Machine Learning in Systematic Review: A Review of Performance, Validity, and Usability

Joshua R Carpenter

Department of Statistics, Brigham Young University

WRTG 316: Technical Communication

Professor Shaylee Erickson

31 July 2021

Abstract

Researchers in evidence synthesis have been exploring ways to shorten the process of systematic review without compromising validity. Many tools already exist to automate parts of the process. This narrative review considers the performance, validity, and usability of existing machine learning methods in systematic review and identifies areas for development. ML methods can be used to achieve work savings in SR without compromising scientific rigor, but it will take significant work before they reach widespread acceptance. Developers of ML tools for SR should focus on making their tools user friendly, feature rich, and flexible.

Systematic review (SR) is becoming increasingly time-consuming and costly as the body of scientific literature expands (Lau, 2019). Researchers in evidence synthesis have been exploring ways to shorten the process without compromising the validity of the reviews. Some of the most time-consuming and repetitive parts of the review process seem more suited to a computer than a human but, at present, are performed manually. Advances in the field of machine learning (ML) invite the possibility of performing such complex but repetitive tasks automatically. Automated tools have been developed for use in several parts of the SR process including screening, data extraction, risk of bias assessment, search, qualitative synthesis, and meta-analysis. This review will focus primarily on the first two, which are the most developed: screening and data extraction.

There are several tools available and in development that utilize machine learning to assist with literature review such as Abstrackr, EPPI-Reviewer, and RobotReviewer, but these tools have not yet gained widespread acceptance or use. This is likely because it is unclear whether these programs introduce bias into the selection process. They can also be difficult to use, and many technical writers struggle to incorporate machine learning tools into their workflow. In their current state, all the available programs still require a human to check their work. Although the available tools may save some time and effort, the vision is to create a tool that can automate much more of the literature review process.

In this review, I seek to provide guidance to developers of ML tools for SR as to what needs improvement in order to bring about more widespread adoption of these tools. To do so, I assess existing ML tools and processes for SR in three key areas: performance, validity, and usability. I found that existing ML tools for SR can achieve significant work savings without compromising the rigor of a review, that it is possible to show that ML tools are valid and unbiased, and that a focus on developing flexible tools that work together may increase adoption.

I reviewed articles written since 2015 on the subject of ML in SR, hand selected the most pertinent articles, and extracted results and researchers' opinions pertaining to those three evaluation metrics. Because of the scoping nature of the review, research questions were decided post hoc. Unlike a scoping review, however, my search was not extensive; I reviewed in detail only a convenient selection of articles. For those reasons, this review should be considered biased and not a comprehensive overview.

Performance

The first question to answer is whether ML tools for SR actually work. The performance of ML in SR is the simplest area to study and evaluate formally, so multiple articles explore this area, particularly screening tools. Performance in SR automation is usually a trade between precision and recall. Performance in screening is usually measured in percentage of relevant articles detected (recall) and percentage of articles that do not need to be screened (work savings). Performance in data extraction is usually measured by the percentage of relevant sentences extracted and the percentage of data points within sentences correctly classified. Also important is the relevance of sentences extracted. In the case of SR, the most important indicator is recall because it is very important to detect all relevant articles or information.

Some studies evaluating automated screening technology have run a simulated screening on a set of pre-labeled articles and compared the results of the real screening with the results of the machine assisted simulation. Others actually screened articles using one or more existing ML screening tools and screened the same articles in some other way as a comparator. Chai et al. (2021) did both a simulated and a real-world analysis of Research Screener. Gartlehner et al. (2019) compared automated and semi-automated screening with DistillerAI to a gold standard of screening with 5 reviewers. Gates et al. (2020) did various retrospective simulations using Abstrackr. Gates et al. (2019) also did a retrospective simulation comparing Abstrackr, DistillerSR, and RobotAnalyst. Callaghan and Müller-Hansen (2020) developed statistical stopping criteria and tested them under real-world conditions and Brockmeier et al. (2019) explored improving recall and precision using PICO recognition. Each study had similar conclusions.

Several studies showed that automated screening can achieve high recall and still save work (Callaghan & Müller-Hansen, 2020; Chai et al., 2021; Gates et al., 2020). Although Gartlehner et al. (2019) and Gates et al. (2019) had very poor recall, their methods used minimal human assistance in the screening process. A more human-led, machine-assisted approach seems to yield better results. For example, Gates et al. (2020) found that using ML to assist one of two dual-independent screeners achieved work-savings with "relatively small risk of missing relevant studies" and Chai et al. (2021) claim that Research Screener can identify one hundred percent of relevant articles with at least fifty percent work savings. This last claim, however, is considerably different from findings of other studies and is not backed by formal hypothesis testing. Callaghan and Müller-Hansen (2020), on the other hand, do use formal hypothesis testing to validate their new method for developing stopping criteria. Although Callaghan and Müller-Hansen achieved much lower work savings (average 17%) than Chai et al., their evaluation is much more robust. All studies did achieve high recall with some work savings using a machine-assisted approach.

Although it is clear that ML screening tools save some work in testing, it is unclear to what extent. Work savings in each of the studies ranged from 4% to 95% under different circumstances, with different tools, and with different types of reviews. Although Gates et al. (2019) had up to 95% work savings, they had very poor recall, so that number can be safely discounted, and although Tsou et al. (2020) had work savings as low as 4%, that review was a low outlier. The number reported by Callaghan and Müller-Hansen (2020), 17%, seems most trustworthy and most realistic because they closely simulated a real-world situation and backed their results with formal hypothesis testing. Under some circumstances, work savings may be much higher. For example, Gates et al. (2020) found that work savings were higher for reviews involving large volumes of literature, and Tsou et al. (2020) speculate that ML would perform better when a high quality training set is available, such as when updating reviews or when checking for missed citations. Potential work savings for some reviews may be as high as 50% or 60% as reported by Chai et al. (2021) and Tsou et al. (2020) respectively.

Researchers are still exploring ways to improve work savings without compromising recall. Brockmeier et al. (2019) saw improvement in performance by adding PICO recognition to their screening algorithm; Tsou et al. (2020) suggests that cleaning references beforehand to remove potentially confusing articles could improve performance; and Gartlehner et al. (2019) and Callaghan and Müller-Hansen (2020) point out that work saved would increase with better stopping criteria. Chai et al. (2021) also note that most ML tools for SR use out of date ML methods and newer methods already shown to perform better in general may improve performance in SR tools as well. Their study of Research Screener seems to validify that view, as it uses more modern ML methods and performs much better than existing tools. Taking steps such as these may help ML in SR to reach the maximum potential work savings.

Validity

The SR process follows a rigorous standard of quality control, and many question the ability of ML to fit into that strict framework. To be accepted in the SR community, ML tools must be proven valid. Specifically, researchers and guideline developers want to know if ML tools can produce the same quality of work as a human reviewer (O'Connor et al., 2019). Health guideline developers have expressed skepticism that a machine can make the sort of “value judgements” that are required in the SR process (Arno et al., 2021) and researchers have said that they view some ML tools as “untrustworthy” (Gates et al., 2019). The widespread adoption of ML tools for SR, therefore, depends on demonstrating their validity to reviewers and guideline developers alike.

The biggest concern about ML tools for SR is that they will miss relevant information and therefore cause systematic bias. However, as mentioned in the performance section, several studies have shown that work can be saved without any cost to recall (Callaghan & Müller-Hansen, 2020; Chai et al., 2021; Gates et al., 2020; Tsou et al., 2020). Callaghan and Müller-Hansen point out that the worst possible consequence of their method is low work savings. This is a compelling argument for its adoption because it does not threaten the validity of the review and is still almost guaranteed to save work. That argument could also be applied to many other ML methods, given the right constraints. Therefore, the issue of thoroughness is really an issue of finding the right constraints and not an issue of coming up with a valid ML method.

Another major concern with ML tools, especially for guideline developers, is transparency (Arno et al., 2021). Interested parties want to be able to see what the tool is doing to be sure that it does not miss anything. This is completely in line with the idea of SR where method is very important, and each part of the method needs to be known and verified. The need for transparency is augmented by a general distrust of ML and artificial intelligence (Arno et al., 2021; O'Connor et al., 2019). Brockmeier et al. (2019) say that trust can be increased by showing what the ML is doing. Their idea is that the user can see how much each PICO term influenced the inclusion decision. Beller et al. (2018) suggest, to the end of transparency and advancement, that "every automation technique should be shared."

In parallel with the other two concerns is the need for formal validation. Like everything in the scientific community, ML tools for SR must be thoroughly verified by replicable methods before they are accepted in SR's. Although many studies have demonstrated the performance of ML tools for SR, very few have studied their validity. The most thorough statistical analysis of the validity of ML in SR was done by Callaghan and Müller-Hansen (2020). The methods they used allow researchers to say with a given level of confidence what percentage of relevant studies were included after screening a certain number. They were able to achieve consistent recall and still save work. Future studies should follow that model to further validate existing ML tools for SR.

Usability

Usability can be a major concern for the adoption of ML in SR. If tools are too complex to use, it will be hard for users to embrace them. While many studies have been done on the performance of ML tools for SR, relatively few have studied the user experience (UX). The one study in this review which specifically studied user experience, Gates et al. (2019), had differing results for each of the three tools they studied, but found that each had things to improve. They found that Abstracker was simple and easy to use but had a rudimentary user interface (UI); DistillerSR had a clean UI but seemed overly complicated; and Robot Analyst was “pretty” but hard to use and had many error messages. Participants mentioned that all three tools were slow and that the export formats were impractical. These factors may be impeding adoption of ML tools for SR.

One of the primary impediments to adoption of ML tools for SR is that they disrupt the SR workflow (O'Connor et al., 2019). Beller et al. (2018) emphasize the importance of flexibility in combining different tools to create a cohesive workflow. They claim that to suit the needs of different literature reviews, each tool should be flexible and provide common export formats and an API. Reviewers who participated in the UX study by Gates et al. (2019) confirmed this need. They found that predictions were exported in an impractical format and were difficult to download. Achieving better integration between tools may require standardization of export formats and increased communication between developers.

Better usability is also dependent on communication between developers and users. Beller et al. (2018) propose that development should be a cycle of deployment and user feedback. As we allow users to drive development, we may be able to produce tools that better meet their needs. For example, Chai et al. (2021) point out that many ML screening tools do not provide convenient features commonly found in commercial software such as team management and conflict resolution. Gates et al. (2019) suggest that a greater focus on usability by developers may encourage adoption. The results of Gates et al. (2019) and Beller et al. (2018) demonstrate that while performance is important, developers also need to adhere to good principles of UX design.

Another impediment to adoption brought up by O'Connor et al. (2019) is the technical complexity or barrier to set up of some tools. Health guideline developers surveyed by Arno et al. (2021) opine that if the learning curve is too steep, it will discourage users. In order for ML tools for SR to diffuse past early adopters, programmers should target their tools towards non-technical users. Some programmers are already making an effort to lower the barrier to entry. For example, many tools like Rayyan (Olofsson et al., 2017) and Research Screener (Chai et al., 2021) have web interfaces where the tools can be used with no set up at all. Respondents with little technical experience said that Rayyan was easy to learn and easy to use (Olofsson et al., 2017). Other researchers, such as Westgate (2019), recognize the need for easy access and are working to provide it. Increasing the accessibility of ML tools for SR is a work in progress, but there are many tools already available and usable by those with little technical experience.

Finally, it is important that ML tools automate as much as possible of the SR process. Beller et al. (2018) claim that automation can assist with all parts of the SR process, from searching to evidence synthesis. As each part slowly becomes more automated, time saved will continue to increase. While total automation is far beyond our present capability (Marshall & Wallace, 2019), developers should seek to reduce human involvement. Screening tools should be optimized to avoid large training sets as mentioned by Chai et al. (2021) and data extraction tools should push towards less supervised learning such as the system designed by Blake and Lucic (2015). When the process becomes fully automated, usability will become immaterial (Gates et al., 2019).

Discussion

Most researchers agree that the potential benefits of ML in SR are highly desirable; they are, however, skeptical that those benefits can be achieved in practice. Several impediments to the practical application of ML methods in SR are the necessity for large training sets, indefinite stopping criteria, lack of features, and installation difficulty (Chai et al., 2021). In short, ML tools are only useful as far as they are user friendly, transparent, flexible, and integrable into a standard SR workflow.

As discussed in the performance section, ML screening methods can provide significant work savings; however, Tsou et al. (2020) point out that it is difficult to tell if those theoretical time-savings translate into real value. To save time by screening with ML there must be a cutoff point where it is safe for researchers to stop screening. Unfortunately, it is unclear when that point is reached (Marshall & Wallace, 2019). Stopping criteria are an essential part of applying any ML algorithm to screening, but little research has been done to develop them (Callaghan & Müller-Hansen, 2020). Thus, for practical work savings to be achieved, more rigorous stopping criteria are needed.

In order to maximize utility, ML tools should be developed for all parts of the literature review process (Beller et al., 2018). Tools need to provide a flexible amount of automation so that users can try them before committing fully (Arno et al., 2021). A good example of this is Brockmeier et al. (2019), whose method could prove itself useful just by highlighting relevant PICO terms. Several tools also allow the user to define a custom cutoff probability which allows users to use tools confidently. I also recommend tools that involve minimum user work and maximum transparency.

ML tools have practical value, especially in a dual reviewer screening. Gates et al. (2020) showed that ML tools can safely be used to assist one of two screeners, with virtually no effect on the studies included in the review. The usability of ML tools is also a key factor to their usefulness because workload savings are often counteracted by long setup time and technical difficulties. ML tools must also work together with existing tools so that researchers can easily adopt them into their workflow. No researcher will change their entire method to adopt a single new tool. They will only use it if they can continue with their normal workflow. Most researchers agree that the potential saving of time and resources by ML tools for SR would be wonderful but are skeptical the desired work savings can be achieved without sacrificing more important things. When there is a cohesive set of software that integrates together to help users perform literature reviews from start to finish without reducing quality, then reviewers will be willing to use that software. Until then it is the job of early adopters and developers to communicate in a circular way to develop fully functional tools.

Conclusion

Although machine learning (ML) tools for systematic review (SR) have great potential they are clearly still in development. Most of the ML tools for SR discussed in this paper, especially screening tools, achieve significant theoretical work savings. However, it is not clear if those theoretical works savings translate into real work savings in a practical setting. To achieve practical value from ML in SR, further research should develop more rigorous stopping criteria and formally demonstrate the validity of ML methods in SR. Developers of ML tools for SR should focus on making their tools user friendly, feature rich, and flexible. ML methods can be used to achieve work savings in SR without compromising scientific rigor, but it will take significant work before they reach widespread acceptance.

References

Arno, A., Elliott, J., Wallace, B., Turner, T., & Thomas, J. (2021). The views of health guideline developers on the use of automation in health evidence synthesis. *Syst Rev*, *10*(1), 16. https://doi.org/10.1186/s13643-020-01569-2

Beller, E., Clark, J., Tsafnat, G., Adams, C., Diehl, H., Lund, H., Ouzzani, M., Thayer, K., Thomas, J., Turner, T., Xia, J., Robinson, K., Glasziou, P., Ahtirschi, O., Christensen, R., Elliott, J., Graziosi, S., Kuiper, J., Moustgaard, R., O'Connor, A., Riis, J., Soares-Weiser, K., Vergara, C., & Wedel-Heinen, I. (2018). Making progress with the automation of systematic reviews: Principles of the International Collaboration for the Automation of Systematic Reviews (ICASR). *Systematic Reviews*, *7*(1). https://doi.org/10.1186/s13643-018-0740-7

Blake, C., & Lucic, A. (2015). Automatic endpoint detection to support the systematic review process. *J Biomed Inform*, *56*, 42-56. https://doi.org/10.1016/j.jbi.2015.05.004

Brockmeier, A. J., Ju, M., Przybyła, P., & Ananiadou, S. (2019). Improving reference prioritisation with PICO recognition. *BMC Med Inform Decis Mak*, *19*(1), 256. https://doi.org/10.1186/s12911-019-0992-8

Callaghan, M. W., & Müller-Hansen, F. (2020). Statistical stopping criteria for automated screening in systematic reviews. *Syst Rev*, *9*(1), 273. https://doi.org/10.1186/s13643-020-01521-4

Chai, K. E. K., Lines, R. L. J., Gucciardi, D. F., & Ng, L. (2021). Research Screener: a machine learning tool to semi-automate abstract screening for systematic reviews. *Syst Rev*, *10*(1), 93. https://doi.org/10.1186/s13643-021-01635-3

Gartlehner, G., Wagner, G., Lux, L., Affengruber, L., Dobrescu, A., Kaminski-Hartenthaler, A., & Viswanathan, M. (2019). Assessing the accuracy of machine-assisted abstract screening with DistillerAI: a user study. *Syst Rev*, *8*(1), 277. https://doi.org/10.1186/s13643-019-1221-3

Gates, A., Gates, M., Sebastianski, M., Guitard, S., Elliott, S. A., & Hartling, L. (2020). The semi-automation of title and abstract screening: a retrospective exploration of ways to leverage Abstrackr's relevance predictions in systematic and rapid reviews. *BMC Med Res Methodol*, *20*(1), 139. https://doi.org/10.1186/s12874-020-01031-w

Gates, A., Guitard, S., Pillay, J., Elliott, S. A., Dyson, M. P., Newton, A. S., & Hartling, L. (2019). Performance and usability of machine learning for screening in systematic reviews: a comparative evaluation of three tools. *Syst Rev*, *8*(1), 278. https://doi.org/10.1186/s13643-019-1222-2

Lau, J. (2019). Editorial: Systematic review automation thematic series. *Systematic Reviews*, *8*(1). https://doi.org/10.1186/s13643-019-0974-z

Marshall, I. J., & Wallace, B. C. (2019). Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Syst Rev*, *8*(1), 163. https://doi.org/10.1186/s13643-019-1074-9

O'Connor, A. M., Tsafnat, G., Thomas, J., Glasziou, P., Gilbert, S. B., & Hutton, B. (2019). A question of trust: can we build an evidence base to gain trust in systematic review automation technologies? *Syst Rev*, *8*(1), 143. https://doi.org/10.1186/s13643-019-1062-0

Olofsson, H., Brolund, A., Hellberg, C., Silverstein, R., Stenström, K., Österberg, M., & Dagerhamn, J. (2017). Can abstract screening workload be reduced using text mining? User experiences of the tool Rayyan. *Research Synthesis Methods*, *8*(3), 275-280. https://doi.org/10.1002/jrsm.1237

Tsou, A. Y., Treadwell, J. R., Erinoff, E., & Schoelles, K. (2020). Machine learning for screening prioritization in systematic reviews: comparative performance of Abstrackr and EPPI-Reviewer. *Syst Rev*, *9*(1), 73. https://doi.org/10.1186/s13643-020-01324-7

Westgate, M. J. (2019). revtools: An R package to support article screening for evidence synthesis. *Res Synth Methods*, *10*(4), 606-614. https://doi.org/10.1002/jrsm.1374