Using AI to help healthcare professionals stay up-to-date with medical research

Kim de Bie[1*] , **Nishant Kishore**[2*] , **Anthony Rentsch**[3*] , **Pablo Rosado**[4] and **Andrea Sipka**[4]
[1]University of Amsterdam
[2]Department of Epidemiology, Harvard T.H. Chan School of Public Health
[3]Institute for Applied Computational Science, Harvard University
[4]The Alan Turing Institute
[*]These authors contributed equally to this work
kimdebie@outlook.com, {nish.kishore, anthony.rentsch, pabloarosado, mala.lacerta}@gmail.com

Abstract

Staying up-to-date with current medical research can be a challenge for doctors and other medical decision-makers. Systematic reviews are one of the key tools that doctors use to stay informed. These are meta-analyses of all the relevant research with the intention of answering one specific question within the healthcare domain. Cochrane produces systematic reviews of medical research that are globally considered as a gold standard for high-quality healthcare information. However, because of the high volume of papers published and the fact that Cochrane's review authors are volunteers, it can take up to three years to write and publish one of these reviews. Our research focuses on speeding up this process. We propose a hybrid human-AI system to establish the topical area of a newly published paper faster, easing the process of searching for papers to include in a review.

1 Introduction

Doctors face challenging decisions about how to best care for their patients on a daily basis. While they leverage their clinical expertise, they must also keep their practice up-to-date with the latest standards developed through evidence-based medicine. Keeping track of research relevant to all aspects of one's clinical practice can easily become very time-consuming, and therefore medical professionals often rely on *systematic reviews*. A systematic review provides a robust synthesis of research evidence across healthcare studies [Higgins *et al.*, 2019]. It summarises the results of research related to a particular health issue and provides a high level of evidence on the effectiveness of healthcare interventions.[1]

Systematic reviews are based on an extensive search for relevant literature. This process is a *high recall search* task, as it is crucial that all high-quality studies related to the topic of interest are included in a review. Omitting an important finding directly impacts the recommended standards of care and therefore could lead to sub-optimal treatment provided to patients. The literature search process is largely conducted manually and therefore can take a very long time. As a result, the time required to perform a systematic review is often multiple years. This is problematic for both the authors of a review, for whom writing a review equals an investment of a large amount of (often volunteer) time, as well as for medical decision-makers, whom often require evidence within much shorter time frames [Higgins *et al.*, 2019].

The growing volume of medical research has made it increasingly difficult to produce systematic reviews that are timely and that are kept up-to-date with the latest findings [Bastian *et al.*, 2010]. Research has shown that while the conclusions of most reviews might be valid for five years, the findings of about a quarter might be out of date within two years, and 7% were outdated at the time of their publication [Shojania *et al.*, 2007]. Shortening the time it takes to create or update one of these reviews means that healthcare professionals can receive up-to-date information on state-of-the-art treatments faster, which improves the quality of their work.

Recently, in light of the COVID-19 pandemic, the United States' White House Office of Science and Technology put forward a call to action for AI experts to develop text-mining methods that would identify answers to pressing questions related to COVID-19 in tens of thousands of relevant scholarly articles.[2] Our work demonstrates the value of taking a hybrid human-AI approach to performing this type of search and offers insights into how to approach and evaluate such a system.

1.1 Cochrane

Cochrane is a not-for-profit organization that creates, publishes and maintains systematic reviews of health care interventions. Cochrane Reviews are considered the gold standard for high-quality healthcare information [Higgins *et al.*, 2019].

To manage the process of writing reviews, Cochrane is organized into 54 topical Review Groups (such as Stroke, Tobacco Addiction and Oral Health) which together maintain over 7,500 systematic reviews. Overseeing and executing this work are more than 37,000 contributors - many of whom are highly trained and skilled healthcare professionals, and most of whom are volunteers - who are located in 130 countries.

[1] Code for this project is accessible at https://github.com/alan-turing-institute/DSSG19-Cochrane-PUBLIC.

[2] https://www.whitehouse.gov/briefings-statements/call-action-tech-community-new-machine-readable-covid-19-dataset/

2 Problem description

To create a new or update an existing systematic review, researchers must submit a proposal to one of Cochrane's 54 Review Groups detailing a specific medical topic, such as the optimal treatment for a particular disease. Once approved, the researchers conduct a search for studies to be included in the review, with the understanding that every single relevant paper should be found. This is because missing an important paper could be costly, as it could contain a particularly relevant finding that would affect the treatment recommendations made in the review - and therefore what is seen as the 'gold standard' of treatment by doctors globally.

We focus our attention on the high-recall search task of looking for relevant studies. Currently this process is centered around performing keyword searches on several different medical research repositories (e.g. PubMed) and manually sorting the retrieved papers into relevant and irrelevant. However, this process is perceived as slow and leads to a large duplication of effort. Because of the high recall required, the keyword searches have to be very broad. This leads to a very high number of search results, which may be overlapping across different search processes, and many of the retrieved studies do not meet Cochrane's standards for inclusion. For example, only randomized control trials (RCTs) are typically included in a Cochrane review, but search results may unearth scores of observational studies that must be weeded out.

To improve the efficiency of the process and to prevent volunteers from repeatedly looking at the same low-quality studies, Cochrane intends to move to a multi-stage process [Higgins *et al.*, 2019]. Here, published papers are automatically pulled into the Cochrane database and regularly sorted into Review Group-specific registers. Under this system, when the time comes to look for relevant papers for a new systematic review, researchers will only have to look as far as their Review Group's register.

Despite the foreseen advantages and improvements in efficiency, the new process has the potential to create a new bottleneck. To maintain high-quality Review Group registers, large numbers of newly published research papers have to be manually reviewed for their relevance to the Review Group, their quality and their potential to ever be included in a new systematic review. This places a large burden upon the Cochrane members tasked with maintaining these databases. To alleviate this potential bottleneck, we developed a system that could help to prioritize which new papers to include in a Review Group's register.

3 Our approach

The aim of this project is to relieve Cochrane volunteers from some of the burden of manually reviewing very large amounts of research, most of which is eventually discarded as being irrelevant to their topic area. At the same time, the new system must maintain an extremely high recall ($> 95\%$), while keeping precision at an acceptable level (95% for most Review Groups). We consider this infeasible for an AI system operating in isolation. In addition, in a study about the attitudes of review authors towards automation, authors express that fully-automated decisions are not acceptable to them [Arno *et al.*, 2020]. Instead, we propose a hybrid human-AI system [Amershi *et al.*, 2019] in which the respective qualities of the human (i.e. the Cochrane volunteer) and AI are leveraged.

The task of the human-AI system is to perform multi-class, multi-label classification: for each paper, we aim to discover the appropriate Review Group label, where multiple labels are possible (i.e. a paper may belong to several groups at once). Review Groups operate independently and may have different wishes with regards to the specificity of their classification model. It is therefore desirable that each group maintains a large degree of control over their own part of the system. In addition, the nature of the data collected by each group and its quality may vary.

Taking these considerations into account, we use a one-vs.-all classification methodology: we construct 54 different classifiers, each of which predicts whether a given paper belongs in a Review Group. Eventually, we categorize papers into one of three categories: 'not relevant', 'relevant' and 'maybe relevant'. Papers for which the classifier exhibits a high level of confidence can be automatically kept (and classified as 'relevant'), and papers for which the classifier produces a very low probability can be automatically discarded ('not relevant'). At the same time, papers with intermediate levels of confidence are classified as 'maybe relevant' and can be further scrutinized by Cochrane volunteers.

3.1 Data and features

To train our classifiers, we rely on a database of roughly 1 million historical papers already ingested into the Cochrane system. Historically, Cochrane has aimed to collect all high-quality medical research into its database, regardless of whether papers were relevant to a particular Review Group. For each paper, we know which Review Group(s) have added the paper to their specialized register, if any. This allows us to construct positive and negative labels based on whether the paper was included in a Review Group's register. In addition, we have access to variables including the papers' title, authors, abstract and the name of the publication venue (journal or conference proceedings).

From the data, we construct numerous features. These include term-frequency inverse document frequency (TF-IDF) representations and word embeddings pre-trained on medical texts [Pyysalo *et al.*, 2013]. These are constructed from the titles, abstracts and/or venue names. We also experimented with features based on citation data, e.g. the Review Group labels of the cited works of a given paper.

3.2 Building a classification pipeline

The available data may differ greatly in size and quality across Review Groups. Therefore, it seems plausible that there is not a single machine learning model architecture that is superior across all Review Groups. Instead, we train a large set of models and select the best model for each Review Group. Thus, for example, one Review Group may end up with a Logistic Regressor using TF-IDF features, while another uses a Light-GBM architecture [Ke *et al.*, 2017] along with word embeddings. Note that the selection of the models happens automatically and is updated when the model is

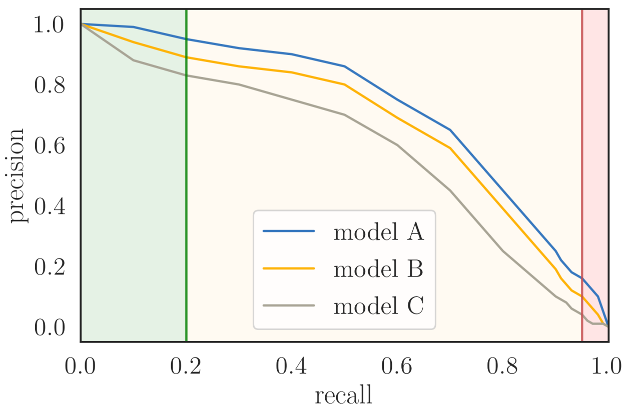


Figure 1: Example precision/recall plot. We compare three model architectures for a Review Group. Model A performs best. The papers with predicted probability scores in the red area (here with recall .95) can be automatically discarded, while those with predicted probabilities in the green area (precision .95) can be kept. The papers in the orange area must be manually inspected.

retrained (e.g. when new data becomes available). The flexibility of our pipeline, which allows each Review Group to have a completely independent classifier in terms of model architecture and features, leads to an overall increased system performance. The full pipeline procedure is outlined in Algorithm 1.

Algorithm 1 The model selection pipeline

```
Input: Models, Feature sets, Hyperparameters
Output: trained model for each RG
for RG $\in$ Review Groups do
    for model $\in$ Models do
        for feature set $\in$ Feature sets do
            for hyperparameter set $\in$ Hyperparameters do
                Train model with selected features/parameters.
                Calculate performance on test set.
            end for
        end for
    end for
    Select best model/feature/hyperparameter performance.
end for
return trained model for each RG
```

While demands for recall and precision are generally very high, Review Groups differ in their exact requirements and must be able to tweak their model's settings such that a specific precision/recall may be expected from the system. Therefore, we restrict our models to only include architectures that produce probabilities. To reiterate, we use a one-vs.-all architecture, where each Review Group has a separate model that predicts the probability of membership of a paper to that group. Review Groups can set the probability thresholds for 'not relevant' and 'relevant' to their desired level, so that their model's performance can be expected to match group-specific demands. See Figure 1 for a visual illustration.

The models used in our pipeline include Logistic Regression, ElasticNet [Zou and Hastie, 2005], Random Forest [Breiman, 2001], AdaBoost [Freund and Schapire, 1995], XGBoost [Chen and Guestrin, 2016] and LightGBM [Ke *et al.*, 2017]. Although the ensemble methods (Random Forest to LightGBM) returned superior results in some cases, it must be taken into account that these methods generally require significantly more training time and heavier computational resources. Decisions about model selection should consider whether the performance gap is large enough to justify a potential extra investment in resources, which may be of extra concern for not-for-profit organizations such as Cochrane. For implementation details of the models, we refer to our GitHub repository.[1]

4 Experimental setup and results

As outlined in Algorithm 1, we test dozens of model configurations - combinations of algorithms, features, and hyperparameter sets - for each Review Group. For each configuration, we measure performance as the *precision at various levels of recall*. We know that a recall of $> 95\%$ is required by most Review Groups, meaning that at least $> 95\%$ of all relevant papers must be retrieved. In addition, most Groups require a precision of 95%, meaning that having up to 5% of irrelevant (and thus wrongly-labeled) papers in the repository is acceptable.

Given these constraints, we aim to minimize the time that Review Groups spend on manually categorizing papers. As stated, we classify papers into 'not relevant', 'relevant' and 'maybe relevant', where only the 'maybe relevant' papers are manually scrutinized by Cochrane members. Therefore, the best-performing architecture for a Review Group is one that renders the 'maybe relevant' group as small as possible given the constraints on precision and recall, and this is what we optimize for in our experimental setup.

As shown in Figure 1, we construct a precision-recall curve for each Review Group. On this curve, we select the probability thresholds that correspond to the Group's desired precision and recall. We use these thresholds to classify papers into the three categories. We then select the architecture that leads to the smallest proportion of papers classified as 'maybe relevant', i.e. the largest reduction in manual scrutinizing of papers. Note that we test performance using a five-fold cross-validation set-up, where precision at a given recall level is averaged across five held-out test folds. Allowing each Review Group to set these thresholds independently allows Groups to specify their tolerance for risk versus their desire to reduce their workload.

4.1 Results: Reducing Review Group workload

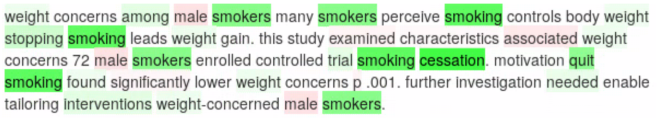
Based on our current results, we estimate that our approach would substantially decrease the workload of keeping Review Group registers up-to-date under Cochrane's new system. While there is some heterogeneity across Review Groups, we find that on average, $\mathbf{77}\%$ of papers can be automatically discarded (and are assigned the label 'not relevant') and $\mathbf{1}\%$ can be automatically kept (and are labelled as 'relevant'). This means that for the average Review Group, only $\mathbf{22}\%$ of all papers should undergo further review (as they are labelled as 'maybe relevant'), even when we specify that the the expected recall for each Review Group should be 95% and precision should be 95%.

We find that the optimal model architecture differs strongly across the 54 Review Groups with no architecture standing out as consistently better than another. In addition, as new papers are continuously added, the system may be retrained regularly and different settings may become optimal after retraining. In this way, we foresee that the model architectures across Review Groups may change substantially over time.

It is practically impossible to compare these results to the current practice of executing keyword searches in various repositories for each systematic review independently. However, these results suggest that using the proposed architecture could result in a reduction in the time it takes to keep a specialized register up-to-date with new research studies. This is crucial for the feasibility of Cochrane's envisioned new process, which is intended to speed up the process of writing systematic reviews significantly.

4.2 The importance of interpretability

The system we propose explicitly combines the strengths of human volunteers and AI in a hybrid system. To further simplify the task of manually reviewing those papers for which the AI cannot make a conclusive categorization, we see a large role for interpretability methods. Specifically, we implement LIME [Ribeiro *et al.*, 2016] to explain the predictions of our classifier[3]. Figure 2 shows an example of a title and abstract for a paper relevant to the Tobacco Review Group. LIME highlights the words that contributed most positively (in green) and most negatively (in red) to the prediction of the classifier. This draws the volunteer's attention to the most important parts of the text, and may therefore help her reach a decision faster. Also, it confirms the validity of our models: we consistently identify substantively relevant words as important predictors.



weight concerns among male smokers many smokers perceive smoking controls body weight stopping smoking leads weight gain. this study examined characteristics associated weight concerns 72 male smokers enrolled controlled trial smoking cessation. motivation quit smoking found significantly lower weight concerns p .001. further investigation needed enable tailoring interventions weight-concerned male smokers.

Figure 2: Example of a title/abstract that was labeled as belonging to the Tobacco Review Group. Words highlighted in green contributed positively to our model's prediction that this paper belongs to the Tobacco group, while words highlighted in red had a negative contribution. Stopwords have been removed from the text and all words have been converted to lowercase.

5 Discussion

Beyond its direct relevance to the creation of systematic reviews by Cochrane, this project harbors more general learnings about creating AI solutions in a 'social-good' context. Firstly, while Cochrane, like many not-for-profit organizations, recognizes the potential benefit of AI to its work, fully-automated black-box solutions are simply not acceptable to its domain experts. To guarantee that important studies are not omitted from a systematic review, Review Groups require to maintain a very high degree of control in the literature selection process. As such, we found that the potential strengths of AI in this context could only be leveraged if the designed solution left ample space for domain experts to manage the system's settings (by altering the thresholds for manual review) and to inspect its functioning (through explanations).

Secondly, over the course of our conversations with Cochrane, it became increasingly clear that our proposed solution would only ever be adopted (and therefore only ever provide real added value) if it followed existing organizational structures. In our case, the diversity in demands across the different Review Groups meant that we could not simply implement a fully automated blanket solution, but that our solution had to be able to adapt to these different demands. This shows that successful design and implementation requires that time is spent by the developers of AI systems to get to know the organizational context in which their solution will be embedded. Lastly (and relatedly), in designing hybrid human-AI systems, the heterogeneity of the human component of the system must be taken into account. It is easy to perceive the 'human' in human-AI as a single entity, but it is important to recognize that the humans any system will interact with (or in our case, the members of 54 Review Groups) are likely to all have their individual requirements from the system.

6 Conclusion

We have presented a hybrid human-AI system that classifies medical research papers into one or more Review Groups, which can help to speed up the process of writing a systematic review by Cochrane. While a few steps removed from clinical practice itself, Cochrane's systematic reviews form a cornerstone in the creation and maintenance of standards of care by doctors and health policymakers globally. The current COVID-19 pandemic has made the importance of having access to up-to-date healthcare research all too clear, and has also shown how difficult this can be when the volume of research is very large. This project shows the potential for using AI as a tool to help tackle this challenge.

Acknowledgments

This project was executed during the Data Science for Social Good Summer Fellowship 2019 at the University of Warwick and The Alan Turing Institute, hosted in partnership with the University of Chicago and the DSSG Foundation. We would like to thank all those involved with the fellowship for their invaluable contributions to this project. We would also like to thank our collaborators at Cochrane, in particular James Thomas, for their input.

References

[Amershi *et al.*, 2019] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

[3] We used the ELI5 implementation: https://eli5.readthedocs.io

[Arno *et al.*, 2020] Anneliese Downey Arno, Julian Elliott, Byron Wallace, Tari Turner, and James Thomas. The views of health guideline developers on the use of automation in health evidence synthesis. preprint. *Research Square*, 2020.

[Bastian *et al.*, 2010] Hilda Bastian, Paul Glasziou, and Iain Chalmers. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS medicine*, 7(9), 2010.

[Breiman, 2001] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[Chen and Guestrin, 2016] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[Freund and Schapire, 1995] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995.

[Higgins *et al.*, 2019] JPT Higgins, J Thomas, J Chandler, M Cumpston, T Li, MJ Page, and VA Welch (editors). *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons, 2019.

[Ke *et al.*, 2017] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, pages 3146–3154, 2017.

[Pyysalo *et al.*, 2013] Sampo Pyysalo, Filip Ginter, Hans Moen, Tapio Salakoski, and Sophia Ananiadou. Distributional semantics resources for biomedical text processing. *Proceedings of LBM*, pages 39–44, 2013.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ”why should i trust you?” explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[Shojania *et al.*, 2007] Kaveh G Shojania, Margaret Sampson, Mohammed T Ansari, Jun Ji, Steve Doucette, and David Moher. How quickly do systematic reviews go out of date? a survival analysis. *Annals of internal medicine*, 147(4):224–233, 2007.

[Zou and Hastie, 2005] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.