BMJ Open

A modern era systematic review using artificial intelligence: considerations to ensure methodological quality

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2023-072254.R1
Article Type:	Communication
Date Submitted by the Author:	22-Feb-2023
Complete List of Authors:	van Dijk, Sanne H B; University of Twente Technical Medical Centre, Health Technology & Services Research; Medisch Spectrum Twente, Pulmonary Medicine Brusse-Keizer, Marjolein; Medisch Spectrum Twente, Medical School Twente; University of Twente Technical Medical Centre, Health Technology & Services Research Bucsán, Charlotte; Medisch Spectrum Twente, Pulmonary Medicine; University of Twente Faculty of Behavioural Sciences, Cognition, Data & Education van der Palen, Job; Medisch Spectrum Twente, Medical School Twente; University of Twente Faculty of Behavioural Sciences, Cognition, Data & Education Doggen, Carine J.M.; University of Twente Technical Medical Centre, Health Technology & Services Research; Rijnstate Hospital, Clinical Research Centre Lenferink, Anke; University of Twente Technical Medical Centre, Health Technology & Services Research; Medisch Spectrum Twente, Pulmonary Medicine
Primary Subject Heading:	Communication
Secondary Subject Heading:	Research methods
Keywords:	Systematic Review, STATISTICS & RESEARCH METHODS, Information technology < BIOTECHNOLOGY & BIOINFORMATICS

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in BMJ Open and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

TITLE PAGE

Title:

A modern era systematic review using artificial intelligence: considerations to ensure methodological quality

Authors and affiliations:

Sanne H B van Dijk[1,2], Marjolein G J Brusse-Keizer[1,3], Charlotte C Bucsán[2,4], Job van der Palen[3,4], Carine J M Doggen[1,5], Anke Lenferink[1,2,5]

[1] Health Technology & Services Research, Technical Medical Centre, University of Twente, Enschede, the Netherlands

[2] Department of Pulmonary Medicine, Medisch Spectrum Twente, Enschede, the Netherlands

[3] Medical School Twente, Medisch Spectrum Twente, Enschede, the Netherlands

[4] Cognition, Data & Education, Faculty of Behavioural, Management & Social Sciences, University of Twente, Enschede, the Netherlands

[5] Clinical Research Centre, Rijnstate Hospital, Arnhem, the Netherlands

Corresponding author:

Anke Lenferink, a.lenferink@utwente.nl

ABSTRACT

Systematic reviews provide a structured overview of the available evidence in medical-scientific research. However, due to the increasing medical-scientific research output, it is a time-consuming task to conduct systematic reviews. To accelerate this process, artificial intelligence (AI) can be used in the screening of titles and abstracts. In this communication paper, we suggest how to conduct a transparent and reliable systematic review using the AI tool 'ASReview' in the title and abstract screening. Using the tool in our review resulted in much time saved: 23% of the articles were screened with the AI tool, leaving 77% of the articles unseen. Considerations to ensure methodological quality when using AI in systematic reviews included: the choice of when to use AI, the need of both deduplication and inter-reviewer agreement, how to choose a stopping criterion, and the quality of reporting. The AI tool is an important innovation for current systematic reviewing practice, as long as it is appropriately used and methodological quality can be assured.

STRENGTHS AND LIMITATIONS OF THIS STUDY

- Potential pitfalls regarding the use of artificial intelligence in systematic reviewing were identified.
- Remedies for each pitfall were provided to ensure methodological quality.
- A time-efficient approach is suggested on how to conduct a transparent and reliable systematic review using an artificial intelligence tool.
- The artificial intelligence tool described in the paper was not evaluated for its accuracy.

BACKGROUND

Medical-scientific research output has grown exponentially since the very first medical papers were published (1–3). The output in the field of clinical medicine increased and keeps doing so (4). To illustrate, a quick PubMed search for "cardiology" shows a fivefold increase in annual publications from 10,420 (2007) to 52,537 (2021). Although the medical-scientific output growth rate is not higher when compared to other scientific fields (1–3), this field creates the largest output (3). Staying updated by reading all published articles is therefore not feasible. However, systematic reviews facilitate up-to-date and accessible summaries of evidence, as they synthesise previously published results in a transparent and reproducible manner (5,6). Hence, conclusions can be drawn that provide the highest considered level of evidence in medical research (5,7). Therefore, systematic reviews are not only crucial in science, but they have a large impact on clinical practice and policy-making as well (6). They are, however, highly labour-intensive to conduct due to the necessity of screening a large amount of research. Thus, efficient and innovative reviewing methods are desired (8).

An open-source artificial intelligence (AI) tool 'ASReview' (9) was published in 2021 to facilitate the title and abstract screening process in systematic reviews. Applying this tool facilitates researchers to conduct systematic reviews: simulations already showed its time-saving potential (9–11). We used the tool in the study selection of our own systematic review and came across scenarios that needed consideration to prevent loss of methodological quality. In this communication paper, we provide a reliable and transparent AI-supported systematic reviewing approach.

METHODS

We first describe how the AI tool was used in a systematic review conducted by our research group. For more detailed information regarding searches and eligibility criteria of the review, we refer to the protocol (PROSPERO registry: CRD42022283952) (12). Subsequently, when deciding on the AI screening-related methodology, we applied appropriate remedies against foreseen scenarios and their pitfalls to maintain a reliable and transparent approach. These potential scenarios, pitfalls and remedies will be discussed in the result section.

In our systematic review, the AI tool 'ASReview' (version 0.17.1) (9) was used for the screening of titles and abstracts by the first reviewer (SvD). The tool uses an active researcher-in-the-loop machine learning algorithm to rank the articles from high to low probability of eligibility for inclusion by text mining. The AI tool offers several classifier models by which the relevancy of the included articles can be determined (9). In a simulation study using six large systematic review datasets on various topics, a Naïve Bayes (NB) and a term frequency-inverse document frequency (TF-IDF) outperformed other model settings (10). The NB classifier estimates the probability of an article being relevant, based on TF-IDF measurements. TF-IDF measures the originality of a certain word within the article relative to the total number of articles the word appears in (13). This combination of NB and TF-IDF were chosen for our systematic review.

Before the AI tool can be used for the screening of relevant articles, its algorithm needs training with at least one relevant and one irrelevant article (i.e., prior knowledge). It is assumed that the more prior knowledge, the better the algorithm is trained at the start of the screening process, and the faster it will identify relevant articles (9). In our review, the prior knowledge consisted of three relevant articles (14–16) selected from a systematic review on the topic (17) and three randomly picked irrelevant articles .

After training with the prior knowledge, the AI tool made a first ranking of all unlabelled articles (i.e., articles not yet decided on eligibility). Unlabelled articles were ranked from highest to lowest probability of being relevant by the classifier, based on the prior knowledge (i.e., articles already decided on). The first reviewer read the title and abstract of the highest ranked article and made a decision ('relevant' or 'irrelevant') following the eligibility criteria. This decision was added to the prior knowledge and the AI tool made a new ranking. Again, the next top ranked article was proposed to the reviewer, who made a decision regarding eligibility. The process of AI taking additional prior knowledge into account for each ranking and a reviewer making decisions was repeated until the predefined stopping criterion of – in our case - 100 subsequent irrelevant articles was reached.

The articles that were labelled relevant during the title and abstract screening were each screened on full text independently by two reviewers (SvD & MBK, AL, JvdP, CD, CB) to minimise the influence of subjectivity on inclusion. Disagreements regarding inclusion were solved by a third independent reviewer.

RESULTS

How to maintain reliability and transparency when using AI in systematic reviewing

A summary of the potential scenarios, and their pitfalls and remedies, when using the AI tool in a systematic review is given in Table 1. These potential scenarios should not be ignored, but acted upon to maintain reliability and transparency. Figure 1 shows when and where to act upon during the screening process, from literature search results to publishing the review.

Table 1 Per-scenario overview of potential pitfalls and how to prevent these when using ASReview in a systematic review

Potential scenario		Pitfall	Remedy
①	Only a small (i.e., manually feasible) number of articles (with possibly a high proportion relevant) available for screening	Time wasted by considering AI-related choices, and no time saved by using AI	Do not use AI: conduct manual screening
②	Presence of duplicate articles in ASReview	Unequal weighing of labelled articles in AI-supported screening	Apply deduplication methods before using AI
③	Reviewer's own opinion, expertise or mistakes influence(s) AI algorithm on article selection	Not all relevant articles are included, potentially introducing selection bias	Reviewer training in title and abstract screening Perform (partial) double screening and check inter-reviewer agreement
④	AI-supported screening is stopped before or a long time after all relevant articles are found	Not all relevant articles are included, potentially introducing selection bias, or time is wasted	Formulate a data-driven stopping criterion (i.e., number of consecutive irrelevant articles)
⑤	AI-related choices not (completely) described	Irreproducible results, leading to a low-quality systematic review	Describe and substantiate the choices that are made
⑥	Study selection is not transparent	Irreproducible results (black box algorithm), leading to a low-quality systematic review	Publish open data (i.e., extracted file with all decisions)

In our systematic review, by means of broad literature searches in several scientific databases, a first set of potentially relevant articles was identified. This yielded 8,456 articles, enough to apply the AI tool in the title and abstract screening (scenario ① was avoided, see

Table 1). Subsequently, this complete set of articles was uploaded in reference manager EndNote X9 (18) and Covidence (19), where 3,761 duplicate articles were removed. Deduplication is usually applied in systematic reviewing (20), but is increasingly important prior to the use of AI. Since multiple decisions regarding a duplicate article weigh more than one, this will disproportionately influence classification and possibly the results (Table 1, scenario ②). In our review, a deduplicated set of articles was uploaded in the AI tool. Prior to the actual AI-supported title and abstract screening, the reviewers (SvD & AL, MBK) trained themselves with a small selection of 74 articles. The first reviewer became familiar with the ASReview software, and all three reviewers learnt how to apply the eligibility criteria, to minimise personal influence on the article selection (Table 1, scenario ③).

Defining the stopping criterion used in the screening process is left to the reviewer (9). An optimal stopping criterion in active learning is considered a perfectly balanced trade-off between a certain cost (in terms of time spent) of screening one more article versus the predictive performance (in terms of identifying a new relevant article) that could be increased by adding one more decision (21). The optimal stopping criterion in systematic reviewing would be the moment that screening additional articles will not result in more relevant articles being identified (22). Therefore, in our review, we predetermined a data-driven stopping criterion for the title and abstract screening as '100 consecutive irrelevant articles' in order to prevent the screening from being stopped before or a long time after all relevant articles were identified (Table 1, scenario ④).

Due to the fact that the stopping criterion was reached after 1,063 of the 4,695 articles, only a part of the total number of articles was seen. Therefore, this approach might be sensitive to possible mistakes when articles are screened by only one reviewer, influencing the algorithm, possibly resulting in an incomplete selection of articles (Table 1, scenario ③) (23). As a remedy, second reviewers (AL, MBK) checked 20% of the titles and abstracts seen by the

first reviewer. This 20% had a comparable ratio regarding relevant versus irrelevant articles over all articles seen. The percentual agreement and Cohen's Kappa (κ), a measure for the inter-reviewer agreement above chance, were calculated to express the reliability of the decisions taken (24). The decisions were agreed in 96% and κ was 0.83. A κ equal of at least 0.6 is generally considered high (24), and thus it was assumed that the algorithm was reliably trained by the first reviewer.

The reporting of the use of the AI tool should be transparent. If the choices made regarding the use of the AI tool are not entirely reported (Table 1, scenario ⑤), the reader will not be able to properly assess the methodology of the review, and review results may even be graded as low-quality due to the lack of transparent reporting. The ASReview tool offers the possibility to extract a data file providing insight into all decisions made during the screening process, in contrast to various other "black box" AI-reviewing tools (9). This file will be published alongside our systematic review to provide full transparency of our AI-supported screening. This way, the screening with AI is reproducible (remedy to scenario ⑥, Table 1).

Results of AI-supported study selection in a systematic review

We experienced an efficient process of title and abstract screening in our systematic review. Whereas the screening was performed with a database of 4,695 articles, the stopping criterion was reached after 1,063 articles, so 23% were seen. Figure 2 shows the proportion of articles identified as being relevant at any point during the AI-supported screening process. It can be observed that the articles are indeed prioritised by the active learning algorithm: in the beginning, relatively many relevant articles were found, but this decreased as the stopping criterion (vertical red line, Figure 2) was approached. During the screening, 142 articles were labelled relevant. After the inter-reviewer agreement check, 142 articles proceeded to the full text reviewing phase, of which 65 were excluded because these were no articles with an original

research format, and three because the full text could not be retrieved. After full text reviewing of the remaining 74 articles, 18 articles from 13 individual studies were included in our review. After snowballing, one additional article from a study already included was added.

DISCUSSION

In our systematic review, the AI tool considerably reduced the number of articles in the screening process. Since the AI tool is offered open source, many researchers may benefit from its time-saving potential in selecting articles. Choices in several scenarios regarding the use of AI, however, are still left open to the researcher, and need consideration to prevent pitfalls. These include the choice whether or not to use AI, the importance of deduplication, double screening to check inter-reviewer agreement, a data-driven stopping criterion to optimally utilise the algorithm's predictive performance, and quality of reporting of the AI-related methodology chosen. This communication paper is, to our knowledge, the first elaborately explaining and discussing these choices regarding the application of this AI tool in a systematic review.

The main advantage of using the AI tool is the amount of time saved. Indeed, in our study, only 23% of the total number of articles were screened before the predefined stopping criterion was met. Assuming that all relevant articles were found, the AI tool saved 77% of the time for title and abstract screening. An additional advantage is that research questions previously unanswerable due to the insurmountable number of articles to screen in a 'classic' (i.e., manual) review, now actually are possible to answer. An example of the latter is a review screening over 60,000 articles (25), which would probably never have been performed without AI supporting the article selection.

Since the introduction of the ASReview tool in 2021 it was applied in seven published reviews (25–31). An important note to make is that only one (25) clearly reported AI-related choices in the methods and a complete and transparent flowchart reflecting the study selection process in the results section. Two reviews reported a relatively small number (< 400) of articles to screen (26,27), of which more than 75% of the articles were screened before the stopping criterion was met, so the amount of time saved was limited. Also, three reviews

reported many initial articles (> 6,000) (25,28,29) and one reported 892 articles (31), of which only 5 to 10% needed to be screened. So in these reviews, the AI tool saved an impressive amount of screening time. In our systematic review, 3% of the articles were labelled relevant during the title and abstract screening and eventually, less than 1% of all initial articles were included. These percentages are low, and are in line with the three above-mentioned reviews (1-2% and 0-1%, respectively) (25,28,29). Still, relevancy and inclusion rates are much lower when compared with 'classic' systematic reviews. A study evaluating the screening process in 25 'classic' systematic reviews showed that approximately 18% was labelled relevant and 5% was actually included in the reviews (32). This difference is probably due to more narrow literature searches in 'classic' reviews for feasibility purposes compared with AI-supported reviews, resulting in a higher proportion of included articles.

In this paper we show how we applied the AI tool, but we did not evaluate it in terms of accuracy. This means that we have to deal with a certain degree of uncertainty. Despite the data-driven stopping criterion there is a chance that relevant articles were missed, as 77% was automatically excluded. Considering this might have been the case, firstly, this could be due to wrong decisions of the reviewer that would have undesirably influenced the training of the algorithm by which the articles were labelled as (ir)relevant and the order in which they were presented to the reviewer. Relevant articles could have therefore remained unseen if the stopping criterion was reached before they were presented to the reviewer. As a remedy, of the 20% of the articles that screened by the first reviewer, relevancy was also assessed by another reviewer to assess inter-reviewer reliability, which was high. It should be noted, though, that 'classic' title and abstract screening is not necessarily better than using AI, as medical-scientific researchers tend to assess one out of nine abstracts wrongly (32). Secondly, the AI tool may not have properly ranked highly relevant to irrelevant articles. However, given that simulations proved this AI tool's accuracy before (9–11) this was not considered plausible. Since our study

applied, but did not evaluate, the AI tool, we encourage future studies evaluating the performance of the tool. This could not only enrich the knowledge about the AI tool, but also increase certainty about using it.

Although various researcher-in-the-loop AI tools for title and abstract screening have been developed over the years (9,23,33), they often do not develop into usable mature software (33), which impedes AI to be permanently implemented in research practice. For medical-scientific research practice, it would therefore be helpful if large systematic review institutions, like Cochrane and PRISMA, would consider to 'officially' make AI part of systematic reviewing practice. When guidelines on the use of AI in systematic reviews are made available and widely recognised, AI-supported systematic reviews can be uniformly and transparently reported. Only then we can really benefit from AI's time-saving potential and reduce our research time waste.

CONCLUSION

Our experience with the AI tool during the title and abstract screening was positive as it has highly accelerated the literature selection process. However, users should consider applying appropriate remedies to scenarios that may form a threat to the methodological quality of the review. We provided an overview of these scenarios, their pitfalls and remedies. These encourage reliable use and transparent reporting of AI in systematic reviewing. To ensure the continuation of conducting systematic reviews in the future, and given their importance for medical guidelines and practice, we consider this tool as an important addition in the review process.

DECLARATIONS

Contributors

SvD proposed the methodology and conducted the study selection. MBK, CD and AL critically reflected on the methodology. MBK and AL contributed substantially to the study selection. CB, JvdP and CD contributed to the study selection. The manuscript was primarily prepared by SvD and critically revised by all authors. All authors read and approved the final manuscript.

Funding

The systematic review is conducted as part of the RE-SAMPLE project. RE-SAMPLE has received funding from the European Union's Horizon 2020 research and innovation programme (grant agreement no. 965315).

Competing interests

None declared.

Protocol registration

PROSPERO CRD42022283952

REFERENCES

1. Bornmann L, Mutz R. Growth Rates of Modern Science: A Bibliometric Analysis Based on the Number of Publications and Cited References. J Am Soc Inf Sci Technol. 2015;66(11):2215–22.
2. Bornmann L, Haunschild R, Mutz R. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. Humanit Soc Sci Commun. 2021;8(1).
3. Michels C, Schmoch U. The growth of science and database coverage. Scientometrics. 2012;93(3):831–46.
4. Haghani M, Abbasi A, Zwack CC, Shahhoseini Z, Haslam N. Trends of research productivity across author gender and research fields: A multidisciplinary and multi-country observational study [Internet]. Vol. 17, PLoS ONE. 2022. Available from: http://dx.doi.org/10.1371/journal.pone.0271998
5. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. J Clin Epidemiol [Internet]. 2021;134:178–89. Available from: https://doi.org/10.1016/j.jclinepi.2021.03.001
6. Gurevitch J, Koricheva J, Nakagawa S, Stewart G. Meta-analysis and the science of research synthesis. Nature. 2018;555(7695):175–82.
7. Burns P, Rohrich R, Chung K. The Levels of Evidence and their role in Evidence-Based Medicine. Plast Reconstr Surg [Internet]. 2011;128(1):305–10. Available from: http://www.alpbc.eu/
8. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: How will we ever keep up? PLoS Med. 2010;7(9).
9. van de Schoot R, de Bruin J, Schram R, Zahedi P, de Boer J, Weijdema F, et al. An

open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell [Internet]. 2021;3(February):125–33. Available from: http://dx.doi.org/10.1038/s42256-020-00287-7

10. Ferdinands G, Schram R, de Bruin J, Bagheri A, Oberski D, Tummers L, et al. Active learning for screening prioritization in systematic reviews: A simulation study. 2020.
11. Ferdinands G. AI-Assisted Systematic Reviewing: Selecting Studies to Compare Bayesian Versus Frequentist SEM for Small Sample Sizes. Multivariate Behav Res [Internet]. 2020;56(1):153–4. Available from: https://doi.org/10.1080/00273171.2020.1853501
12. van Dijk SHB, Brusse-Keizer MGJ, Bucsán CC, Doggen CJM, van der Palen J, Lenferink A. Distinguishing acute heart failure from exacerbations of COPD: An AI-supported systematic literature review. PROSPERO [Internet]. 2022;CRD42022283952. Available from: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42022283952
13. Havrlant L, Kreinovich V. A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation). Int J Gen Syst [Internet]. 2017;46(1):27–36. Available from: http://dx.doi.org/10.1080/03081079.2017.1291635
14. Wang R, Cao Z, Li Y, Yu K. Utility of N-terminal pro B-type natriuretic peptide and mean platelet volume in differentiating congestive heart failure from chronic obstructive pulmonary disease. Int J Cardiol [Internet]. 2013;170(2):e28–9. Available from: http://dx.doi.org/10.1016/j.ijcard.2013.10.048
15. Ouanes I, Jalloul F, Ayed S, Dachraoui F, Ouanes-Besbes L, Fekih Hassen M, et al. N-terminal proB-type natriuretic peptide levels aid the diagnosis of left ventricular dysfunction in patients with severe acute exacerbations of chronic obstructive

pulmonary disease and renal dysfunction. Respirology. 2012;17(4):660–6.

16. Andrijevic I, Milutinov S, Lozanov Crvenkovic Z, Matijasevic J, Andrijevic A, Kovacevic T, et al. N-Terminal Prohormone of Brain Natriuretic Peptide (NT-proBNP) as a Diagnostic Biomarker of Left Ventricular Systolic Dysfunction in Patients with Acute Exacerbation of Chronic Obstructive Pulmonary Disease (AECOPD). Lung [Internet]. 2018;196(5):583–90. Available from: http://dx.doi.org/10.1007/s00408-018-0137-3

17. Hawkins NM, Khosla A, Virani SA, McMurray JJV, FitzGerald JM. B-type natriuretic peptides in chronic obstructive pulmonary disease: A systematic review. BMC Pulm Med [Internet]. 2017;17(1). Available from: http://dx.doi.org/10.1186/s12890-016-0345-7

18. Clarivate Analytics. EndNote X9. 2018.

19. Covidence systematic review software [Internet]. Melbourne, Australia: Veritas Health Innovation; Available from: www.covidence.org

20. McKeown S, Mir ZM. Considerations for conducting systematic reviews: evaluating the performance of different methods for de-duplicating references. Syst Rev. 2021;10(1):4–11.

21. Ishibashi H, Hino H. Stopping Criterion for Active Learning Based on Error Stability. 2021;1:1–32. Available from: http://arxiv.org/abs/2104.01836

22. Wang W, Cai W, Zhang Y. Stability-Based Stopping Criterion for Active Learning. Proc - IEEE Int Conf Data Mining, ICDM. 2014;2015-Janua(January):1019–24.

23. Blaizot A, Veettil SK, Saidoung P, Moreno-Garcia CF, Wiratunga N, Aceves-Martins M, et al. Using artificial intelligence methods for systematic review in health sciences: A systematic review. Res Synth Methods. 2022;13(3):353–62.

24. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med [Internet].

2012;22(3):276–82. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/

25. Bernardes RC, Botina LL, Araújo R dos S, Guedes RNC, Martins GF, Lima MAP. Artificial Intelligence-Aided Meta-Analysis of Toxicological Assessment of Agrochemicals in Bees. Front Ecol Evol. 2022;10(May).

26. Silva GFS, Fagundes TP, Teixeira BC, Chiavegatto Filho ADP. Machine Learning for Hypertension Prediction: a Systematic Review. Curr Hypertens Rep [Internet]. 2022;(3). Available from: https://doi.org/10.1007/s11906-022-01212-6

27. Miranda L, Paul R, Pütz B, Koutsouleris N, Müller-Myhsok B. Systematic Review of Functional MRI Applications for Psychiatric Disease Subtyping. Front Psychiatry. 2021;12(October).

28. Schouw HM, Huisman LA, Janssen YF, Slart RHJA, Borra RJH, Willemsen ATM, et al. Targeted optical fluorescence imaging: a meta-narrative review and future perspectives. Eur J Nucl Med Mol Imaging [Internet]. 2021;48(13):4272–92. Available from: https://doi.org/10.1007/s00259-021-05504-y

29. Bakkum L, Schuengel C, Sterkenburg PS, Frielink N, Embregts PJCM, de Schipper JC, et al. People with intellectual disabilities living in care facilities engaging in virtual social contact: A systematic review of the feasibility and effects on well-being. J Appl Res Intellect Disabil. 2022;35(1):60–74.

30. Huang Y, Procházková M, Lu J, Riad A, Macek P. Family Related Variables' Influences on Adolescents' Health Based on Health Behaviour in School-Aged Children Database, an AI-Assisted Scoping Review, and Narrative Synthesis. Front Psychol. 2022;13(August).

31. Zhang W, Huang S, Lam L, Evans R, Zhu C. Cyberbullying definitions and measurements in children and adolescents: Summarizing 20 years of global efforts.

Front Public Heal. 2022;10.

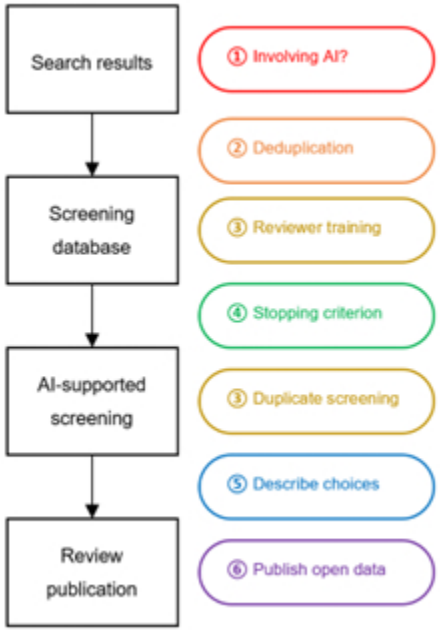
32. Wang Z, Nayfeh T, Tetzlaff J, O'Blenis P, Murad MH. Error rates of human reviewers during abstract screening in systematic reviews. PLoS One [Internet]. 2020;15(1):1–8. Available from: http://dx.doi.org/10.1371/journal.pone.0227742

33. Marshall IJ, Wallace BC. Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. Syst Rev. 2019;8(1):1–10.

FIGURE CAPTIONS

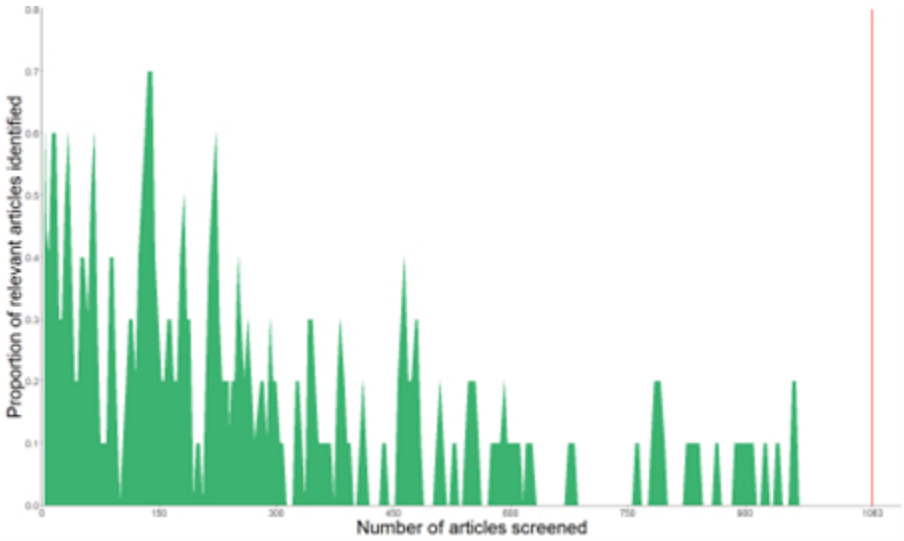
Figure 1 Flowchart showing when and where to act upon when using ASReview in systematic reviewing

Figure 2 Proportion of relevant articles identified after a certain number of titles and abstracts were screened using the AI tool



Flowchart showing when and where to act upon when using ASReview in systematic reviewing

147x211mm (38 x 38 DPI)



Proportion of relevant articles identified after a certain number of titles and abstracts were screened using the AI tool

409x242mm (28 x 28 DPI)

BMJ Open

Artificial intelligence in systematic reviews: promising when appropriately used

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2023-072254.R2
Article Type:	Communication
Date Submitted by the Author:	16-Jun-2023
Complete List of Authors:	van Dijk, Sanne H B; University of Twente Technical Medical Centre, Health Technology & Services Research; Medisch Spectrum Twente, Pulmonary Medicine Brusse-Keizer, Marjolein; Medisch Spectrum Twente, Medical School Twente; University of Twente Technical Medical Centre, Health Technology & Services Research Bucsán, Charlotte; Medisch Spectrum Twente, Pulmonary Medicine; University of Twente Faculty of Behavioural Sciences, Cognition, Data & Education van der Palen, Job; Medisch Spectrum Twente, Medical School Twente; University of Twente Faculty of Behavioural Sciences, Cognition, Data & Education Doggen, Carine J.M.; University of Twente Technical Medical Centre, Health Technology & Services Research; Rijnstate Hospital, Clinical Research Centre Lenferink, Anke; University of Twente Technical Medical Centre, Health Technology & Services Research; Medisch Spectrum Twente, Pulmonary Medicine
Primary Subject Heading:	Communication
Secondary Subject Heading:	Research methods
Keywords:	Systematic Review, STATISTICS & RESEARCH METHODS, Information technology < BIOTECHNOLOGY & BIOINFORMATICS

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in BMJ Open and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

TITLE PAGE

Title:

Artificial intelligence in systematic reviews: promising when appropriately used

Authors and affiliations:

Sanne H B van Dijk[1,2], Marjolein G J Brusse-Keizer[1,3], Charlotte C Bucsán[2,4], Job van der Palen[3,4], Carine J M Doggen[1,5], Anke Lenferink[1,2,5]

[1] Health Technology & Services Research, Technical Medical Centre, University of Twente, Enschede, the Netherlands

[2] Department of Pulmonary Medicine, Medisch Spectrum Twente, Enschede, the Netherlands

[3] Medical School Twente, Medisch Spectrum Twente, Enschede, the Netherlands

[4] Cognition, Data & Education, Faculty of Behavioural, Management & Social Sciences, University of Twente, Enschede, the Netherlands

[5] Clinical Research Centre, Rijnstate Hospital, Arnhem, the Netherlands

Corresponding author:

Anke Lenferink, a.lenferink@utwente.nl

ABSTRACT

Background: Systematic reviews provide a structured overview of the available evidence in medical-scientific research. However, due to the increasing medical-scientific research output, it is a time-consuming task to conduct systematic reviews. To accelerate this process, artificial intelligence (AI) can be used in the review process. In this communication paper, we suggest how to conduct a transparent and reliable systematic review using the AI tool 'ASReview' in the title and abstract screening.

Methods: Use of the AI tool consisted of several steps. First, the tool required training of its algorithm with several prelabelled articles prior to screening. Next, using a researcher-in-the-loop algorithm, the AI tool proposed the article with the highest probability of being relevant. The reviewer then decided on relevancy of each article proposed. This process was continued until the stopping criterion was reached. All articles labelled relevant by the reviewer were screened on full text.

Results: Considerations to ensure methodological quality when using AI in systematic reviews included: the choice of whether to use AI, the need of both deduplication and inter-reviewer agreement, how to choose a stopping criterion, and the quality of reporting. Using the tool in our review resulted in much time saved: only 23% of the articles were proposed by the AI tool.

Conclusion: The AI tool is a promising innovation for the current systematic reviewing practice, as long as it is appropriately used and methodological quality can be assured.

STRENGTHS AND LIMITATIONS OF THIS STUDY

- Potential pitfalls regarding the use of artificial intelligence in systematic reviewing were identified.

- Remedies for each pitfall were provided to ensure methodological quality. - A time-efficient approach is suggested on how to conduct a transparent and reliable systematic review using an artificial intelligence tool.

- The artificial intelligence tool described in the paper was not evaluated for its accuracy.

BACKGROUND

Medical-scientific research output has grown exponentially since the very first medical papers were published (1–3). The output in the field of clinical medicine increased and keeps doing so (4). To illustrate, a quick PubMed search for "cardiology" shows a fivefold increase in annual publications from 10,420 (2007) to 52,537 (2021). Although the medical-scientific output growth rate is not higher when compared to other scientific fields (1–3), this field creates the largest output (3). Staying updated by reading all published articles is therefore not feasible. However, systematic reviews facilitate up-to-date and accessible summaries of evidence, as they synthesise previously published results in a transparent and reproducible manner (5,6). Hence, conclusions can be drawn that provide the highest considered level of evidence in medical research (5,7). Therefore, systematic reviews are not only crucial in science, but they have a large impact on clinical practice and policy-making as well (6). They are, however, highly labour-intensive to conduct due to the necessity of screening a large amount of articles, which results in a high consumption of research resources. Thus, efficient and innovative reviewing methods are desired (8).

An open-source artificial intelligence (AI) tool 'ASReview' (9) was published in 2021 to facilitate the title and abstract screening process in systematic reviews. Applying this tool facilitates researchers to conduct systematic reviews: simulations already showed its time-saving potential (9–11). We used the tool in the study selection of our own systematic review and came across scenarios that needed consideration to prevent loss of methodological quality. In this communication paper, we provide a reliable and transparent AI-supported systematic reviewing approach.

METHODS

We first describe how the AI tool was used in a systematic review conducted by our research group. For more detailed information regarding searches and eligibility criteria of the review, we refer to the protocol (PROSPERO registry: CRD42022283952) (12). Subsequently, when deciding on the AI screening-related methodology, we applied appropriate remedies against foreseen scenarios and their pitfalls to maintain a reliable and transparent approach. These potential scenarios, pitfalls and remedies will be discussed in the result section.

In our systematic review, the AI tool 'ASReview' (version 0.17.1) (9) was used for the screening of titles and abstracts by the first reviewer (SvD). The tool uses an active researcher-in-the-loop machine learning algorithm to rank the articles from high to low probability of eligibility for inclusion by text mining. The AI tool offers several classifier models by which the relevancy of the included articles can be determined (9). In a simulation study using six large systematic review datasets on various topics, a Naïve Bayes (NB) and a term frequency-inverse document frequency (TF-IDF) outperformed other model settings (10). The NB classifier estimates the probability of an article being relevant, based on TF-IDF measurements. TF-IDF measures the originality of a certain word within the article relative to the total number of articles the word appears in (13). This combination of NB and TF-IDF were chosen for our systematic review.

Before the AI tool can be used for the screening of relevant articles, its algorithm needs training with at least one relevant and one irrelevant article (i.e., prior knowledge). It is assumed that the more prior knowledge, the better the algorithm is trained at the start of the screening process, and the faster it will identify relevant articles (9). In our review, the prior knowledge consisted of three relevant articles (14–16) selected from a systematic review on the topic (17) and three randomly picked irrelevant articles .

After training with the prior knowledge, the AI tool made a first ranking of all unlabelled articles (i.e., articles not yet decided on eligibility) from highest to lowest probability of being relevant. The first reviewer read the title and abstract of the number one ranked article and made a decision ('relevant' or 'irrelevant') following the eligibility criteria. Next, the AI tool took into account this additional knowledge and made a new ranking. Again, the next top ranked article was proposed to the reviewer, who made a decision regarding eligibility. This process of AI making rankings and the reviewer making decisions, which is also called 'researcher-in-the-loop', was repeated until the predefined data-driven stopping criterion of – in our case - 100 subsequent irrelevant articles was reached. After the reviewer rejected what the AI tool puts forward as 'most probably relevant' a hundred times, it was assumed that there were no relevant articles left in the unseen part of the dataset.

The articles that were labelled relevant during the title and abstract screening were each screened on full text independently by two reviewers (SvD & MBK, AL, JvdP, CD, CB) to minimise the influence of subjectivity on inclusion. Disagreements regarding inclusion were solved by a third independent reviewer.

RESULTS

How to maintain reliability and transparency when using AI in title and abstract screening

A summary of the potential scenarios, and their pitfalls and remedies, when using the AI tool in a systematic review is given in Table 1. These potential scenarios should not be ignored, but acted upon to maintain reliability and transparency. Figure 1 shows when and where to act upon during the screening process reflected by the PRISMA flowchart (18), from literature search results to publishing the review.

In our systematic review, by means of broad literature searches in several scientific databases, a first set of potentially relevant articles was identified, yielding 8,456 articles, enough to expect the AI tool to be efficient in the title and abstract screening (scenario ① was avoided, see Table 1). Subsequently, this complete set of articles was uploaded in reference manager EndNote X9 (19) and review manager Covidence (20), where 3,761 duplicate articles were removed. Given that EndNote has quite low sensitivity in identifying duplicates, additional deduplication in Covidence was considered beneficial (21). Deduplication is usually applied in systematic reviewing (21), but is increasingly important prior to the use of AI. Since multiple decisions regarding a duplicate article weigh more than one, this will disproportionately influence classification and possibly the results (Table 1, scenario ②). In our review, a deduplicated set of articles was uploaded in the AI tool. Prior to the actual AI-supported title and abstract screening, the reviewers (SvD & AL, MBK) trained themselves with a small selection of 74 articles. The first reviewer became familiar with the ASReview software, and all three reviewers learnt how to apply the eligibility criteria, to minimise personal influence on the article selection (Table 1, scenario ③).

Defining the stopping criterion used in the screening process is left to the reviewer (9). An optimal stopping criterion in active learning is considered a perfectly balanced trade-off between a certain cost (in terms of time spent) of screening one more article versus the

predictive performance (in terms of identifying a new relevant article) that could be increased by adding one more decision (22). The optimal stopping criterion in systematic reviewing would be the moment that screening additional articles will not result in more relevant articles being identified (23). Therefore, in our review, we predetermined a data-driven stopping criterion for the title and abstract screening as '100 consecutive irrelevant articles' in order to prevent the screening from being stopped before or a long time after all relevant articles were identified (Table 1, scenario ④).

Due to the fact that the stopping criterion was reached after 1,063 of the 4,695 articles, only a part of the total number of articles was seen. Therefore, this approach might be sensitive to possible mistakes when articles are screened by only one reviewer, influencing the algorithm, possibly resulting in an incomplete selection of articles (Table 1, scenario ③) (24). As a remedy, second reviewers (AL, MBK) checked 20% of the titles and abstracts seen by the first reviewer. This 20% had a comparable ratio regarding relevant versus irrelevant articles over all articles seen. The percentual agreement and Cohen's Kappa (κ), a measure for the inter-reviewer agreement above chance, were calculated to express the reliability of the decisions taken (25). The decisions were agreed in 96% and κ was 0.83. A κ equal of at least 0.6 is generally considered high (25), and thus it was assumed that the algorithm was reliably trained by the first reviewer.

The reporting of the use of the AI tool should be transparent. If the choices made regarding the use of the AI tool are not entirely reported (Table 1, scenario ⑤), the reader will not be able to properly assess the methodology of the review, and review results may even be graded as low-quality due to the lack of transparent reporting. The ASReview tool offers the possibility to extract a data file providing insight into all decisions made during the screening process, in contrast to various other "black box" AI-reviewing tools (9). This file will be

published alongside our systematic review to provide full transparency of our AI-supported screening. This way, the screening with AI is reproducible (remedy to scenario ⑥, Table 1).

Table 1 Per-scenario overview of potential pitfalls and how to prevent these when using ASReview in a systematic review

Potential scenario		Pitfall	Remedy
①	Only a small (i.e., manually feasible*) number of articles (with possibly a high proportion relevant) available for screening	Time wasted by considering AI-related choices, software training, and no time saved by using AI	Do not use AI: conduct manual screening
②	Presence of duplicate articles in ASReview	Unequal weighing of labelled articles in AI-supported screening	Apply deduplication methods before using AI
③	Reviewer's own opinion, expertise or mistakes influence(s) AI algorithm on article selection	Not all relevant articles are included, potentially introducing selection bias	Reviewer training in title and abstract screening Perform (partial) double screening and check inter-reviewer agreement
④	AI-supported screening is stopped before or a long time after all relevant articles are found	Not all relevant articles are included, potentially introducing selection bias, or time is wasted	Formulate a data-driven stopping criterion (i.e., number of consecutive irrelevant articles)
⑤	AI-related choices not (completely) described	Irreproducible results, leading to a low-quality systematic review	Describe and substantiate the choices that are made
⑥	Study selection is not transparent	Irreproducible results (black box algorithm), leading to a low-quality systematic review	Publish open data (i.e., extracted file with all decisions)

* What is considered manually feasible is highly context-dependent (i.e., the intended workload and/or number reviewers available)

Results of AI-supported study selection in a systematic review

We experienced an efficient process of title and abstract screening in our systematic review. Whereas the screening was performed with a database of 4,695 articles, the stopping criterion was reached after 1,063 articles, so 23% were seen. Figure 2A shows the proportion of articles identified as being relevant at any point during the AI-supported screening process. It can be

observed that the articles are indeed prioritised by the active learning algorithm: in the beginning, relatively many relevant articles were found, but this decreased as the stopping criterion (vertical red line) was approached. Figure 2B compares the screening progress when using the AI tool versus manual screening. The moment the stopping criterion was reached, approximately 32 records would have been found when the titles and abstract would have been screened manually, compared to 142 articles labelled relevant using the AI tool. After the inter-reviewer agreement check, 142 articles proceeded to the full text reviewing phase, of which 65 were excluded because these were no articles with an original research format, and three because the full text could not be retrieved. After full text reviewing of the remaining 74 articles, 18 articles from 13 individual studies were included in our review. After snowballing, one additional article from a study already included was added.

DISCUSSION

In our systematic review, the AI tool considerably reduced the number of articles in the screening process. Since the AI tool is offered open source, many researchers may benefit from its time-saving potential in selecting articles. Choices in several scenarios regarding the use of AI, however, are still left open to the researcher, and need consideration to prevent pitfalls. These include the choice whether or not to use AI by weighing the costs versus the benefits, the importance of deduplication, double screening to check inter-reviewer agreement, a data-driven stopping criterion to optimally utilise the algorithm's predictive performance, and quality of reporting of the AI-related methodology chosen. This communication paper is, to our knowledge, the first elaborately explaining and discussing these choices regarding the application of this AI tool in an example systematic review.

The main advantage of using the AI tool is the amount of time saved. Indeed, in our study, only 23% of the total number of articles were screened before the predefined stopping criterion was met. Assuming that all relevant articles were found, the AI tool saved 77% of the time for title and abstract screening. However, time should be invested to become acquainted with the tool. Whether the expected screening time saved outweighs this time investment is context-dependent (e.g., researcher's digital skills, systematic reviewing skills, topic knowledge). An additional advantage is that research questions previously unanswerable due to the insurmountable number of articles to screen in a 'classic' (i.e., manual) review, now actually are possible to answer. An example of the latter is a review screening over 60,000 articles (26), which would probably never have been performed without AI supporting the article selection.

Since the introduction of the ASReview tool in 2021, it was applied in seven published reviews (26–32). An important note to make is that only one (26) clearly reported AI-related choices in the methods and a complete and transparent flowchart reflecting the study selection

process in the results section. Two reviews reported a relatively small number (< 400) of articles to screen (27,28), of which more than 75% of the articles were screened before the stopping criterion was met, so the amount of time saved was limited. Also, three reviews reported many initial articles (> 6,000) (26,29,30) and one reported 892 articles (32), of which only 5 to 10% needed to be screened. So in these reviews, the AI tool saved an impressive amount of screening time. In our systematic review, 3% of the articles were labelled relevant during the title and abstract screening and eventually, less than 1% of all initial articles were included. These percentages are low, and are in line with the three above-mentioned reviews (1-2% and 0-1%, respectively) (26,29,30). Still, relevancy and inclusion rates are much lower when compared with 'classic' systematic reviews. A study evaluating the screening process in 25 'classic' systematic reviews showed that approximately 18% was labelled relevant and 5% was actually included in the reviews (33). This difference is probably due to more narrow literature searches in 'classic' reviews for feasibility purposes compared with AI-supported reviews, resulting in a higher proportion of included articles.

In this paper we show how we applied the AI tool, but we did not evaluate it in terms of accuracy. This means that we have to deal with a certain degree of uncertainty. Despite the data-driven stopping criterion there is a chance that relevant articles were missed, as 77% was automatically excluded. Considering this might have been the case, firstly, this could be due to wrong decisions of the reviewer that would have undesirably influenced the training of the algorithm by which the articles were labelled as (ir)relevant and the order in which they were presented to the reviewer. Relevant articles could have therefore remained unseen if the stopping criterion was reached before they were presented to the reviewer. As a remedy, in our own systematic review, of the 20% of the articles screened by the first reviewer, relevancy was also assessed by another reviewer to assess inter-reviewer reliability, which was high. It should be noted, though, that 'classic' title and abstract screening is not necessarily better than using

AI, as medical-scientific researchers tend to assess one out of nine abstracts wrongly (33). Secondly, the AI tool may not have properly ranked highly relevant to irrelevant articles. However, given that simulations proved this AI tool's accuracy before (9–11) this was not considered plausible. Since our study applied, but did not evaluate, the AI tool, we encourage future studies evaluating the performance of the tool across different scientific disciplines and contexts, since research suggests that the tool's performance depends on the context, for example, the complexity of the research question (34). This could not only enrich the knowledge about the AI tool, but also increase certainty about using it. Also, future studies should investigate the effects of choices made regarding the amount of prior knowledge that is provided to the tool, the number of articles defining the stopping criterion, and how duplicate screening is best performed, to guide future users of the tool.

Although various researcher-in-the-loop AI tools for title and abstract screening have been developed over the years (9,24,35), they often do not develop into usable mature software (35), which impedes AI to be permanently implemented in research practice. For medical-scientific research practice, it would therefore be helpful if large systematic review institutions, like Cochrane and PRISMA, would consider to 'officially' make AI part of systematic reviewing practice. When guidelines on the use of AI in systematic reviews are made available and widely recognised, AI-supported systematic reviews can be uniformly conducted and transparently reported. Only then we can really benefit from AI's time-saving potential and reduce our research time waste.

CONCLUSION

Our experience with the AI tool during the title and abstract screening was positive as it has highly accelerated the literature selection process. However, users should consider applying appropriate remedies to scenarios that may form a threat to the methodological quality of the

review. We provided an overview of these scenarios, their pitfalls and remedies. These encourage reliable use and transparent reporting of AI in systematic reviewing. To ensure the continuation of conducting systematic reviews in the future, and given their importance for medical guidelines and practice, we consider this tool as an important addition in the review process.

DECLARATIONS

Contributors

SvD proposed the methodology and conducted the study selection. MBK, CD and AL critically reflected on the methodology. MBK and AL contributed substantially to the study selection. CB, JvdP and CD contributed to the study selection. The manuscript was primarily prepared by SvD and critically revised by all authors. All authors read and approved the final manuscript.

Funding

The systematic review is conducted as part of the RE-SAMPLE project. RE-SAMPLE has received funding from the European Union's Horizon 2020 research and innovation programme (grant agreement no. 965315).

Competing interests

None declared.

Protocol registration

PROSPERO CRD42022283952

REFERENCES

1. Bornmann L, Mutz R. Growth Rates of Modern Science: A Bibliometric Analysis Based on the Number of Publications and Cited References. J Am Soc Inf Sci Technol. 2015;66(11):2215–22.
2. Bornmann L, Haunschild R, Mutz R. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. Humanit Soc Sci Commun. 2021;8:224.
3. Michels C, Schmoch U. The growth of science and database coverage. Scientometrics. 2012;93(3):831–46.
4. Haghani M, Abbasi A, Zwack CC, Shahhoseini Z, Haslam N. Trends of research productivity across author gender and research fields: A multidisciplinary and multi-country observational study. Vol. 17, PLoS ONE. 2022. e0271998 p.
5. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. J Clin Epidemiol. 2021;134:178–89.
6. Gurevitch J, Koricheva J, Nakagawa S, Stewart G. Meta-analysis and the science of research synthesis. Nature. 2018;555(7695):175–82.
7. Burns P, Rohrich R, Chung K. The Levels of Evidence and their role in Evidence-Based Medicine. Plast Reconstr Surg. 2011;128(1):305–10.
8. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: How will we ever keep up? PLoS Med. 2010;7(9):e1000326.
9. van de Schoot R, de Bruin J, Schram R, Zahedi P, de Boer J, Weijdema F, et al. An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell. 2021;3(February):125–33.
10. Ferdinands G, Schram R, de Bruin J, Bagheri A, Oberski D, Tummers L, et al. Active

learning for screening prioritization in systematic reviews: A simulation study. 2020.

11. Ferdinands G. AI-Assisted Systematic Reviewing: Selecting Studies to Compare Bayesian Versus Frequentist SEM for Small Sample Sizes. Multivariate Behav Res. 2020;56(1):153–4.

12. van Dijk SHB, Brusse-Keizer MGJ, Bucsán CC, Doggen CJM, van der Palen J, Lenferink A. Distinguishing acute heart failure from exacerbations of COPD: An AI-supported systematic literature review. PROSPERO. 2022;CRD42022283952.

13. Havrlant L, Kreinovich V. A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation). Int J Gen Syst. 2017;46(1):27–36.

14. Wang R, Cao Z, Li Y, Yu K. Utility of N-terminal pro B-type natriuretic peptide and mean platelet volume in differentiating congestive heart failure from chronic obstructive pulmonary disease. Int J Cardiol. 2013;170(2):e28–9.

15. Ouanes I, Jalloul F, Ayed S, Dachraoui F, Ouanes-Besbes L, Fekih Hassen M, et al. N-terminal proB-type natriuretic peptide levels aid the diagnosis of left ventricular dysfunction in patients with severe acute exacerbations of chronic obstructive pulmonary disease and renal dysfunction. Respirology. 2012;17(4):660–6.

16. Andrijevic I, Milutinov S, Lozanov Crvenkovic Z, Matijasevic J, Andrijevic A, Kovacevic T, et al. N-Terminal Prohormone of Brain Natriuretic Peptide (NT-proBNP) as a Diagnostic Biomarker of Left Ventricular Systolic Dysfunction in Patients with Acute Exacerbation of Chronic Obstructive Pulmonary Disease (AECOPD). Lung. 2018;196(5):583–90.

17. Hawkins NM, Khosla A, Virani SA, McMurray JJV, FitzGerald JM. B-type natriuretic peptides in chronic obstructive pulmonary disease: A systematic review. BMC Pulm Med. 2017;17(1).

18. Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. Campbell Syst Rev. 2022;18(2):1–12.

19. Clarivate Analytics. EndNote X9. 2018.

20. Covidence systematic review software. Melbourne, Australia: Veritas Health Innovation;

21. McKeown S, Mir ZM. Considerations for conducting systematic reviews: evaluating the performance of different methods for de-duplicating references. Syst Rev. 2021;10(1):4–11.

22. Ishibashi H, Hino H. Stopping Criterion for Active Learning Based on Error Stability. 2021;1:1–32. Available from: http://arxiv.org/abs/2104.01836

23. Wang W, Cai W, Zhang Y. Stability-Based Stopping Criterion for Active Learning. In: Proceedings - IEEE International Conference on Data Mining, ICDM. IEEE; 2014. p. 1019–24.

24. Blaizot A, Veettil SK, Saidoung P, Moreno-Garcia CF, Wiratunga N, Aceves-Martins M, et al. Using artificial intelligence methods for systematic review in health sciences: A systematic review. Res Synth Methods. 2022;13(3):353–62.

25. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med. 2012;22(3):276–82.

26. Bernardes RC, Botina LL, Araújo R dos S, Guedes RNC, Martins GF, Lima MAP. Artificial Intelligence-Aided Meta-Analysis of Toxicological Assessment of Agrochemicals in Bees. Front Ecol Evol. 2022;10:845608.

27. Silva GFS, Fagundes TP, Teixeira BC, Chiavegatto Filho ADP. Machine Learning for Hypertension Prediction: a Systematic Review. Curr Hypertens Rep. 2022;(3).

28. Miranda L, Paul R, Pütz B, Koutsouleris N, Müller-Myhsok B. Systematic Review of Functional MRI Applications for Psychiatric Disease Subtyping. Front Psychiatry. 2021;12:665536.

29. Schouw HM, Huisman LA, Janssen YF, Slart RHJA, Borra RJH, Willemsen ATM, et al. Targeted optical fluorescence imaging: a meta-narrative review and future perspectives. Eur J Nucl Med Mol Imaging. 2021;48:4272–92.

30. Bakkum L, Schuengel C, Sterkenburg PS, Frielink N, Embregts PJCM, de Schipper JC, et al. People with intellectual disabilities living in care facilities engaging in virtual social contact: A systematic review of the feasibility and effects on well-being. J Appl Res Intellect Disabil. 2022;35(1):60–74.

31. Huang Y, Procházková M, Lu J, Riad A, Macek P. Family Related Variables' Influences on Adolescents' Health Based on Health Behaviour in School-Aged Children Database, an AI-Assisted Scoping Review, and Narrative Synthesis. Front Psychol. 2022;13:871795.

32. Zhang W, Huang S, Lam L, Evans R, Zhu C. Cyberbullying definitions and measurements in children and adolescents: Summarizing 20 years of global efforts. Front Public Heal. 2022;10:1000504.

33. Wang Z, Nayfeh T, Tetzlaff J, O'Blenis P, Murad MH. Error rates of human reviewers during abstract screening in systematic reviews. PLoS One. 2020;15(1):e0227742.

34. Muthu S. The efficiency of machine learning-assisted platform for article screening in systematic reviews in orthopaedics. Int Orthop. 2023;47:551–6.

35. Marshall IJ, Wallace BC. Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. Syst Rev. 2019;8(1):1–10.

FIGURE CAPTIONS

Figure 1 Flowchart showing when and where to act upon when using ASReview in systematic reviewing

Footnote: Adapted the PRISMA flowchart from Haddaway et al. (18).

Figure 2 Relevant articles identified after a certain number of titles and abstracts were screened using the AI tool compared with manual screening