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
Rigmor C Berg
Jose Francisco Meneses-Echavez
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: June 7th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1733027/v1

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.

Study design and setting: 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 that merely adds ML use to existing procedures. 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.

What Is New

- We propose a protocol to systematically measure any resource savings of using machine learning to produce evidence syntheses.
- We propose a definition of “non-recommended” use of machine learning and will test the hypothesis that this does not save resources.

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 will compare use and non-use of ML with respect to resource use (person-hours of effort) and time-to-completion.
- 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.

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 (Antman, Lau et al. 1992, Oxman and Guyatt 1993).

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 (Nussbaumer-Streit, Ellen et al. 2021), fifteen months has been an estimate from both a systematic review (Borah, Brown et al. 2017) and a simulation study (Pham, Jovanovic et al. 2021). Cochrane suggests reviewers should prepare to spend one to two years (Cochrane Community), yet only half of reviews are completed within two years of protocol publication (Andersen, Gulen et al. 2020). 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 (Shojania, Sampson et al. 2007). 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 (Elliott, Synnot et al. 2017).

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 (Aum and Choe 2021), Stansfield and colleagues to update living reviews (Stansfield, Stokes et al. 2022), and Verdugo-Paiva and colleagues to update an entire COVID-19 database (Verdugo-Paiva, Vergara et al. 2022).

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 (Thomas 2013). Since then, an increasing amount of review software with ML functionalities are available (van der Mierden, Tsaioun et al. 2019, Harrison, Griffin et al. 2020), including functionalities that map to the most time-intensive phases (Clark, McFarlane et al. 2021, Nussbaumer-Streit, Ellen et al. 2021). 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 (O'Mara-Eves, Thomas et al. 2015); we reported similar or perhaps more (60–90%) time savings in 2021 (Muller, Ames et al. 2021). 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 (Thomas, McDonald et al. 2021). We have also reported that categorizing studies using automated clustering used 33% of the time compared to manual categorization (Muller, Ames et al. 2021).

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. We are unaware of any studies that have compared the use of ML and human-based review methods with respect to these outcomes. 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” (Clark, Glasziou et al. 2020) (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 *Fig. 1* for a schematic.

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, reaching the “plateau” that indicates all relevant studies have been identified, but then continuing to use two blinded human reviewers to screen thousands of remaining and likely irrelevant studies.

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. 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. This protocol therefore distinguishes between “non-recommended” ML that merely adds additional tasks to normal, manual procedures, and “recommended” ML that corresponds to some level of automation that replaces manual procedures.

We do not mean to say 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. There may be cases in which the use of ML alongside human activity is expected to be beneficial, for example if it is expected that important studies may be easy to miss even by humans, or to help new and inexperienced reviewers learn (Jardim, Rose et al. in press). 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.

Methods

We have two research questions:

RQ 1: Is there a difference in resource use (i.e., person-time) by 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?

We hypothesize that reviews that use fully manual procedures (without ML) are more resource-intensive and take more time to complete than reviews that use ML. We further hypothesize that non-recommended use of ML (see below) will not lead to reduced resource use or time-to-completion, while recommended use of ML will. For each research question we will therefore make three comparisons:

1. Use of recommended ML versus no ML (primary analysis)
2. Use of recommended ML versus non-recommended ML (secondary analysis)
3. Use of any ML versus no ML (secondary analysis)

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, JME will code each review as having used recommended versus non-recommended ML and the ML team lead (AEM) will confirm the coding; 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). Non-recommended ML will be defined as ML that deviates from ML team guidance or direct recommendation. For example, we will label a review as having used non-recommended 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). 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 also has detailed notes on all technical assistance provided to review teams; these notes flag whether project leaders have deviated from using ML according to teaching and recommendations.

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

- Synthesis type planned (none, such as in scoping reviews; pairwise meta-analysis or qualitative synthesis; or network meta-analysis).
- Review type (health technology assessment [HTA] or non-HTA).
- Type of ML use and review phase.
- Field (health/medicine or welfare).

JME 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).

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 the included 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 modelled 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, outcomes 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. 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 1, 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 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 1
— 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
Non-recommended	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
Non-recommended	XXX		
Any	XXX	XXX (XXX to XXX)	0.XXX
None	XXX		

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

2. Estimates 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 used non-recommended ML, and that did not use ML.

If there is a small number of observations in the non-recommended group, this may not be analyzable. In this case, we will only report the primary comparison.

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 (Sterne, Hernán et al. 2016) 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.

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.

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 Error! Reference source not found..

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.

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, previous members of this machine learning team, for their instrumental expertise in building up the infrastructure for an innovation-based team.

References

1. Andersen, M. Z., S. Gulen, S. Fonnes, K. Andresen and J. Rosenberg (2020). "Half of Cochrane reviews were published more than 2 years after the protocol." J Clin Epidemiol **124**: 85–93.
2. Antman, E. M., J. Lau, B. Kupelnick, F. Mosteller and T. C. Chalmers (1992). "A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts: treatments for myocardial infarction." Jama **268**(2): 240–248.
3. Arno, A., J. Elliott, B. Wallace, T. Turner and J. Thomas (2021). "The views of health guideline developers on the use of automation in health evidence synthesis." Syst Rev **10**(1): 16.
4. Aum, S. and S. Choe (2021). "srBERT: automatic article classification model for systematic review using BERT." Syst Rev **10**(1): 285.

5. Blaizot, A., S. K. Veettil, P. Saidoung, C. F. Moreno-Garcia, N. Wiratunga, M. Aceves-Martins, N. M. Lai and N. Chaiyakunapruk (2022). "Using artificial intelligence methods for systematic review in health sciences: A systematic review." Res Synth Methods.
6. Borah, R., A. W. Brown, P. L. Capers and K. A. Kaiser (2017). "Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry." BMJ Open **7**(2): e012545.
7. Clark, J., P. Glasziou, C. Del Mar, A. Bannach-Brown, P. Stehlik and A. M. Scott (2020). "A full systematic review was completed in 2 weeks using automation tools: a case study." J Clin Epidemiol **121**: 81–90.
8. Clark, J., C. McFarlane, G. Cleo, C. Ishikawa Ramos and S. Marshall (2021). "The Impact of Systematic Review Automation Tools on Methodological Quality and Time Taken to Complete Systematic Review Tasks: Case Study." JMIR medical education **7**(2): e24418-e24418.
9. Cochrane Community. (2022). "Proposing and registering new Cochrane Reviews." Retrieved 1.4.2022, 2022.
10. Elliott, J. H., A. Synnot, T. Turner, M. Simmonds, E. A. Akl, S. McDonald, G. Salanti, J. Meerpohl, H. MacLehose, J. Hilton, D. Tovey, I. Shemilt and J. Thomas (2017). "Living systematic review: 1. Introduction-the why, what, when, and how." J Clin Epidemiol **91**: 23–30.
11. Harrison, H., S. J. Griffin, I. Kuhn and J. A. Usher-Smith (2020). "Software tools to support title and abstract screening for systematic reviews in healthcare: an evaluation." BMC Med Res Methodol **20**(1): 7.
12. Jardim, P. S. J., C. J. Rose, H. M. R. Ames, J. F. Meneses-Echavez, S. Van de Velde and A. E. Muller (in press). "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.
13. Muller, A., H. Ames, J. Himmels, P. Jardim, L. Nguyen, C. Rose and S. Van de Velde (2021). 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: 83.
14. Muller, A., H. Ames, P. Jardim and C. Rose (2021). "Machine learning in systematic reviews: Comparing automated text clustering with Lingo3G and human researcher categorization in a rapid review." Res Synth Methods.
15. Nussbaumer-Streit, B., M. Ellen, I. Klerings, R. Sfetcu, N. Riva, M. Mahmić-Kaknjo, G. Poulentzas, P. Martinez, E. Baladia, L. E. Ziganshina, M. E. Marqués, L. Aguilar, A. P. Kassianos, G. Frampton, A. G. Silva, L. Affengruber, R. Spjker, J. Thomas, R. C. Berg, M. Kontogiani, M. Sousa, C. Kontogiorgis and G. Gartlehner (2021). "Resource use during systematic review production varies widely: a scoping review." J Clin Epidemiol.
16. O'Connor, A. M., G. Tsafnat, S. B. Gilbert, K. A. Thayer, I. Shemilt, J. Thomas, P. Glasziou and M. S. Wolfe (2019). "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 **8**(1): 57.

17. O'Mara-Eves, A., J. Thomas, J. McNaught, M. Miwa and S. Ananiadou (2015). "Using text mining for study identification in systematic reviews: a systematic review of current approaches." Syst Rev **4**(1): 5.
18. Oxman, A. D. and G. H. Guyatt (1993). "The science of reviewing research a." Annals of the New York Academy of Sciences **703**(1): 125–134.
19. Pham, B., J. Jovanovic, E. Bagheri, J. Antony, H. Ashoor, T. T. Nguyen, P. Rios, R. Robson, S. M. Thomas, J. Watt, S. E. Straus and A. C. Tricco (2021). "Text mining to support abstract screening for knowledge syntheses: a semi-automated workflow." Syst Rev **10**(1): 156.
20. Shojania, K. G., M. Sampson, M. T. Ansari, J. Ji, S. Doucette and D. Moher (2007). "How quickly do systematic reviews go out of date? A survival analysis." Ann Intern Med **147**(4): 224–233.
21. Stansfield, C., G. Stokes and J. Thomas (2022). "Applying machine classifiers to update searches: Analysis from two case studies." Res Synth Methods **13**(1): 121–133.
22. Sterne, J. A., M. A. Hernán, B. C. Reeves, J. Savović, N. D. Berkman, M. Viswanathan, D. Henry, D. G. Altman, M. T. Ansari, I. Boutron, J. R. Carpenter, A.-W. Chan, R. Churchill, J. J. Deeks, A. Hróbjartsson, J. Kirkham, P. Jüni, Y. K. Loke, T. D. Pigott, C. R. Ramsay, D. Regidor, H. R. Rothstein, L. Sandhu, P. L. Santaguida, H. J. Schünemann, B. Shea, I. Shrier, P. Tugwell, L. Turner, J. C. Valentine, H. Waddington, E. Waters, G. A. Wells, P. F. Whiting and J. P. Higgins (2016). "ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions." BMJ **355**: i4919.
23. Thomas, J. (2013). "Diffusion of innovation in systematic review methodology: Why is study selection not yet assisted by automation?" OA Evidence-Based Medicine **1**(12): 6.
24. Thomas, J., S. McDonald, A. Noel-Storr, I. Shemilt, J. Elliott, C. Mavergames and I. J. Marshall (2021). "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 **133**: 140–151.
25. Thomas, J. and C. Stansfield (2018). Automation technologies for undertaking HTAs and systematic reviews. European Association for Health Information and Libraries (EAHIL) Conference, Cardiff, Wales.
26. van der Mierden, S., K. Tsaioun, A. Bleich and C. H. C. Leenaars (2019). "Software tools for literature screening in systematic reviews in biomedical research." ALTEX - Alternatives to animal experimentation **36**(3): 508–517.
27. Verdugo-Paiva, F., C. Vergara, C. Avila, J. Castro, J. Cid, V. Contreras, I. Jara, V. Jimenez, M. H. Lee, M. Munoz, A. M. Rojas-Gomez, P. Roson-Rodriguez, K. Serrano-Arevalo, I. Silva-Ruz, J. Vasquez-Laval, P. Zambrano-Achig, G. Zavadzki and G. Rada (2022). "COVID-19 L.OVE repository is highly comprehensive and can be used as a single source for COVID-19 studies." J Clin Epidemiol.

Figures

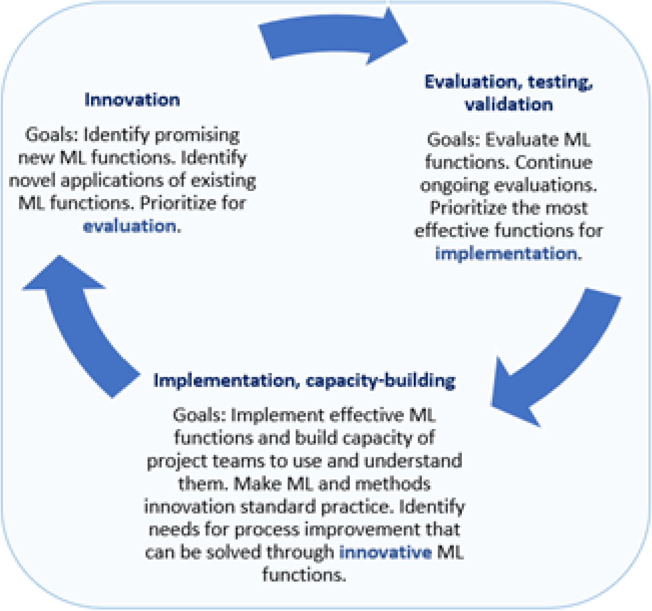
Innovation
Goals: Identify promising new ML functions. Identify novel applications of existing ML functions. Prioritize for evaluation.
Evaluation, testing, validation
Goals: Evaluate ML functions. Continue ongoing evaluations. Prioritize the most effective functions for implementation.
Implementation, 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

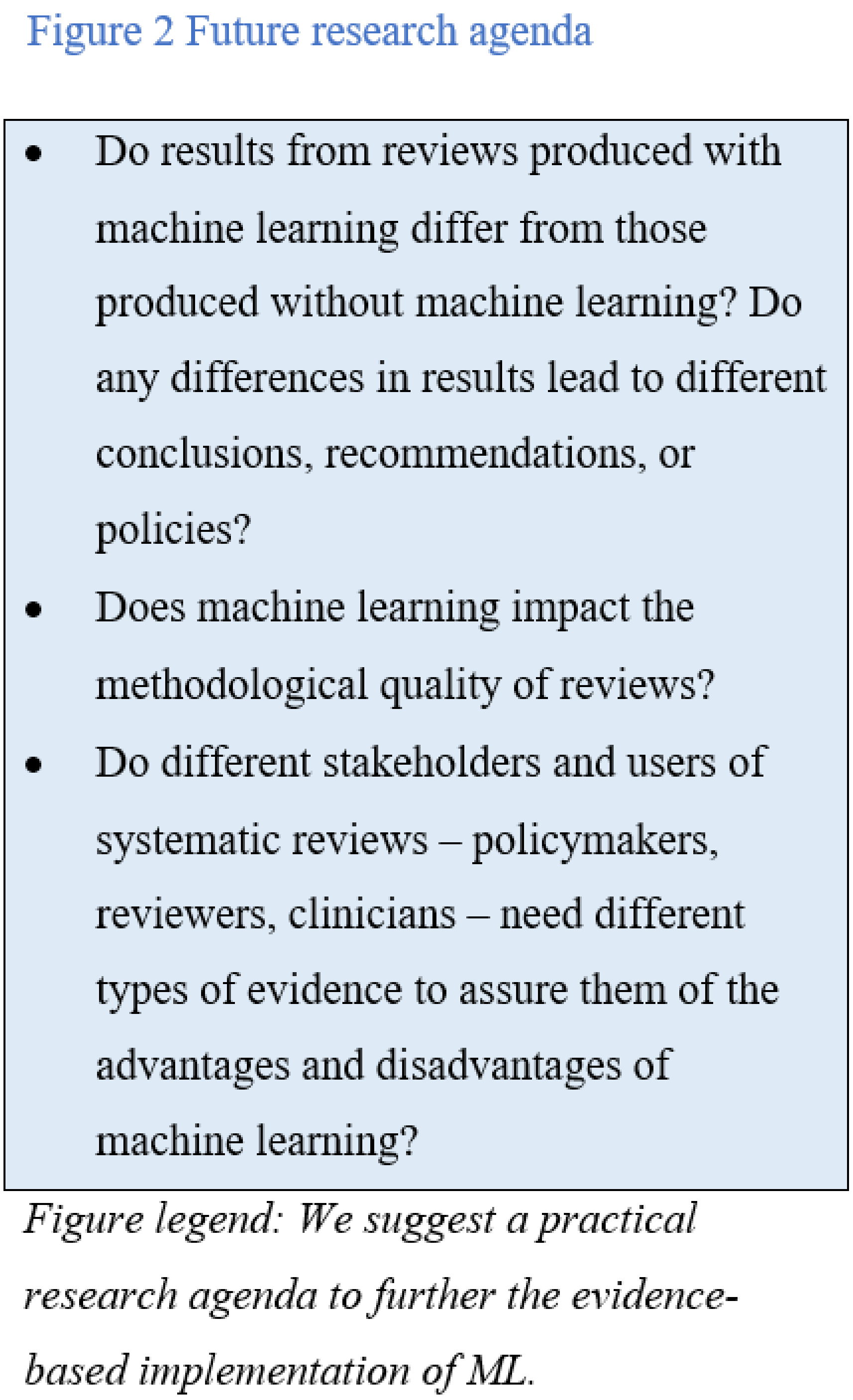

Figure 2 Future research agenda

- 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