seems untrustworthy; program was extremely slow; took several seconds to register a decision; randomly loses all previous decisions on a page; colours would disappear and unclear if decisions had been lost; could not find a way to tag records; steady stream of error messages; could not trust if decisions were being registered; logging in took multiple attempts; sometimes the wrong record opened; program said files were uploaded but they were not; error stops all screening and asks you to contact admin, then resolves itself; takes longer to screen studies due to time waiting for abstracts to load; slowest program of the three; worry that on-screen predictions could bias the screening process; required a few tries to download the predictions	"Logging in was sometimes difficult; the program would seem to shut down and I would need to reset my browser history to get it working again. Once logged in, the server was extremely slow, and there were multiple pop-ups and error messages, none of which I could quite figure out the meaning of, that were highly distracting." "I was getting several error messages while screening, and was unable to complete the screens" "I sort of like that the predictions appear on-screen alongside the records as you are screening; however, I sort of worry whether this would bias the human's screening decisions." "I would not use this program, mostly because it is cumbersome and does not seem trustworthy at all." "There was a steady stream (no exaggeration) of error messages coming up while using the program, such that I could not trust whether it was actually registering any of my decisions or not." "I would never use this program for a systematic review. It is completely unreliable and there is no practical way to download the screening decisions or predictions."
Ease and speed of obtaining the predictions *Positives:* predictions were available quickly *Negatives:* applying the predictions was slow	"Seemed to be quick to apply predictions." "Applying the predictions was slow, which was not too surprising by that point."

Properties and focused codes	Supporting quote(s)
Practicality of the export file(s) *Positives:* Exporting the decisions was not hard *Negatives:* download does not contain the screening decisions and predictions; output download is not practical; predictions had to be added to the download file manually; export file is not usable; report was impossible to understand; record of decisions was poorly organized; difficult to decipher; text download file is useless; did not know what to do with the export file	"Exporting the decisions was not hard, but required a few tries since the first time the program seemed to be thinking for awhile but nothing downloaded. The next time it worked." "When it comes to downloading the predictions I could not find any practical way to do so, such that they would be in a format that could actually be used. The download is actually a .txt file, which can be converted to .ris and opened in EndNote. That said, there is nowhere in this EndNote file where the screening decisions can be found. These have to be added in manually." "[…] record of decisions in notepad poorly organized and difficult to decipher."