Artificial intelligence supports literature screening in medical guideline development: towards up-to-date medical guidelines.

Wouter Harmsen[2], Janke de Groot[2] , Albert Harkema[1], Ingeborg van Dusseldorp[2], Jonathan de Bruin[3], Sofie van den Brand[1], Rens van de Schoot[1]

[1] Department of Methodology and Statistics, Faculty of Social and Behavioral Sciences, Utrecht University, Utrecht, the Netherlands

[2] Knowledge Institute of the Dutch Association of Medical Specialists, Utrecht, the Netherlands

[3] Department of Research and Data Management Services, Information Technology Services, Utrecht University, Utrecht, the Netherlands

Funding: This project was funded by the Dutch Research Council in the "Corona: Fast-track data" (2020/SGW/00909334).

Corresponding author: Rens van de Schoot: Department of Methods and Statistics, Utrecht University, P.O. Box 80.140, 3508TC, Utrecht, The Netherlands; Tel.: +31 302534468; E-mail address: a.g.j.vandeschoot@uu.nl.

Abstract

Objectives: In a time of exponential growth of new evidence supporting clinical decision making, combined with a labor-intensive process of selecting this evidence, there is a need for methods to speed up current processes in order to keep medical guidelines up-to-date. The purpose of this study was to evaluate the performance and feasibility of active learning to support the selection of relevant publications within the context of medical guideline development.

Design: We used a mixed methods design. The manual process of literature selection by two independent clinicians was evaluated in 14 searches by calculating Cohen's Kappa (κ) for interrater reliability. This was followed by a series of simulations investigating the performance of random reading versus using screening prioritization based on active learning..

Main outcome measures: Work Saved over Sampling at 95% recall (WSS@95), percentage Relevant Records Found at reading only 10% of the total number of records (RRF@10) and average time to discovery (ATD). Finally, results were discussed in a reflective dialogue with guideline developers.

Results: Mean κ for manual title-abstract selection by clinicians was 0.50 and varied between -0.01 to 0.87 based on 5021 abstracts. WSS@95 ranged from 50.15% (SD=17.7) based on selection by clinicians, to 69.24% (SD=11.5) based on the selection by research methodologist up to 75.76% (SD=12.2) based on the final full-text inclusion. A similar pattern was seen for RRF@10 ranging from 48.31% (SD= 23.3) to 62.8% (SD=21.20) and to 65.58% (SD=23.25). ATD ranged from 20 to 67 abstracts.

Conclusion: Tools, implementing active learning, such as ASReview, can speed up the process of literature screening within guideline development.

Keywords: guideline development, medical guidelines, text data, natural language processing, active learning, machine learning, systematic reviewing

Introduction

Producing and updating trustworthy medical guidelines is a deliberative process that requires substantial investment of time and resources.[1] In the Netherlands, medical guidelines in specialist care are being developed and revised in a co-production between clinicians and research methodologists. In the Netherlands, there are over 650 medical guidelines, answering to approximately 12.000 clinical questions. An essential element in guideline development is a systematic synthesis of the evidence. This systematic appraisal includes the formulation of clinical questions, selection of relevant sources, a systematic literature review, grading the body of evidence using GRADE,[2] and finally translating the evidence into recommendations for clinical practice. [3]

The synthesis of evidence starts with the translation of a clinical question into a research question through the PICO acronym (Patient, Intervention, Comparison and Outcomes). Hereafter, a medical information specialist systematically searches literature in different databases. Then, literature screening is performed independently by two clinicians who label relevant publications based on in- and exclusion criteria in title-abstract. Once the relevant publications have been selected, a research methodologist, who is more experienced in selecting relevant publications from large datasets, supports further title-abstract selection, assessing the methodological quality of the selected papers. Since a single literature search can easily result in hundreds to thousands of publications, literature screening is time-consuming, with an estimated 0.9 minutes and 7 minutes per reference per reviewer on abstract screening and full text screening, respectively.[4] In an era of exponential growth of new evidence, combined with a labor-intensive process, there is a need for methods to speed up current processes in order to keep medical guidelines up-to-date.

The rapidly evolving field of Artificial Intelligence *(AI)* has allowed the development of tools that assist in finding relevant texts for search tasks.[5] A well-established approach to increase the efficiency of title and abstract screening is screening prioritization [6,7] via *active learning*.[8] Active Learning is found to be extremely effective for systematic reviewing [9–20]

With machine learning models, relevance scores for each publication can be computed. Then, assessors label title-abstracts (relevant versus irrelevant) for each most relevant record and the model iteratively updates its predictions based on the given labels and prioritizes articles that are most likely to be relevant.

Introducing active learning could save tremendous amount of work and time and may open a new window of opportunity in the context of evidence-based guideline development. However, active learning works under the strong assumption that given labels are correct.[21,22] While in research with experienced reviewers, this may be straightforward, in the daily practice of guideline development, working with clinical questions and clinicians, this may be more complex. Most clinicians are not experienced with title-abstract screening and often screen in addition to their daily work. With large numbers of abstracts and limited time, clinicians can become distracted or fatigued, introducing variability in the quality of their annotations. This variability in human performance may hinder the applicability of active learning in guideline development. Given the potential of active learning, but also the more complex context of guideline development, the purpose of this practice-based study was to evaluate the performance and feasibility of active learning for literature screening, and find out how much effort can be saved in the context of medical guideline development.

In what follows, we present the workflow for manual literature screening in guideline development and introduce the setup of active learning. We first compared the performance (i.e., work saved) of literature screening between active learning and the manual selection by simulating 14 clinical questions through three stages of the review process 1) screening by clinicians, 2) screening by clinicians and research methodologist, and 3) final full-text inclusions after expert consensus. We then discuss the performance of active learning in a reflective dialogue and evaluate reasons that facilitate or hamper the performance of active learning in the discussion section. For ease of interpretation, we visualize the performance of active learning for all datasets times the three levels

in so-called, recall plots. Finally, since this is the first study to report on the performance of active learning in guideline development, we also propose new directions for future research.

Methods

Datasets

We selected 14 clinical questions from recently published clinical guidelines containing manually labeled datasets, providing a wide range of type and complexity of clinical questions, see Table 1. The datasets were derived from different guidelines, published between 2019 en 2021, covering different types of questions, e.g., diagnostic, prognostic and intervention type of questions. In order to be sure that the guidelines had been authorized and thus finished, we selected those that are openly published on the Dutch Medical Guideline Database [Richtlijnendatabase.nl]. Per clinical question, two clinicians independently labeled title-abstracts using prespecified in- and exclusion criteria. To evaluate manual literature screening, interrater agreement for categorical items was calculated according to Cohen's Kappa index measure ($\kappa = \frac{P0-Pe}{1-Pe}$). Cohen's Kappa gives relevant information on how manual title-abstract screening is being done (e.g., use of in- and exclusion criteria). Furthermore, in Table 1 we provide the number of abstracts screened, number of title-abstract inclusions, number of final full-text inclusions and total time spent screening all title-abstracts.

The datasets contained (at least) title and abstract of the paper plus the labels relevant/irrelevant for each of the annotators, clinician and research methodologist, and the column with the final inclusion. Duplicates and papers with missing abstracts were removed from the dataset. All datasets can be found on the Open Science Framework page of the project: https://osf.io/vt3n4/.

Simulation

Active Learning

The simulation was conducted with the command line interface of ASReview version v0.16.[23] We used Naïve Bayes as the classifier with TF-IDF as feature extraction technique and the default balancing strategy. Each protocol was simulated with the relevant records as indicated by (1) the clinicians, (2) the clinicians and research methodologist, and (3) the final inclusions, resulting in 3x14=42 simulations.

We analyzed the model performance of active learning by calculating the following three outcome measures: Firstly, the Work Saved over Sampling (WSS), which indicates the reduction in publications needed to be screened at a given level of recall [Cohen, 2006]. WSS is typically measured at a recall level of 95%, WSS@95 reflects the amount of work saved by using active learning at the cost of failing to identify 5% of relevant publications. Note that humans typically fail to find about 10%.[24] Secondly, we computed the metric Relevant Records Found (RRF), which represents the proportion of relevant publications that are found after screening a prespecified percentage of all publications. Here we calculated RRF@10 which represents the percentage of relevant publications found after screening only 10% of all publications. Thirdly, we calculate average time to discovery (ATD), the fraction of non-reviewed relevant publications during the review (except the relevant publications in the initial dataset). The ATD is an indicator of the performance throughout the entire screening process instead of performance at some arbitrary cutoff value. The ATD is computed by taking the average of the Time to Discovery (TD) of all relevant publications. The TD for a given relevant publication *i* is computed as the fraction of publications needed to screen to detect *i*.[10]

For the training data for each simulation, an equal number of runs was induced equal to the number of relevant records in the dataset with each relevant record being a prior inclusion for one run and 10 randomly chosen irrelevant records. In each run, and for every dataset within a protocol, the same 10 irrelevant records have been used. This is done because the starting paper used for the first iteration of the model can have an influence in the performance which is of importance for computing the ATD. In addition, we plotted recall curves to visualize model performance throughout

the entire simulation. Recall curves give information in two directions; they display the number of publications that need to be screened and the number of relevant publications.

All scripts to reproduce the simulations are available at: http://doi.org/10.5281/zenodo.5031390.[25]

Results

Manual Screening

The selected datasets cover seven different medical fields and include intervention, diagnostic and prognostic type of questions. Twenty-four clinicians independently screened 5021 abstracts and selected a total of 339 potentially relevant publications which took 3766 minutes. Mean κ of interrater agreement was 0.50 and varied between -0.01 to 0.87.

From the 339 relevant publications labeled by the clinicians, 166 (=49%) were excluded by the research methodologist due to methodological concerns, and 45 (=13,3%) were additionally excluded based on full-text selection, leaving 128 publications for final full-text inclusion.

Simulation

We ran a total of 42 simulations, but to explain the results we discuss the results for one dataset in detail: *Distal_radius_fractures_approach*. Out of the 195 records identified in the search, 11 (5,64%) were indicated as relevant by the clinicians, 6 (3.08%) by the research methodologist and ultimately only 5 (2.56%) were included in the final protocol. In Figure 1, first row, the number of relevant records found for each simulation run is displayed as a function of the number of records screened for each of the three levels (clinician, research methodologist, final decision). The vertical line indicates when 95% of the relevant records has been found. Zooming in on WSS@95 for full text inclusions, on average, after screening 43% of the records (n= 83), all records (5 out of 5) would have been found. If you would screen records in a random order, at this point you would have found 3 of the relevant records and finding 5 of the relevant records would take on average 186 records. In other words, the time that can be saved using active learning expressed as the percentage of records

that do not have to be screened is 61% (sd= 5.43), while still identifying 95% of the relevant records. The RRF@10 is 20% (sd= 11.18), meaning that after screening 10% of records, 20% of the relevant records have been identified.

Figure 1 presents recall curves for *all* simulations and as can observed the recall curves differ across datasets but always outperform randomly reading the records which is the standard approach. Simulation results are presented in Table 2 and showed that the Work Saved over Sampling (WSS@95) was lowest for clinicians and ranged from 32.31% to 97.99%, with a mean of 50.15% (SD= 17,74); followed by the research methodologist, it ranged from 45.34% to 95.7%, with a mean of 69.24% (SD=11.51); and simulating the full-text inclusions resulted in the highest WSS@95 that ranged from 61.41% to 96.68% (0.92), with a mean of 75.76% (SD=12.16).

>> FIGURE 1 <<

A similar pattern emerged for RRF@10 which, for clinicians, ranged from 28.10% to 85.95%, with a mean of 48.31% (SD= 23.32); for the research methodologist, it ranged from 25.00% to 100%, with a mean of 62.78% (SD=21.20); and simulating full-text inclusions gave a RRF@10 that ranged from 20.00% to 100% (0.92), with a mean of 65.58% (SD=23.25). ATD Ranged from screening 20 to 62 abstracts.

Discussion

The purpose of this practice-based study was to evaluate the performance and feasibility of active learning to support the selection of relevant publications within the context of guideline development. To do so, we evaluated the performance of active learning on labeled datasets from 14 clinical questions and discussed the results with professional guideline developers. The main conclusion is that tools, implementing active learning, such as ASReview, can speed up the process of literature screening within guideline development. The main results of our simulation show a 13% to 98% reduction in the number of papers needed to screen compared to manual screening. When

ASReview was used based on the manual screening by clinicians, the average WSS@95 was 50%. After additional assessment by the research methodologist the average WSS@95 increased to 69%, with a further increase to 75% after final full text inclusion. So, the performance of active learning increases with more accurate title-abstract labelling, which underline the importance of strict in- and exclusion criteria.

In a reflective dialogue of two 3.5 hours sessions seven guideline developers critically appraised the accuracy of the labeled datasets and performance of active learning. The discussion revealed that almost half (=49%) of the selected publications by the clinicians did not meet the predefined inclusion criteria, e.g., PICO-criteria or study design and were therefore re-labeled as irrelevant. This emphasizes the need for methodological support in title-abstract screening, but also that in- and exclusion in guideline development is not always as straight forward as in systematic reviews for research purposes.

In the reflective dialogue we also discussed the performance of ASReview in specific datasets. The recall plots for the dataset *Distal_radius_fractures_approach,* showed that 5 papers were identified as relevant by the clinicians, but were deemed irrelevant by the research methodologists. Especially one paper hampers the performance of active learning and was always found last in the simulation. This paper describes a literature overview, and although it matches the PICO-criteria, the study was excluded because of methodological reasons, i.e. it describes empiric research. For other datasets (i.e., *Shoulder_replacement_surgery, Total_knee_replacement* and *Shoulder_dystocia_positioning),* active learning seems to have difficulty finding systematic reviews and observational studies compared to randomized control trials. As discussed, this may be inherent to the way the abstract is structured; where RCTs often describe a strict comparison, this may be less evident for systematic reviews and observational studies.

Our results are in line with the assumption that active learning works under the strong assumption that given labels are correct.[21,22,26] During our reflective dialogue session, the notion of 'noisy

labels' was introduced for the initial screening process. This notion was confirmed in the low to moderate interrater reliability of the manual title-abstract screening, with an average kappa of 0.5 in line with other recent findings.[27] While independent selection of papers is an important step to reduce bias, it is a time-consuming process depending on the level of experience by the reviewer and clarity of the inclusion/exclusion criteria. Our next question was to find the 'noise' in the manual screening process. When looking at the differences between the selections made by the clinicians and the professional guideline developers, some interesting themes emerged. Guideline developers realized that clinicians often include publications not only based on the PICO-criteria, but also out of personal interest or fear of leaving out important data. Indeed, many articles when re-examined did not fall within the PICO-criteria nor the pre-defined criteria regarding methodological concerns (e.g., RCT vs case control studies or cohort studies). On average there was a 49% drop of inclusion when the research methodologist re-evaluated the original inclusion made by the clinicians.

A question of interest for future study is when to trust that all relevant literature on the topic has been retrieved, not only based on our results, but also others.[28] In this study we plotted recall curves to visualize the performance of active learning and we organized discussion meetings trying to reason why some publications were more difficult to find. Looking at the examples this often happened when the search had followed a slightly different process. In the current workflow, due to limited resources, pragmatic choices are being made to not include all individual studies when a recent systematic review is available. For active learning models it takes time to 'learn' this adapted (non-logical) strategy. For instance plateaus occur in some of the recall plots, and after a series of irrelevant records have been identified suddenly a new relevant record was found. Interestingly, when time is being saved by working with active learning tools, these pragmatic choices might not be necessary anymore and may therefore actually lead to a much larger and a more complete set of inclusions than the manual workflow.

Strengths and weaknesses.

While an obvious weakness concerns the size of this study, the obvious strength includes the evaluation within the daily practice of guideline development using real world data from previous developed guidelines. While there are studies reporting on tools implementing active learning in systematic reviews, there is only little evidence on implementation of such tools in daily practice.[29–31] Our 'real world' data provided us with new challenges, not seen before because it is most frequently tested in research settings,[11,24] leaving out more pragmatic and human-interest choices that influence literature screening.

This type of practice-based study has shown potential ways to use and improve current practice. In our sample, ASReview was able to detect most relevant studies with a significant reduction of the number of abstracts needed to be screened. The system performed better when the inclusion and exclusion criteria were adhered to in a stricter way. This finding brought us to look at our own workflow needing more attention to guide the clinicians in the systematic selection of papers. This is not only beneficial when using ASReview, where the principle of 'better in, better out' seems to apply, but also when using the manual selection of papers. After abstract screening, almost half of the inclusions were incorrect which is higher than error rates reported in systematic reviews; with a mean error rate of nearly 11% over 25 systematic reviews.[24] Methods to improve literature screening have been described in recent papers,[27,32,33] and include recommendations to include reflection and group discussion resulting in a more iterative process, practical tips like taking regular breaks and coding in small batches at a time to prevent fatigue, but also setting up very clear inclusion criteria and adjusting the codebook during the process if needed. While the inclusion by two independent reviewers is often assumed the best way to reduce bias, these authors also advise to regularly assess interrater reliability as part of reflective and learning practice.

We also defined some remaining questions for future research. As described above, in guideline development, research questions do not always yield prior inclusion papers, while the performance of active learning partially depends on relevant starting papers to learn from. A possible solution,

that needs to be explored, might be to start with a dummy abstract containing all relevant elements from the PICO. At the same time, we need more samples of research questions in clinical guidelines to further evaluate the use of AI tools in different types of questions and contexts. In this study, we used only one tool in a limited set of retrospective data, future studies should include different AI tools within the actual process of guideline development, to further evaluate the human-machine interaction and how this affects the process of guideline development.

Conclusions

This study shows a reduction of 50-75% in abstracts that needed to be screened to find and select all relevant literature for inclusion in medical guidelines when using ASReview. A next step would be to evaluate how to apply active learning in the workflow of guideline development, and what it means for both the timeframe to develop new recommendations and the transparency and quality of these evidence based recommendations.

Availability of data and materials

All scripts that were used during this study, including preprocessing, analysing and simulation scripts for results, figures and tables published in this paper can be found on Zenodo: http://doi.org/10.5281/zenodo.5031390.[25]. The 14 systematic review datasets are openly available at the Open Science Framework [REF]

277

278 **Table 1**. Descriptive characteristics of the purposefully selected datasets (n=14)

#	Guideline topic	Medical specialty	Type of question	N	Screening time (min)	K
1	Radial fractures approach	General surgery	Intervention	195	225	0.31
2	Radial fractures closed reduction	General surgery	Prognostic	277	294	0.55
3	Halux Valgus prognostic	Orthopedic surgery	Prognostic	640	327	0.64
4	Head and neck cancer bone	Otolaryngology	Diagnostic	311	253	0.87
5	Head and neck cancer imaging	Otolaryngology	Diagnostic	56	72	0.61
6	Obstetric emergency training	Obstetrics	Intervention	188	275	0.61
7	Post-intensive care treatment	Rehabilitation	Intervention	435	388	0.05
8	Pregnancy medication	Obstetrics	Intervention	428	243	0.66
9	Shoulder replacement diagnostic	Radiology	Intervention	215	123	0.59
10	Shoulder replacement surgery	Orthopedic surgery	Intervention	335	270	0.57
11	Shoulder dystocia positioning	Gynecology	Diagnostic	342	366	0.49
12	Shoulder dystocia recurrence	Gynecology	Intervention	397	172	-0,01
13	Total knee replacement	Orthopedic surgery	Intervention	480	262	0.55
14	Vascular access	General surgery	Intervention	722	496	0.51

Table 2. Results from simulation analyses for datasets labeled by clinicians, research methodologist and full-text selection.

#	N	$Select_{Cl}$	$Select_{Ex}$	$Select_{FT}$	$WSS95_{Cl}$	$WSS95_{Ex}$	$WSS95_{FT}$	$RRF10_{Cl}$	$RRF10_{Ex}$	$RRF10_{FT}$
1	195	11	6	5	32.31 (6.37)	57.70 (4.80)	61.41 (2.14)	29.09 (8.31)	30.00 (16.73)	20.00 (11.18)
2	277	8	4	4	43.33 (5.47)	59.40 (6.82)	62.31 (9.15)	28.57 (13.23)	25.00 (16.67)	33.33 (27.22)
3	640	20	14	12	55.76 (2.54)	73.55 (1.54)	77.40 (2.45)	43.16 (5.56)	52.75 (7.90)	62.88 (10.59)
4	311	34	20	11	73.15 (1.43)	72.98 (2.48)	78.12 (4.10)	66.22 (4.10)	71.32 (11.39)	73.64 (9.24)
5	56	18	9	8	48.89 (0.00)	70.12 (3.35)	70.28 (3.13)	28.10 (2.52)	45.83 (12.50)	41.07 (5.05)
6	188	18	12	7	40.33 (2.47)	45.34 (8.99)	86.76 (1.78)	47.06 (6.69)	40.15 (11.92)	78.57 (15.85)
7	435	109	22	6	32.70 (1.25)	66.10 (1.08)	64.19 (11.75)	25.63 (3.31)	62.55 (5.93)	46.67 (20.66)
8	428	45	45	45	66.42 (1.26)	66.34 (1.08)	66.92 (1.23)	60.45 (5.34)	61.26 (5.52)	60.86 (5.31)
9	342	3	1	1	97.99 (0.70)	NA	NA	100.00 (0)	NA	NA
10	397	6	4	4	74.78 (2.53)	93.78 (0.87)	93.13 (0.15)	63.33 (8.16)	100.00 (0)	100.00 (0)
11	218	6	5	4	79.55 (1.82)	82.80 (0.43)	79.95 (2.51)	46.67 (20.66)	40.00 (13.69)	33.33 (0)
12	335	5	5	4	61.30 (14.84)	61.05 (14.20)	96.68 (0.92)	65.00 (22.36)	60.00 (33.54)	100.00 (0)
13	480	35	16	9	65.09 (4.03)	73.12 (4.00)	95.76 (0.34)	78.74 (10.14)	89.17 (16.67)	100.00 (0)
14	772	21	10	8	34.84 (16.34)	95.72 (0.32)	96.27 (0.48)	85.95 (19.72)	100.00 (0)	100.00 (0)
total	5074	339	173	128	50.15 (17.14)	69.24 (11.51)	75.76 (12.16)	48.31 (23.32)	62.78 (21.20)	65.58 (23.25)

$Select_{Cl, Ex, FT}$ = number of records included by clinician, research methodologist and full-text selection;

$WSS95_{Cl, Ex, FT}$ = Work Saved over Sampling measured at a recall level of 95% for dataset labeled by clinician, research methodologist and full-text selection;

$RRF10_{Cl, Ex, FT}$ = Relevant References Found after screening 10% of all publications (RRF10) for dataset labeled by clinician, research methodologist and full-text selection.

FIGURE 1

NOTE: Y-axis presents the number of relevant papers minus one paper, selected for training data.

* Analyses of the dataset *Shoulder_replacement_diagnostic*, showed no simulations because no relevant papers were included.

References

1 Graham RP, Mancher M, Wolman D, *et al.* COMMITTEE ON STANDARDS FOR DEVELOPING TRUSTWORTHY CLINICAL PRACTICE GUIDELINES. 2011.

2 Guyatt GH, Oxman AD, Vist GE, *et al.* GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;**336**. doi:10.1136/bmj.39489.470347.AD

3 Guyatt GH, Oxman AD, Kunz R, *et al.* Going from evidence to recommendations. *BMJ* 2008;**336**. doi:10.1136/bmj.39493.646875.AE

4 Wang Z, Asi N, Elraiyah TA, *et al.* Dual computer monitors to increase efficiency of conducting systematic reviews. *Journal of Clinical Epidemiology* 2014;**67**. doi:10.1016/j.jclinepi.2014.06.011

5 Harrison H, Griffin SJ, Kuhn I, *et al.* Software tools to support title and abstract screening for systematic reviews in healthcare: an evaluation. *BMC Medical Research Methodology* 2020;**20**. doi:10.1186/s12874-020-0897-3

6 Cohen AM, Ambert K, McDonagh M. Cross-Topic Learning for Work Prioritization in Systematic Review Creation and Update. *Journal of the American Medical Informatics Association* 2009;**16**. doi:10.1197/jamia.M3162

7 Shemilt I, Simon A, Hollands GJ, *et al.* Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. *Research Synthesis Methods* 2014;**5**. doi:10.1002/jrsm.1093

8 Settles B. Active Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 2012;**6**. doi:10.2200/S00429ED1V01Y201207AIM018

9 Cormack G v., Grossman MR. Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. In: *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. New York, NY, USA: : ACM 2014. doi:10.1145/2600428.2609601

10 Ferdinands G, Schram R, Bruin J, *et al. Active learning for screening prioritization in systematic reviews - A simulation study*. 2020. doi:10.31219/osf.io/w6qbg

11 Ferdinands G. AI-Assisted Systematic Reviewing: Selecting Studies to Compare Bayesian Versus Frequentist SEM for Small Sample Sizes. *Multivariate Behavioral Research* 2021;**56**. doi:10.1080/00273171.2020.1853501

12 Gates A, Johnson C, Hartling L. Technology-assisted title and abstract screening for systematic reviews: a retrospective evaluation of the Abstrackr machine learning tool. *Systematic Reviews* 2018;**7**. doi:10.1186/s13643-018-0707-8

13 Gates A, Guitard S, Pillay J, *et al.* Performance and usability of machine learning for screening in systematic reviews: a comparative evaluation of three tools. *Systematic Reviews* 2019;**8**. doi:10.1186/s13643-019-1222-2

14 Miwa M, Thomas J, O'Mara-Eves A, *et al.* Reducing systematic review workload through certainty-based screening. *Journal of Biomedical Informatics* 2014;**51**. doi:10.1016/j.jbi.2014.06.005

15 O'Mara-Eves A, Thomas J, McNaught J, *et al.* Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic Reviews* 2015;**4**. doi:10.1186/2046-4053-4-5

16 van de Schoot R, de Bruin J, Schram R, *et al.* An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence* 2021;**3**. doi:10.1038/s42256-020-00287-7

17 Singh G, Thomas J, Shawe-Taylor J. Improving Active Learning in Systematic Reviews. *ArXiv* 2018;**abs/1801.09496**.

18 Wallace BC, Small K, Brodley CE, *et al.* Deploying an Interactive Machine Learning System in an Evidence-Based Practice Center: Abstrackr. In: *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. New York, NY, USA: : Association for Computing Machinery 2012. 819–24. doi:10.1145/2110363.2110464

19 Wallace BC, Trikalinos TA, Lau J, *et al.* Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics* 2010;**11**. doi:10.1186/1471-2105-11-55

20 Yu Z, Menzies T. FAST2: An intelligent assistant for finding relevant papers. *Expert Systems with Applications* 2019;**120**. doi:10.1016/j.eswa.2018.11.021

21 Sheng VS, Provost F, Ipeirotis PG. KDD '08: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: : Association for Computing Machinery 2018.

22 Ipeirotis PG, Provost F, Sheng VS, *et al.* Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery* 2014;**28**. doi:10.1007/s10618-013-0306-1

23 van de Schoot R, de Bruin J, Schram R, *et al.* ASReview: Active learning for systematic reviews. 2021.https://zenodo.org/record/4647608#.YJOb1LUzaUk (accessed 6 May 2021).

24 Wang Z, Nayfeh T, Tetzlaff J, *et al.* Error rates of human reviewers during abstract screening in systematic reviews. *PLOS ONE* 2020;**15**. doi:10.1371/journal.pone.0227742

25 Harmsen W, de Groot J, Harkema A, *et al.* Scripts for paper on Towards up-to-date medical guidelines. . 2021.

26 ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning. New York, NY, USA: : Association for Computing Machinery 2009.

27 Pérez J, Díaz J, Garcia-Martin J, *et al.* Systematic literature reviews in software engineering—enhancement of the study selection process using Cohen's Kappa statistic. *Journal of Systems and Software* 2020;**168**. doi:10.1016/j.jss.2020.110657

28 Zou J, Li D, Kanoulas E. Technology Assisted Reviews: Finding the Last Few Relevant Documents by Asking Yes/No Questions to Reviewers. *arXiv e-prints* 2018;:arXiv:1810.05414.

29 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? *Systematic Reviews* 2019;**8**. doi:10.1186/s13643-019-1062-0

30 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). *Systematic Reviews* 2018;**7**. doi:10.1186/s13643-017-0667-4

31 Altena AJ, Spijker R, Olabarriaga SD. Usage of automation tools in systematic reviews. *Research Synthesis Methods* 2019;**10**. doi:10.1002/jrsm.1335

32 Belur J, Tompson L, Thornton A, *et al.* Interrater Reliability in Systematic Review Methodology: Exploring Variation in Coder Decision-Making. *Sociological Methods & Research* 2021;**50**. doi:10.1177/0049124118799372

33 Ali N bin, Petersen K. Evaluating strategies for study selection in systematic literature studies. In: *Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement - ESEM '14*. New York, New York, USA: : ACM Press 2014. doi:10.1145/2652524.2652557