The effect of machine learning tools for evidence synthesis on resource use and time-to-completion: protocol for a retrospective pilot study

Ashley Elizabeth Muller (✉ aemu@fhi.no)
Norwegian Institute of Public Health
Rigor C Berg
Jose Francisco Meneses-Eschavez
Heather M. R. Ames
Tiril C. Borge
Patricia Sofia Jacobsen Jardim
Chris Cooper
Christopher James Rose

Method Article

Keywords: systematic reviewing, machine learning, artificial intelligence, automation, research waste, business process management, protocol

Posted Date: February 23rd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-1733027/v2

Abstract

Background

Machine learning (ML) tools exist that can reduce or replace human activities in repetitive or complex tasks. Yet ML is underutilized within evidence synthesis, despite the steadily growing rate of primary study publication and need to periodically update reviews to reflect new evidence. Underutilization may be partially explained by a paucity of evidence on how ML tools can reduce resource use and time-to-completion of reviews.

Methods

This protocol describes how we will answer two research questions using a retrospective study design: Is there a difference in resources used to produce reviews using recommended ML versus not using ML, and is there a difference in time-to-completion? We will also compare recommended ML use to non-recommended ML use. We will retrospectively include all reviews conducted at our institute from 1 August 2020, corresponding to the commission of the first review in our institute that used ML. We will use the results from this study to design a rigorous, multi-institutional, prospective study that will additionally explore review quality.

Conclusion

We invite other evidence synthesis groups to adopt and adapt this protocol and to collaborate with us.

Highlights

- Machine learning (ML) tools for evidence synthesis now exist but little is known about whether they lead to decreased resource use and time-to-completion of reviews.
- We propose a protocol to systematically measure any resource savings of using machine learning to produce evidence syntheses.
- Co-primary analyses will compare “recommended” ML use (in which ML replaces some human activities) and no ML use.
- We will additionally explore differences between “recommended” ML use and “non-recommended” ML use (in which ML does not replace human activities).
- We invite interested groups and researchers to collaborate with us on a future study that assesses methodological quality.

Introduction

Machine learning (ML) tools can reduce the need for humans to conduct repetitive or complex tasks. In simple terms, ML is an area of artificial intelligence developed to use computer systems that learn to complete tasks without explicit instructions. Recent estimates suggest that substantial resources could be saved if ML adoption was increased within evidence synthesis [1]. Despite its potential to reduce resource use (e.g., total person-time) and time-to-completion (e.g., time from project commission to publication), the evidence synthesis field struggles to adopt ML [2, 3]. We suggest this underutilization may be explained by the field having grown to equate human effort with methodological quality, such that automation may be seen as sacrificing quality (see also Arno et al. [4]).

The COVID-19 pandemic appears to have increased the use of ML methods in evidence synthesis [5]. In our own institution's experience, the need to map and process the onslaught of COVID-19 evidence in 2020 was a direct impetus to use ML. ML grew from being used in none of our institution's evidence syntheses before the pandemic, to 26 after the first year [6].

This protocol describes a pilot study that will estimate the effect of ML on resource use and time-to-completion. The *Background* section describes the current evidence on ML within evidence synthesis, the need for this pilot study, and the institutional context in which we will conduct this study. The *Methods* section describes our research questions, procedures, and analysis plan. This study will be conducted retrospectively using evidence syntheses produced by the Cluster for Reviews and Health Technology Assessment at the Norwegian Institute of Public Health. In the *Conclusion*, we describe our ambitions for subsequent work and a call for collaboration. We will use the results from this pilot to design a more rigorous, multinational, prospective study.

Background

Evidence synthesis and machine learning

Systematic reviews aim to identify and summarize all available evidence to draw inferences of causality, prognosis, diagnosis, prevalence, and so on, to inform policy and practice. Reviewers should adhere closely to principles of transparency, reproducibility, and methodological rigor to accurately synthesize the available evidence. These principles are pursued through adhering to explicit and pre-specified processes [7, 8].

As noted above, ML can reduce the need for humans to perform repetitive and complex tasks. "Repetitive and complex" characterizes several systematic review steps, such as assessing the eligibility of thousands of studies according to a set of inclusion criteria, extracting data, and even assessing risk of bias domains using signaling questions. Not only are most tasks repeated many times for each study, but they are often conducted by two trained researchers.

Unsurprisingly, conducting a systematic review is a resource-intensive process. Although the amount of time taken to complete health reviews varies greatly [9], fifteen months has been an estimate from both a systematic review [10] and a simulation study [11]. Cochrane suggests reviewers should prepare to spend one to two years [12], yet only half of reviews are completed within two years of protocol publication [13]. Andersen et al. also report that median time-to-publication has been increasing. A worrying estimate from 2007 is that twenty-five percent of reviews are outdated within two years of publication due to the availability of new findings [14]. Furthermore, resource use does not necessarily end with the publication of a review: many reviews — notably those published by Cochrane and health technology assessments in rapidly-advancing fields such as cancer treatment — must be updated [15].

ML offers the potential to reduce resource use, produce evidence syntheses in less time, and maintain or perhaps exceed current expectations of transparency, reproducibility, and methodological rigor. One example is the training of binary classifiers to predict the relevance of unread studies without human assessment: Aum and Choe recently used a classifier to predict systematic review study designs [16], Stansfield and colleagues to update living reviews [17], and Verdugo-Paiva and colleagues to update an entire COVID-19 database [18].

ML tools have been available for systematic reviewers for at least ten years, yet uptake has been slow. In 2013, Thomas asked why automation tools were not more widely used in evidence synthesis [19]. Since then, an increasing amount of review software with ML functionalities are available [20, 21], including functionalities that map to the most time-intensive phases [1, 9]. The evidence in favor of time savings has grown with respect to specific review phases. O'Mara-Eves and colleagues' review in 2015 found time savings of 40-70% in the screening phase when using various text mining software [22]; we reported similar or perhaps more (60-90%) time savings in 2021 [6]. Automatic classification and exclusion of non-randomized designs with a study design classifier saved Cochrane Crowd from manually screening more than 40% of identified references in 2018 [23]. We have also reported that categorizing studies using automated clustering used 33% of the time compared to manual categorization [24].

While the available estimates of time saved within distinct review phases are impressive, there are two additional outcomes that are more important to quantify: total resource use and time-to-completion. Studying resource use is important because producing evidence syntheses is expensive. Studying time-to-completion is important because answers that are late are not useful. A recent evaluation of AI use in COVID-19 reviews suggests important workload savings can be achieved (more abstracts screened and fewer full-texts inspected per review author) in AI-assisted reviews compared to reviews without AI [25]. Knowing how ML may affect total resource use would help review producers to budget and price their products and services. Knowing how ML may affect time-to-completion would help review producers decide whether to adopt ML in general or for specific projects and, if they do, how project timelines may be affected. Clark et al. conclude their report of a review conducted in two weeks, attributed to full integration of software with and without ML, as well as project management changes, by predicting that adoption of ML will increase if "the increase in efficiency associated with their use becomes more apparent" [26] (page 89).

Context

The Cluster for Reviews and Health Technology Assessments in the Norwegian Institute of Public Health is staffed by about 60 employees and, before the COVID-19 pandemic, produced up to about 50 evidence synthesis products per year. This number has roughly doubled under COVID-19. Cluster management funded the ML team in late 2020 to coordinate implementation, including building the capacity of reviewers to independently use, interpret, and explain relevant ML concepts and tools. This team is tasked with the continuous identification, process evaluation, and implementation of ML tools that can aid the production of evidence synthesis products and tailoring them to institutional procedures and processes; see Figure 1 for a schematic.

***Insert Figure 1 about here

Recommended versus non-recommended use of ML

Fifteen months after the ML team was formed, we noticed that ML is sometimes used in addition to, rather than instead of, fully manual processes. One example of this is screening titles and abstracts with a ranking algorithm. We noticed that some reviewers reach the "plateau" that indicates that there is a very high probability that all relevant studies have been identified, but then use two blinded human reviewers' time to screen potentially thousands of remaining and likely irrelevant studies, without clearly justifying the approach. Several months later, we also noticed that ML is sometimes used to replace manual processes that should not have been replaced. One example is the use of Cochrane's RCT classifier. When used as recommended to classify studies as "may be RCTs" versus "unlikely to be RCTs", the tool has been shown to perform with 99.5% recall [27]. However, some reviews seem to have placed too much trust in the tool and incorrectly excluded a portion of studies classified as "may be RCTs" (in addition to those classified as "unlikely to be RCTs"). We will therefore distinguish between two types of non-recommended ML use: *under-use* and *over-use* of ML. We have chosen these terms to avoid ascribing definitive explanations of non-recommended use. However, we suspect that under-use may be associated with being technologically skeptical or unaware of, or unconvinced by, the evidence in favor of ML use; while over-use may be associated with excessive trust in ML tools or being insufficiently critical of automated processes. See Figure 2 below for our anticipation of the temporality of ML use.

It seems self-evident that introducing a new tool (e.g., ML) — but continuing to perform the tasks the tool seeks to replace — will not result in reduced resource use or decrease time-to-completion. On the other hand, over-use of ML is likely to be associated with reduced resources use and shorter time-to-completion, which would be desirable if review quality does not suffer (see Discussion).

If ML tools can deliver the savings they promise, and are to be adopted, then it is necessary to convince reviewers to adopt these new tools and use them as recommended, neither in addition to unnecessary manual processes, nor in replacement of necessary processes. That we draw the distinction between under- and over-use of ML should underscore that we do not think that every project should use ML, or use it in the same way, but that if ML is adopted to reduce resource use or time-to-completion — as is the overarching aim in our institution — it should replace some human activities. If review quality is to be protected, this replacement should not be unfounded. There may be cases in which the use of ML alongside human activity is expected to be beneficial. For example, if the literature being studied for a particular review is characterized by poor or uncommon reporting, it may be easy for ML tools as well as humans to overlook studies that should be included. Similarly, we have found that new and inexperienced reviewers may be able to better understand a study's content by using ML in parallel with their own reviewing work [28]. Importantly, we do not mean to say that people have no role in evidence synthesis, but that it seems likely that people can make valuable higher-level contributions that machines cannot.

******Insert Figure 2 about here****

Methods

We have two research questions:

RQ 1: Is there a difference in resource use (i.e., person-time) for reviews that use ML compared to those that do not?

RQ 2. Is there a difference in time-to-completion for reviews that use ML compared to those that do not?

For each research question we will make four comparisons:

1. Use of recommended ML versus no ML (primary analysis). We hypothesize that recommended ML use is associated with less resource use and shorter time-to-completion.
2. Use of recommended ML versus non-recommended under-use of ML (secondary analysis). We hypothesize that recommended ML use is associated with less resource use and shorter time-to-completion.
3. Use of recommended ML versus non-recommended over-use of ML (secondary analysis). We hypothesize that recommended ML use is associated with more or the same resource use and longer or the same time-to-completion.

4. Use of any ML versus no ML (secondary analysis). We hypothesize that any ML use is associated with less resource use and shorter time-to-completion.

Procedures and data collection

RCB will identify reviews commissioned on or after 1 August 2020, corresponding to the commission of the first NIPH evidence synthesis that used ML. We anticipate identifying upwards of about 100 reviews, of which approximately 50 are likely to have used any ML. RCB will send a list of all potentially eligible projects to the rest of the team. RCB will separately extract outcome data for the primary and secondary analyses (see above). Resource use will be measured as number of person-hours used from commission until completion (see below) or, for ongoing projects, the number of person-hours used so far. Time-to-completion will be computed from project commission and completion dates (see below). The rest of the project team will initially be blinded to outcome to facilitate unbiased assessment of recommended versus non-recommended ML use (see below).

Norwegian commissioners have varying requirements for the time they require to deliberate on a completed review before allowing NIPH to publish it on its website. Some commissioners require six weeks, and there may be delays, which are not recorded by NIPH. We will therefore use time-to-completion, rather than time-to-publication, to prevent introducing unnecessary variance in this outcome. Time-to-completion will be calculated as the number of weeks from commission to approval for delivery to the commissioner (this includes time used on the peer review process). While we have chosen to measure time in units of week, we anticipate being able to measure commission and completion at the resolution of day (i.e., we are not limited to integer numbers of weeks). Resource use and time-to-completion will be right-censored if a review has not been completed and these outcomes will be coded as missing if they are not available. We will not attempt to impute missing data for statistical analyses, and we expect very few, if any, reviews will have such missing data.

While blinded to the outcomes (resource use and time-to-completion), JFME will tally for each review the ML functions the review used that followed the ML team's recommendations, under-used ML, and over-used ML. The ML team lead (AEM) will confirm the tallies; disagreements will be resolved by discussion.

Recommended ML will be defined as the use of ML in any review phase that is consistent with the ML team's guidance or direct recommendation (i.e., that if ML is used it should replace human activity, with appropriate justification). Because the ML team's guidance evolved over time in line with evidence and experience, we will define recommended ML use with respect to the guidance that was applicable at the time of each review. Non-recommended under-use of ML will be defined as the use of human effort to perform a review function that could have been replaced by ML had the ML team's guidance or direct recommendation been followed. For example, we will label an ML function as non-recommended under-use of ML if the review team performed ML-based screening in addition to manual screening (this goes against the ML team’s guidance and would be expected to increase resource use and delay project completion). Non-recommended over-use of ML will be defined as the use of ML to perform a review function that should have been performed by a human reviewer had the ML team’s guidance or direct recommendation been followed.would be expected to increase resource use and delay project completion). Non-recommended over-use of ML will be defined as the use of ML to perform a review function that should have been performed by a human reviewer had the ML team’s guidance or direct recommendation been followed.

A review will be classified to have: used recommended ML if the majority of ML functions used by the review followed the ML team’s recommendations; under-used ML if the majority of functions under-used ML; and over-used ML if the majority of functions over-used ML. We will classify the study to have used recommended ML if there is a tie (e.g., equal numbers of functions used as recommended and under-used).

ML use is reported within published evidence syntheses in the *Methods* and *Results* sections or in a separate *Use of machine learning* attachment. The ML team has detailed notes on all directions provided to review teams; these notes flag whether project leaders have deviated from guidance.

JFME will extract data on the following covariates from published evidence syntheses or internal sources (see Supplementary file 1):

- Review type (health technology assessment [HTA] or non-HTA).
- Synthesis type planned (none, such as in scoping reviews; pairwise meta-analysis or qualitative synthesis; or network meta-analysis).
- Whether the review is an update (yes/no).
- Number of ML functions used, by type (e.g., study identification, data extraction).
- Review phase of ML use(s).
- Field (health/medicine or welfare).

JFME will also extract the following outcome data that are unlikely to be able to be formally analyzed:

- Commissioner satisfaction, user engagement (e.g., number of downloads).

Table 1 below displays how the included studies will be described.

Table 1 — Shell table illustrating how included studies will be described.

	All reviews	Use of Machine Learning		
		Recommended	Non-recommended	
			Under-use	Over-use
Number of reviews	XXX (100%)	XXX(%)	XXX(%)	XXX(%)
Review type				
HTA	XXX(%)	XXX(%)	XXX(%)	XXX(%)
Non-HTA	XXX(%)	XXX(%)	XXX(%)	XXX(%)
Synthesis type planned				
None	XXX(%)	XXX(%)	XXX(%)	XXX(%)
pairwise meta-analysis or qualitative synthesis	XXX(%)	XXX(%)	XXX(%)	XXX(%)
network meta-analysis	XXX(%)	XXX(%)	XXX(%)	XXX(%)
ML functions (total)				
Ranking algorithm during study identification	XXX(%)	XXX(%)	XXX(%)	XXX(%)
Classifiers during study identification	XXX(%)	XXX(%)	XXX(%)	XXX(%)
Classifiers during data extraction	XXX(%)	XXX(%)	XXX(%)	XXX(%)
Clustering during study identification	XXX(%)	XXX(%)	XXX(%)	XXX(%)
Clustering during data extraction	XXX(%)	XXX(%)	XXX(%)	XXX(%)
OpenAlex during study identification	XXX(%)	XXX(%)	XXX(%)	XXX(%)
RoB assessment	XXX(%)	XXX(%)	XXX(%)	XXX(%)
Automated data extraction	XXX(%)	XXX(%)	XXX(%)	XXX(%)
Other ML	XXX(%)	XXX(%)	XXX(%)	XXX(%)

Statistical analysis

This pilot is a retrospective observational study: reviews were not randomized to use or not use ML and it is likely that ML use was more likely in certain types of reviews. We assume that healthcare reviews will have been less likely to use ML than welfare reviews because HTAs (which fall under healthcare) are more likely to adhere to established review procedures. We also assume that urgent reviews that are likely to have been completed faster — particularly those performed during the first two years of the COVID-19 pandemic — were more likely to lack protocols and would have been more likely to adopt ML to expedite the review process. We will therefore model ML use as an endogenously assigned treatment predicted by field (healthcare or welfare) and pre-specification (i.e., existence of a protocol) in all analyses.

For RQ 1, we will estimate relative number of person-hours used. Because the outcome will be right-censored for ongoing reviews and treatment is endogenously assigned we will use extended interval-data regression via Stata's eintreg command. For RQ 2, we will estimate relative mean time-to-completion (accounting for censoring) using Stata's stteffects command to account for censoring and model endogenous treatment assignment using inverse-probability-weighted regression adjustment.

As described above, resource use and time-to-completion may be lower in reviews that did not plan to perform meta-analyses, which includes qualitative syntheses. We will therefore adjust for planned use of meta-analysis in all analyses and will report adjusted estimates if there is statistical evidence of an association with outcome. If a review was not pre-specified — i.e., did not publish a protocol — we will impute that meta-analysis was not planned, even if meta-analysis was performed in the review, because lack of pre-specification is likely associated with lower resource use and time-to-completion. Because we expect to include very few reviews that used network meta-analysis (NMA), we will exclude NMAs from analysis and report resource use and time-to-completion narratively.

We will present estimates as shown in Table 2, along with sample mean numbers of person-hours and times-to-completion. We will also present Kaplan-Meier curves illustrating times-to-completion for the sample. Using information about which review phases used ML, we will perform exploratory analyses to estimate which specific phases of a review may benefit from ML. We will present point estimates with 95% confidence intervals and two-sided *p*-values. While inference will focus on confidence intervals, we will consider an estimate to be statistically significant if its *p*-value is less than 0.05.

Table 2 — Shell table illustrating how results will be presented

Type of ML Use	Sample Mean[1]	Effect Estimate[2]	*p*-value
Resource Use			
Recommended	XXX	XXX (XXX to XXX)	0.XXX
None	XXX		
Recommended	XXX	XXX (XXX to XXX)	0.XXX
Under-use	XXX		
Recommended	XXX	XXX (XXX to XXX)	0.XXX
Over-use	XXX	XXX (XXX to XXX)	
Any	XXX	XXX (XXX to XXX)	0.XXX
None	XXX		
Time-to-completion			
Recommended	XXX	XXX (XXX to XXX)	0.XXX
None	XXX		
Recommended	XXX	XXX (XXX to XXX)	0.XXX
Under-use	XXX		
Recommended	XXX	XXX (XXX to XXX)	0.XXX
Over-use	XXX	XXX (XXX to XXX)	
Any	XXX	XXX (XXX to XXX)	0.XXX
None	XXX		

ta are mean number of person-hours or weeks to completion. Sample mean resource use may be underestimated due to right-nsoring of ongoing projects.

imates are presented with 95% confidence intervals and adjusted as described in the statistical analysis section.

Additional variables may only be available for a portion of reviews, so we will report them narratively. We will use these variables to describe the reviews themselves that used recommended ML, that under- or over-used ML against recommendations, and that did not use ML. If there are very few reviews that under- or over-used ML, it may not be possible to perform some of the planned comparisons. In this case we will report the analyses that could not be performed and explain why.

We anticipate that many reviews will have used some ML functions as recommended and will have under- and over-used other functions. We will therefore also perform exploratory analyses to investigate the association between such combinations of ML use and the outcomes (resource use and time-to-completion). Specifically, we will model outcome in terms of the proportions of the numbers of ML functions classified as having been used as recommended, under-used, and over-used. This may allow us to predict the resource and time savings that may be possible if all ML functions are used as recommended, or if all of them are under-used, for example.

Conclusion

This study will have three key applications. First, we will be able to provide reasonably robust quantitative estimates of the effect of ML adoption on resource use and time-to-completion which we hope other institutions will be able to use to calculate expected resource savings were they to implement ML. Second, we will have better information for making higher-level organizational decisions about ML. Third, the effect estimates will help us prospectively power a subsequent study.

Limitations

The main limitation is the retrospective non-randomized design. We used ROBINS-I [30] to anticipate risks of bias (Supplementary file 2 reports our assessments). While ROBINS-I was designed for assessing published studies included in a systematic review, we also find such tools useful for identifying potential problems at the protocol stage. Overall, we anticipate a low risk of bias.

The most likely risks are posed by residual confounding that we cannot account for and the post-hoc classification of reviews as having used recommended and non-recommended ML. We will address the confounding issue by modelling treatment as being endogenously assigned, but some risk must remain. Ideally, we would model review type at a finer level of granularity, but this would probably lead to a model that cannot be fitted to a data set of only about 100 observations.

We will address the classification of intervention issue by blinding the researcher who will do the classification to outcome. However, this blinding will be imperfect because the researcher may be familiar with some of the included projects and their approximate duration (and proxies for resource use, such as project team size). We will be able to blind the statistician to intervention.

The sample averages (mean person-hours and mean weeks to completion) are likely to be specific to our institution (a relatively well-resourced national institute in a wealthy country), reflecting our commissions, resources, organizational procedures, and commissioner expectations. However, we anticipate that the relative effect estimates will be broadly generalizable to other institutions and research groups.

Insert Figure 3 about here

We anticipate that resource use and time-to-completion are likely associated with covariates (risk factors) such as the "complexity" of the literature, and we would ideally adjust for this in the analyses. We considered several measurable covariates that may plausibly be associated with the outcomes (e.g., number of included studies and number of pages in the report). However, we found it was generally easy to think of counterexamples (e.g., reports with few included studies that were nonetheless challenging). The only measurable variable that we think may be reliably associated with resource use and time-to-completion is planned meta-analysis. However, we will be able to explore some of these questions once data are available.

In addition to comparing resource use and time-to-completion, we considered comparing review "quality". Ideally, reviews that use ML will have equal or higher quality compared to those that do not use ML, and we hypothesize that non-recommended ML use is associated with lower quality. We surveyed commonly-used tools for assessing review quality (such as AMSTAR-2, ROBIS, and CASP checklist), but concluded that they assess human-centered procedures as proxies for review quality, rather than quality itself. For example, for a review to be judged as high quality, AMSTAR requires that two people perform study selection. The rationale is presumably that there is an unacceptably high risk of error if such tasks are performed by a single reviewer, it is desirable to implement systems that reduce risk, and that this can be done by combining the decisions of two human reviewers. In the machine learning parlance of Kearns and Valiant [29], individual reviewers are "weak learners", which can be "boosted" into a single "strong learner" that makes better decisions than any single reviewer. Rather than assessing whether a review is likely to achieve its objective (e.g., helping a decision-maker make optimal use of appropriate evidence), or an intermediate aim such as reducing the risk of incorrect study omission, commonly-used tools focus on the use of human-centric procedures as proxies. This issue is perhaps illustrated in a study that found no difference in quality between COVID-19 reviews that used AI and matched controls [25].

Call for collaboration for a prospective, multi-institutional study

We will use lessons learned through this pilot study to inform a future, prospective study. The involvement of other organizations will increase sample size and power and will enable independent assessment of another important variable: methodological and hence review quality. The hypothesis that ML does not negatively impact quality must be tested in the future, and we anticipate that participating organizations could assess each other's reviews using a tool such as AMSTAR-2 [31].

Please contact the project leader, Jose Meneses-Echavez (jose.meneses@fhi.no) if you are interested in collaboration. Please provide the following information: feedback on the methods, particularly the proposed outcomes and variables; additional ML tools that your reviews utilize or that you think are missing; the estimated number of reviews that you could provide, if relevant; and interest in assessing the methodological quality of other organizations' reviews. We also invite critical feedback om this protocol.

We conclude by suggesting a future research agenda in Figure 2.

Declarations

Author contributions

CJR conceived the study, wrote the methods section, and edited the manuscript. RCB, CJR, and AEM designed the study and drafted the manuscript. JFME drafted the data collection form, planned data extraction, and will lead the pilot study. HMRA contributed significantly to the introduction. CC contributed to the methods and discussion. PSJJ, TCB, and CC critically reviewed the final draft. All authors have read and approved the final version.

Availability of data and materials

Anonymized data and analysis code will be made publicly available.

Competing interests

The authors declare no competing interests.

Consent to publication

Not applicable.

Ethics Approval

Not applicable.

Funding

The planned evaluation will be implemented as a normal activity of the ML team, which is funded by the Cluster for Reviews and Health Technology Assessment at the Norwegian Institute of Public Health.

Acknowledgements

Thank you to Stijn Van De Velde and Jan Himmels for their instrumental expertise in building up the infrastructure for this innovation-based machine learning team.

Revision history

Version 1 of this protocol was posted to a preprint server on 7 June 2022. After becoming aware of over-use of ML, we revised the protocol on 21 November 2022, during data extraction but before any analysis or unblinding, to treat this forms of non-recommended use separately from another type of non-recommended use, namely under-use. Version 2 was posted to replace the previous version on 17 February 2023.

References

1. Clark J, McFarlane C, Cleo G, Ishikawa Ramos C, Marshall S. The Impact of Systematic Review Automation Tools on Methodological Quality and Time Taken to Complete Systematic Review Tasks: Case Study. JMIR Med Educ. 2021;7(2):e24418-e.

2. O'Connor AM, Tsafnat G, Gilbert SB, Thayer KA, Shemilt I, Thomas J, Glasziou P, Wolfe MS. Still moving toward automation of the systematic review process: a summary of discussions at the third meeting of the International Collaboration for Automation of Systematic Reviews (ICASR). Syst Rev. 2019;8(1):57.

3. Thomas J, Stansfield C, editors. Automation technologies for undertaking HTAs and systematic reviews. European Association for Health Information and Libraries (EAHIL) Conference; 2018; Cardiff, Wales.

4. Arno A, Elliott J, Wallace B, Turner T, Thomas J. The views of health guideline developers on the use of automation in health evidence synthesis. Syst Rev. 2021;10(1):16.

5. Blaizot A, Veettil SK, Saidoung P, Moreno-Garcia CF, Wiratunga N, Aceves-Martins M, Lai NM, Chaiyakunapruk N. Using artificial intelligence methods for systematic review in health sciences: A systematic review. Res Synth Methods. 2022.

6. Muller A, Ames H, Himmels J, Jardim P, Nguyen L, Rose C, Van de Velde S. Implementation of machine learning in evidence syntheses in the Cluster for Reviews and Health Technology Assessments: Final report 2020-2021. Oslo: Norwegian Institute of Public Health; 2021.

7. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts: treatments for myocardial infarction. Jama. 1992;268(2):240-8.

8. Oxman AD, Guyatt GH. The science of reviewing research a. Annals of the New York Academy of Sciences. 1993;703(1):125-34.

9. Nussbaumer-Streit B, Ellen M, Klerings I, Sfetcu R, Riva N, Mahmić-Kaknjo M, Poulentzas G, Martinez P, Baladia E, Ziganshina LE, Marqués ME, Aguilar L, Kassianos AP, Frampton G, Silva AG, Affengruber L, Spjker R, Thomas J, Berg RC, Kontogiani M, Sousa M, Kontogiorgis C, Gartlehner G. Resource use during systematic review production varies widely: a scoping review. J Clin Epidemiol. 2021.

10. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open. 2017;7(2):e012545.

11. Pham B, Jovanovic J, Bagheri E, Antony J, Ashoor H, Nguyen TT, Rios P, Robson R, Thomas SM, Watt J, Straus SE, Tricco AC. Text mining to support abstract screening for knowledge syntheses: a semi-automated workflow. Syst Rev. 2021;10(1):156.

12. Cochrane Community. Proposing and registering new Cochrane Reviews: Cochrane; [updated 2022.

13. Andersen MZ, Gulen S, Fonnes S, Andresen K, Rosenberg J. Half of Cochrane reviews were published more than 2 years after the protocol. J Clin Epidemiol. 2020;124:85-93.

14. Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? A survival analysis. Ann Intern Med. 2007;147(4):224-33.

15. Elliott JH, Synnot A, Turner T, Simmonds M, Akl EA, McDonald S, Salanti G, Meerpohl J, MacLehose H, Hilton J, Tovey D, Shemilt I, Thomas J. Living systematic review: 1. Introduction-the why, what, when, and how. J Clin Epidemiol. 2017;91:23-30.

16. Aum S, Choe S. srBERT: automatic article classification model for systematic review using BERT. Syst Rev. 2021;10(1):285.

17. Stansfield C, Stokes G, Thomas J. Applying machine classifiers to update searches: Analysis from two case studies. Res Synth Methods. 2022;13(1):121-33.

18. Verdugo-Paiva F, Vergara C, Avila C, Castro J, Cid J, Contreras V, Jara I, Jimenez V, Lee MH, Munoz M, Rojas-Gomez AM, Roson-Rodriguez P, Serrano-Arevalo K, Silva-Ruz I, Vasquez-Laval J, Zambrano-Achig P, Zavadzki G, Rada G. COVID-19 L.OVE repository is highly comprehensive and can be used as a single source for COVID-19 studies. J Clin Epidemiol. 2022.

19. Thomas J. Diffusion of innovation in systematic review methodology: Why is study selection not yet assisted by automation? OA Evidence-Based Medicine. 2013;1(12):6.

20. Harrison H, Griffin SJ, Kuhn I, Usher-Smith JA. Software tools to support title and abstract screening for systematic reviews in healthcare: an evaluation. BMC Med Res Methodol. 2020;20(1):7.

21. van der Mierden S, Tsaioun K, Bleich A, Leenaars CHC. Software tools for literature screening in systematic reviews in biomedical research. ALTEX - Alternatives to animal experimentation. 2019;36(3):508-17.

22. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Syst Rev. 2015;4(1):5.

23. Thomas J, McDonald S, Noel-Storr A, Shemilt I, Elliott J, Mavergames C, Marshall IJ. Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane Reviews. Journal of Clinical Epidemiology. 2021;133:140-51.

24. Muller A, Ames H, Jardim P, Rose C. Machine learning in systematic reviews: Comparing automated text clustering with Lingo3G and human researcher categorization in a rapid review. Res Synth Methods. 2021.

25. Tercero-Hidalgo JR, Khan KS, Bueno-Cavanillas A, Fernández-López R, Huete JF, Amezcua-Prieto C, Zamora J, Fernández-Luna JM. Artificial intelligence in COVID-19 evidence syntheses was underutilized, but impactful: a methodological study. Journal of Clinical Epidemiology. 2022;148:124-34.

26. Clark J, Glasziou P, Del Mar C, Bannach-Brown A, Stehlik P, Scott AM. A full systematic review was completed in 2 weeks using automation tools: a case study. J Clin Epidemiol. 2020;121:81-90.

27. Thomas J, McDonald S, Noel-Storr A, Shemilt I, Elliott J, Mavergames C, Marshall IJ. Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane Reviews. J Clin Epidemiol. 2021;133:140-51.

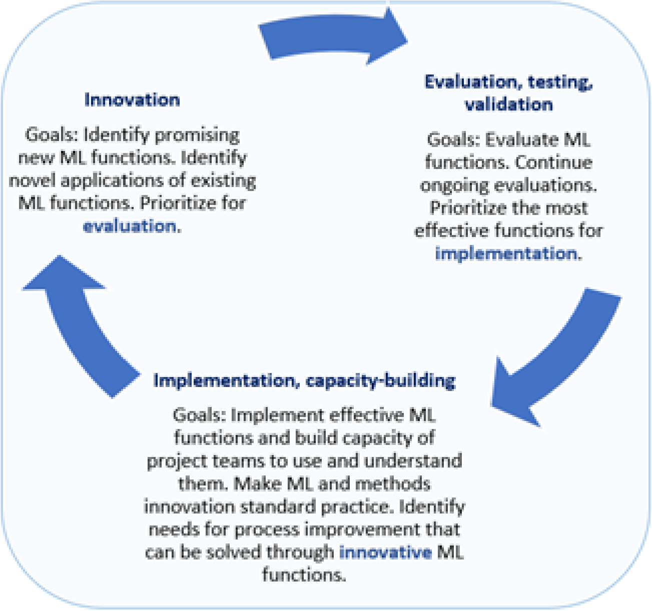
28. Jardim PSJ, Rose CJ, Ames HMR, Meneses-Echavez JF, Van de Velde S, Muller AE. Automating risk of bias assessment in systematic reviews: a real-time mixed methods comparison of human researchers to a machine learning system. BMC Med Res Methodol. in press.

29. Kearns MJ, Valiant LG. Cryptographic limitations on learning Boolean formulae and finite automata. In: Hanson SJ, Remmele W, Rivest RL, editors. Machine Learning: From Theory to Applications: Cooperative Research at Siemens and MIT. Berlin, Heidelberg: Springer Berlin Heidelberg; 1993. p. 29-49.

30. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I, Carpenter JR, Chan A-W, Churchill R, Deeks JJ, Hróbjartsson A, Kirkham J, Jüni P, Loke YK, Pigott TD, Ramsay CR, Regidor D, Rothstein HR, Sandhu L, Santaguida PL, Schünemann HJ, Shea B, Shrier I, Tugwell P, Turner L, Valentine JC, Waddington H, Waters E, Wells GA, Whiting PF, Higgins JP. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355:i4919.

31. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, Moher D, Tugwell P, Welch V, Kristjansson E, Henry DA. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ. 2017;358:j4008.

Figures



Innovation

Goals: Identify promising new ML functions. Identify novel applications of existing ML functions. Prioritize for evaluation.

Evaluation, testing, validation

Goals: Evaluate ML functions. Continue ongoing evaluations. Prioritize the most effective functions for implementation.

Implementation, capacity-building

Goals: Implement effective ML functions and build capacity of project teams to use and understand them. Make ML and methods innovation standard practice. Identify needs for process improvement that can be solved through innovative ML functions.

Figure 1

The machine learning team identifies promising ML tools or applications, evaluates a portion of these, and implements those that are effective

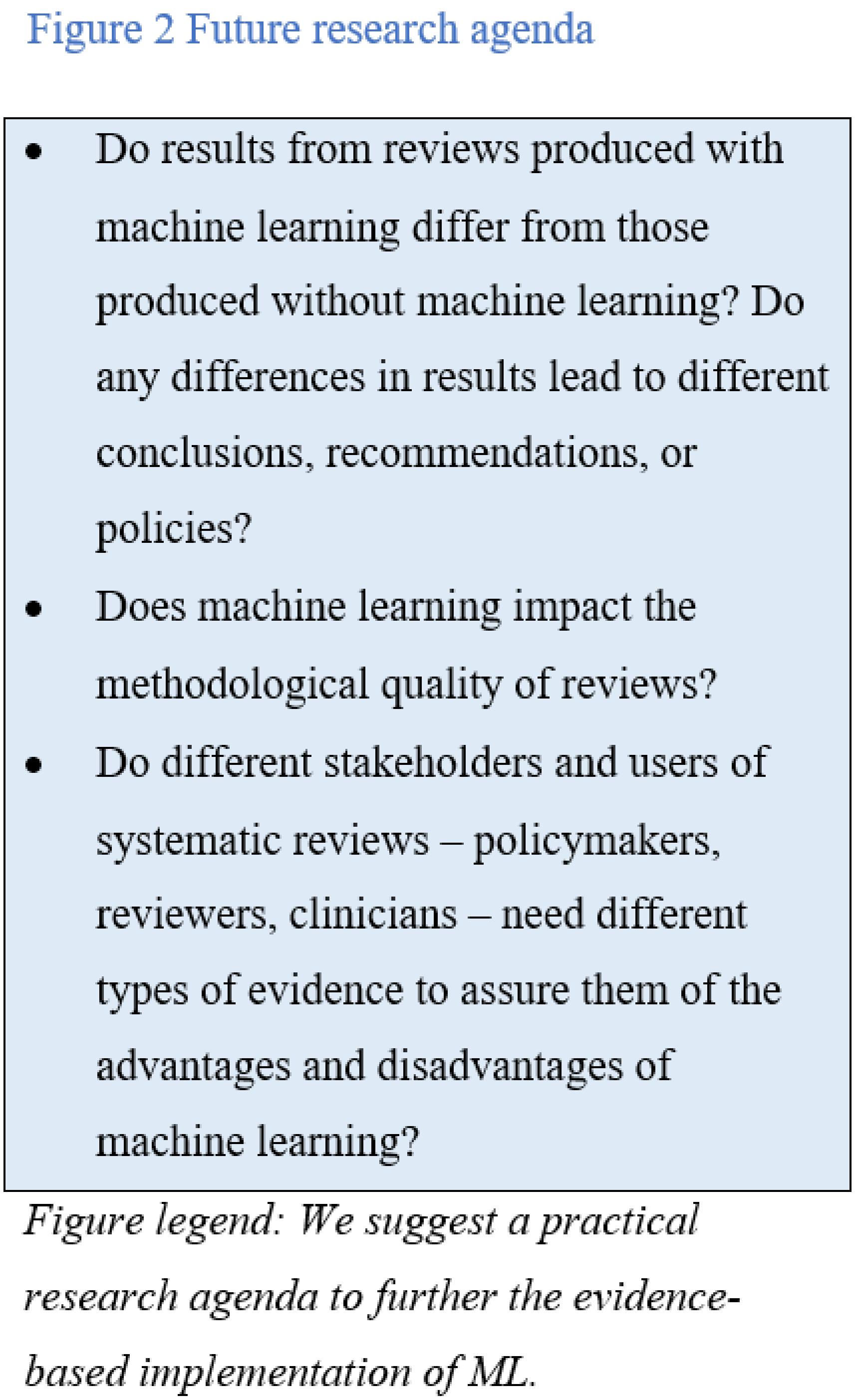


Figure 2 Future research agenda

- Do results from reviews produced with machine learning differ from those produced without machine learning? Do any differences in results lead to different conclusions, recommendations, or policies?
- Does machine learning impact the methodological quality of reviews?
- Do different stakeholders and users of systematic reviews – policymakers, reviewers, clinicians – need different types of evidence to assure them of the advantages and disadvantages of machine learning?

Figure legend: We suggest a practical research agenda to further the evidence-based implementation of ML.

Figure 2

We suggest a practical research agenda to further the evidence-based implementation of ML

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementary1dataextraction.xlsx
- Supplementary2ROBINSI.docx