ELSEVIER

Journal of Clinical Epidemiology 110 (2019) 42−49

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

The risk of conclusion change in systematic review updates can be estimated by learning from a database of published examples

Rabia Bashir[a,*], Didi Surian[a], Adam G. Dunn[a,b]

[a]Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, Sydney, New South Wales 2109, Australia
[b]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA 02115, USA

Accepted 26 February 2019; Published online 5 March 2019

Abstract

Objectives: To determine which systematic review characteristics are needed to estimate the risk of conclusion change in systematic review updates.

Study Design and Setting: We applied classification trees (a machine learning method) to model the risk of conclusion change in systematic review updates, using pairs of systematic reviews and their updates as samples. The classifiers were constructed using a set of features extracted from systematic reviews and the relevant trials added in published updates. Model performance was measured by recall, precision, and area under the receiver operating characteristic curve (AUC).

Results: We identified 63 pairs of systematic reviews and updates, of which 20 (32%) exhibited a change in conclusion in their updates. A classifier using information about new trials exhibited the highest performance (AUC: 0.71; recall: 0.75; precision: 0.43) compared to a classifier that used fewer features (AUC: 0.65; recall: 0.75; precision: 0.39).

Conclusion: When estimating the risk of conclusion change in systematic review updates, information about the sizes of trials that will be added in an update are most useful. Future tools aimed at signaling conclusion change risks would benefit from complementary tools that automate screening of relevant trials.

Keywords: Machine learning; Classification trees; Automation of systematic reviews; Systematic reviews as topic; Clinical trial registries; Updating systematic reviews

1. Introduction

Systematic reviews are used to provide a comprehensive synthesis of clinical evidence to guide clinical decision-making, form the basis for clinical practice guidelines, and suggest directions for new research [1,2]. To fulfill that purpose, systematic reviews need to be kept up to date [3]. As a consequence of the resource-intensive processes involved in producing a systematic review, it can be a challenge to keep up with the rate at which evidence from new trials is made available [4]. At the same time, a substantial proportion of the systematic reviews being published are redundant, unnecessary, or focused away from the clinical questions where accumulating evidence could influence the conclusions in ways that would influence clinical practice [5].

An approach for improving the efficiency of systematic reviews is to make them easier to do, by individually automating the underlying processes—searching, screening, information extraction, and synthesis [6−15]. However, these tools alone are unlikely to help avoid unnecessary or redundant systematic reviews. According to a recent study [16], screening accounts for 25% of the effort required to produce a systematic review, which means that tools for avoiding unnecessary systematic reviews could save a majority of the time costs of undertaking systematic reviews. General guidance on when to update a systematic review considers not only the accumulation of new evidence but also contextual factors like the importance of the topic of the review and the potential impact on guidelines and clinical decision-making. However, there is no evidence that review updates are undertaken faster when new evidence is made available [17]. Similarly, relatively little work has been done to develop statistical methods to quantify the potential for new evidence to change the results of a review, and

Conflicts of interest: None.

Funding: R.B. is supported by a Macquarie University Postgraduate Scholarship.

* Corresponding author.

E-mail address: rabia.bashir@students.mq.edu.au (R. Bashir).

https://doi.org/10.1016/j.jclinepi.2019.02.015
0895-4356/© 2019 Elsevier Inc. All rights reserved.

What is new?

Key findings

- When modeling the risk of a conclusion change in systematic review updates, information about the sizes of new relevant trials was most useful than information extracted from original review.
- Identifying and extracting data from systematic review updates for use as training data is challenging.

What this adds to what was known?

- Existing methods for predicting whether a systematic review conclusion would change were based on measures extracted from primary meta-analyses.
- The risk of conclusion change in a more general set of systematic review updates was modeled in a database of examples of systematic reviews paired with their updates, using features that were relatively simple to extract.

What is the implication, what should change now?

- Future tools aimed at estimating when a systematic review is at risk of a change in conclusion would benefit from being coupled with tools that automate trial screening because information about new relevant trials was found to be most useful for estimating risk.

these methods are mostly confined to examining meta-analyses [4,18−21]. Some tools and checklists that are intended to support the decision to update a systematic review use the availability of new evidence as an input [22,23].

One tool used to support the decision to update a systematic review was developed to take advantage of information about what happened in a database of previously updated systematic reviews to predict whether a primary meta-analysis would change given the accumulation of new evidence [24]. This form of empirically derived tool has the potential to improve the efficiency of systematic review efforts but there are several practical limitations. A focus on meta-analyses means that the tool would be less useful in systematic reviews with no meta-analyses or where new outcomes were added, which may be of particular importance for interventions where new safety issues arise. The approach makes use of continual surveillance of new and relevant studies to determine how much of the currently available evidence is covered by the original meta-analysis. This can be time-consuming, so it would be useful to know if this information is necessary for estimating whether a systematic review conclusion is likely to change. Because systematic reviews are rarely made available in structured and machine-readable formats that would make them amenable to data mining, the tool was limited to learning from examples in one journal where data extraction could be standardized.

Our aim was to determine which systematic review characteristics are useful for estimating the risk that a systematic review would change its conclusion if updated to include new studies. To do this, we extracted information from a set of systematic reviews paired with their published updates to model the risk of a change in conclusion and examined how those features might be operationalized to create a risk-signaling tool.

2. Methods

2.1. Study data

We searched PubMed for systematic reviews published before December 2017 with the aim of identifying pairs of systematic reviews, starting with the most recent systematic review updates and finding their most recent previous version. To identify updates, we limited the search to articles that included the terms "systematic review" and "update" in the titles or abstracts. We then read the abstract to exclude any article that was not a systematic review and used information in the abstract or background to determine whether the systematic review referred to a previous version.

From pairs of identified systematic reviews and updates, we excluded any that were not written in English and any pairs for which either review had a published erratum or were withdrawn. We then excluded pairs where there was a major change in the clinical question answered or where we were unable to extract a minimum set of features. This included updates that had substantially changed the inclusion criteria, included only observational studies, added no new evidence, or did not clearly state the number of participants in the set of included studies. We additionally excluded systematic reviews that added no new studies in the update because they do not exhibit changes in conclusion and are not useful in the models.

2.2. Main outcomes and measures

Classification trees are used in machine learning to model a categorical outcome feature from several input features, producing a decision tree. We chose to use classification trees to model the risk of a change in conclusion because of their simplicity and interpretability—the contribution of the input features to the decision is clear and the tree can be implemented for use in practice more easily than noninterpretable models. For input, we characterized each pair of systematic reviews and updates by four features. The time elapsed because the search date was defined by the number of days between the search date of the systematic review and the search date in the update. We extracted the number of trials and participants from information available in the systematic review. For already published systematic review updates, a relevant trial is one that has already been evaluated by the authors of the systematic review update and included. The coverage score was defined by the total number of participants in the trials included in the systematic review as a proportion of the total number of participants in the trials included in the update. The coverage score is similar to the inverse of participant ratio used by Takwoingi et al. [24].

The primary outcome used in the construction of the classification trees was the presence of a change in conclusion. To determine which of the updates exhibited a conclusion change, two authors (R.B. and A.G.D.) read the systematic reviews and their updates to determine whether there was a change in conclusion, using information extracted from the results and conclusion statements. The level of agreement between the evaluations was measured using Cohen's kappa, and disagreements were resolved by discussion.

To ensure feasibility and to capture a sample of systematic reviews that was reasonably balanced in terms of the main outcome, we only included systematic reviews if their updates added new studies. Systematic reviews that do not include new studies can change conclusions, but we expected that this would be rare. The exclusion of systematic reviews without updates and the under-representation of reviews for which no new evidence was found may introduce biases into the models. A fairer sampling approach would have been to sample from across a general set of systematic reviews (regardless of whether they have an update) and manually search and screen for the availability of new evidence. However, this would severely limit the number of examples that could be used to construct the classifiers.

2.3. Classification tree construction, analysis, and evaluation

We built three classification trees using sets of features extracted from systematic reviews and their updates, to model the risk of conclusion change in systematic review updates. Our rationale for each of the three classifiers was based on how we expected to use the decision tree as a tool, where the classifiers represented differences in how quickly we could estimate the risk of a conclusion change in updates given the amount of time and effort involved to extract the information needed to apply the tool. The first classifier uses all four features: the time elapsed since the search date, the number of trials in the systematic review, the number of participants in the systematic review, and the coverage score based on information from new and relevant trials. This is resource intensive—users need to extract multiple types of information from the review and identify and extract information about the new and relevant trials not included in the review. Although there are a range of methods used to support searching and screening of published articles and trial registrations [14,15,25,26], this still requires manual effort by experts. The second classifier excludes the coverage score and uses the time elapsed since the search date and the number of trials and number of participants in the original systematic review. This reduces the amount of manual effort required to predict the risk of a conclusion change by limiting what is needed to only include information available in the systematic review. The third classifier uses only the time elapsed since the search date and excludes information about the number of trials and participants as well as the coverage score. Early guidelines about how often a systematic review should be updated were based on this feature (often 2 years was given as a reasonable time), and this third classifier represents an empirically derived version of these guidelines. The information is trivial to extract from systematic reviews, making the classifier amenable to automation.

For each of the four sets of features, we used the classification tree method to model the change in conclusion using all pairs of systematic reviews and their updates. The approach produces an interpretable set of rules that splits the set of systematic reviews into increasingly smaller groups comprising mostly changed or mostly unchanged conclusions. The ability to discern systematic reviews with changed conclusions from those with unchanged conclusions might then be useful for signaling when a systematic review has features that are most similar to others that previously exhibited a conclusion change when they were updated. The assumption is that we can learn the features that predict conclusion changes in systematic reviews from a large general database of already published systematic review updates.

To compare the performance of three models, we calculated the precision, recall, and F_1-score of the three classifiers. The precision was defined as the number of correctly identified systematic reviews with conclusion change divided by the total number of reviews, and recall was the number of correctly identified systematic reviews with conclusion change divided by the total number of reviews with conclusion change. The F_1-score is the harmonic mean of precision and recall, which is used as a more robust measure of accuracy in cases where data are unbalanced in terms of positive and negative results. The model produces a risk estimate between 0 and 1 for each systematic review, so to classify the systematic reviews as high-risk or low-risk, we selected the threshold that maximizes the F_1-score across the set of reviews. We also produced receiver operating characteristic (ROC) curves and calculated the area under the ROC (AUC) to compare the performance of the three classifiers. We conducted all experiments using Python 3.6.

2.4. Practical demonstration

To demonstrate how the classification trees might work in practice to signal when a systematic review is at risk of a conclusion change, we selected two additional systematic reviews that included randomized controlled trials and had updates published in 2018. We then applied the tool retrospectively to examine how the estimated risk of a conclusion change varied as new evidence was published over time. To do this, we extracted information about the accumulation of new evidence by identifying the set of trials added to the updates and reconciled information about their timing and relevant characteristics. For each trial, we extracted the number of participants and the date when the results were first publicly reported in a published article, on ClinicalTrials.gov, or on company websites. We used the earliest date if the results of the trial were reported in more than one location and used the number of participants from the published article if the number of participants varied from the registration to the trial report.

To determine how the estimated risk of a conclusion change varied over time, we applied the three classification trees to calculate the risk for each day in the period between the publication date of the systematic review and the search date of the update. The number of trials and number of participants in the review were constant, the time elapsed since the search date increased each day, and the coverage score decreased to align with the public reporting of results for each of the new and relevant trials.

3. Results

We screened 1,047 records returned by the search. Of these, we excluded 207 by screening the titles and abstracts to remove articles that were not systematic reviews. We excluded 656 because they were not updates of previously published systematic reviews. Of the 184 that were updates of systematic reviews, 63 met our inclusion criteria (Fig. 1).

Of the 63 systematic review and update pairs included in the models, 40 (63%) were published in the Cochrane Database of Systematic Reviews. After two investigators independently reviewed the conclusions (Cohen's kappa 0.40) of 63, we found that 20 (32%) exhibited a conclusion change. Across the 63 systematic review updates, the two investigators agreed that there was a conclusion change on 13, agreed that there was no conclusion change on 37, and disagreements in each direction were 6 and 7, respectively.

3.1. Classification tree performance

The first classifier that used all features including the coverage score produced the highest performance: recall 0.75 and precision 0.43 at a threshold of 0.45, the value that maximized the F_1-score (Table 1). The CART based on all features illustrates that changes in conclusion were more common in scenarios where the coverage score was lower, where the systematic review included a smaller number of participants and fewer trials, and where more time had elapsed since the search date of the review (Fig. 2). The second classifier used only features extracted from the systematic review, which reduced the performance from the full classifier: recall 0.75 and precision 0.39 at a threshold

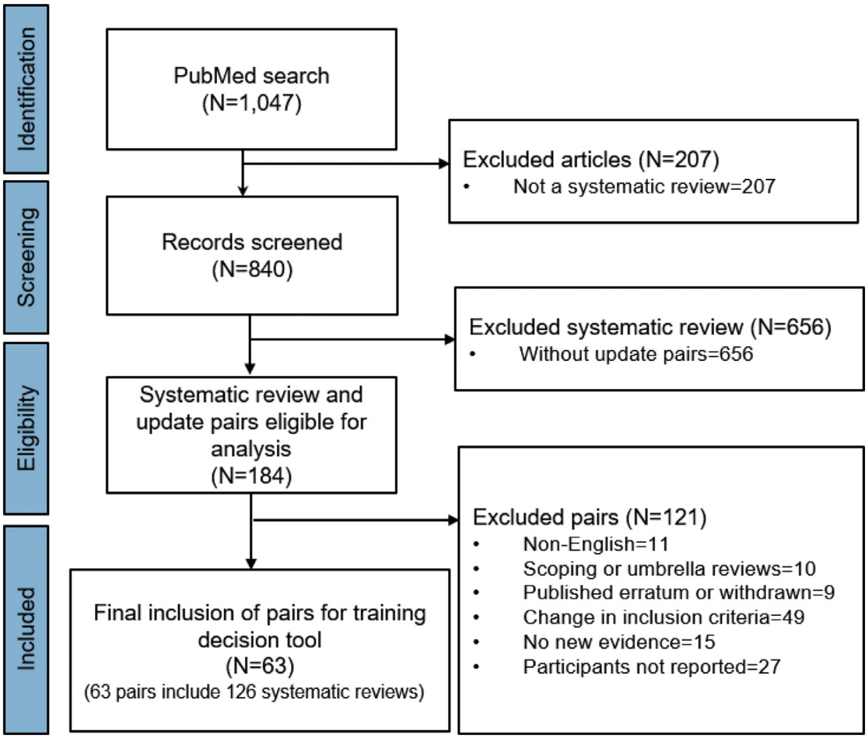
Identification
Screening
Eligibility
Included
PubMed search (N=1,047)
Excluded articles (N=207)
• Not a systematic review=207
Records screened (N=840)
Excluded systematic review (N=656)
• Without update pairs=656
Systematic review and update pairs eligible for analysis (N=184)
Excluded pairs (N=121)
• Non-English=11
• Scoping or umbrella reviews=10
• Published erratum or withdrawn=9
• Change in inclusion criteria=49
• No new evidence=15
• Participants not reported=27
Final inclusion of pairs for training decision tool (N=63)
(63 pairs include 126 systematic reviews)

Fig. 1. PRISMA flow diagram of study selection for a search and screening process that resulted in the inclusion of 63 systematic review and update pairs for constructing the classification tool.

Table 1. Performance of the three classifiers by precision, recall, and area under the receiver operating characteristic curve (AUC)

Classifier	Precision	Recall	AUC
1: Full, including coverage score	0.43	0.75	0.71
2: Partial, including included trial details	0.39	0.75	0.64
3: Partial, including only search date details	0.39	0.75	0.61

of 0.45. The third classifier used only the time elapsed since the search date and exhibited the lowest performance: recall 0.75 and precision 0.39 at a threshold of 0.45. The area under receiver operating characteristic curve (AUC) is determined independent of the choice of threshold and was highest for the first classifier (AUC: 0.71) compared to the second classifier (AUC: 0.64) and the third classifier (AUC: 0.61) (Fig. 3). The results show that information about the sizes of new and potentially relevant trials produced the largest positive impact on the ability to estimate the risk of a conclusion change.

3.2. Practical demonstration

We demonstrated how the models might be used to produce signals of conclusion change risks over time by retrospectively applying them to two systematic reviews. The systematic reviews used in the demonstration were new and were not included in the training data. The threshold value that maximized the F_1-score in the best-performing model was 0.45, so we used this as the value to signal when a systematic review was at higher risk of a conclusion change.

The first systematic review and its update examined evidence for physical therapy to reduce patient length of stay and did not exhibit a change in conclusion [27,28]. The original search date was May 2010 (it was published in September 2011), and the search date for the update was June 2017 (it was published in April 2018). When we used the first classifier (all features), the estimated risk increased to 0.44 after 1.8 years (Fig. 4a). The second classifier produced a signal 6.3 years after the systematic review publication date, estimating the risk at 0.52. The third classifier produced a signal after 4.2 years. A tool based on the first classifier would not have produced a strong signal of risk in conclusion change, and systematic reviewers using the tool may have chosen to delay the update of the systematic review.

The second systematic review was examining intravenous thrombolysis for acute ischemic stroke and exhibited a conclusion change [29,30]. The original search date for the second systematic review was June 2010 (it was published in November 2011), and the search date for the

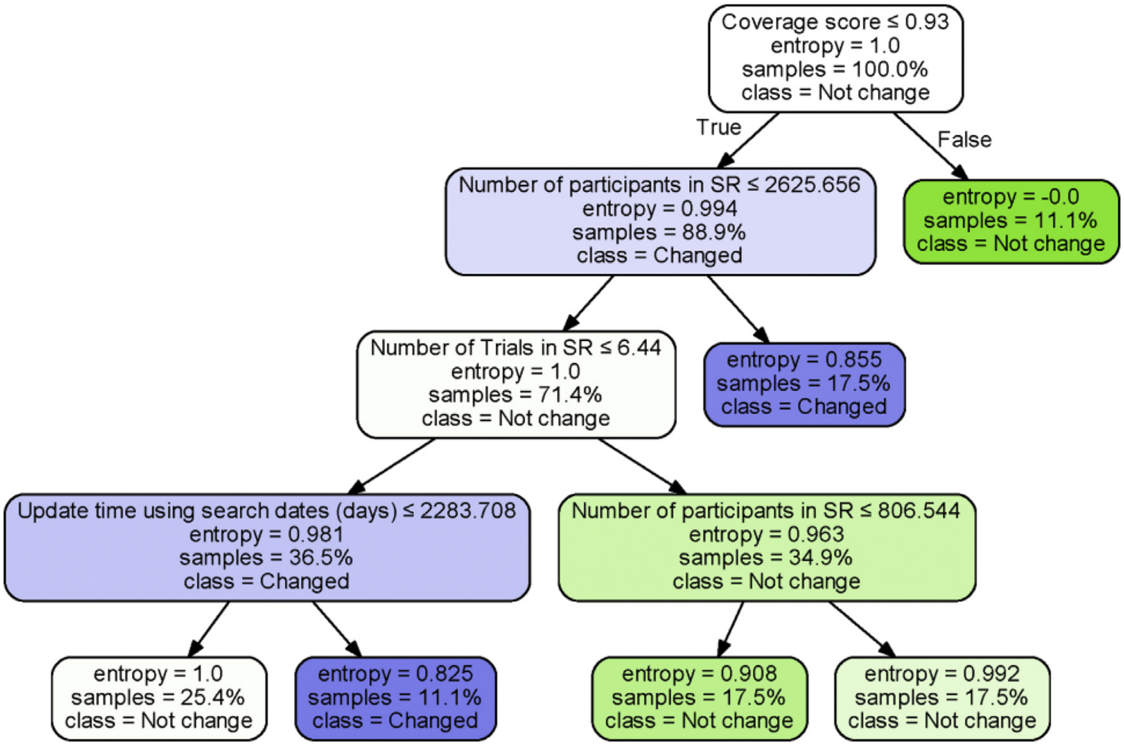

Fig. 2. The classification tree produced by using all features to estimate change in conclusion. Update time is given in days; entropy shows how uniform are all samples of a node; and samples represent the proportion of 63 systematic reviews and update pairs that appear within the node of the tree. The tree shows that conclusion changes were more common in systematic reviews that had more participants in the original review (>2,625) and a lower coverage score (≤0.93); or where there were fewer participants in the original review (≤2,625), fewer trials (≤6), and more time had elapsed since the search date (>2,283 days).

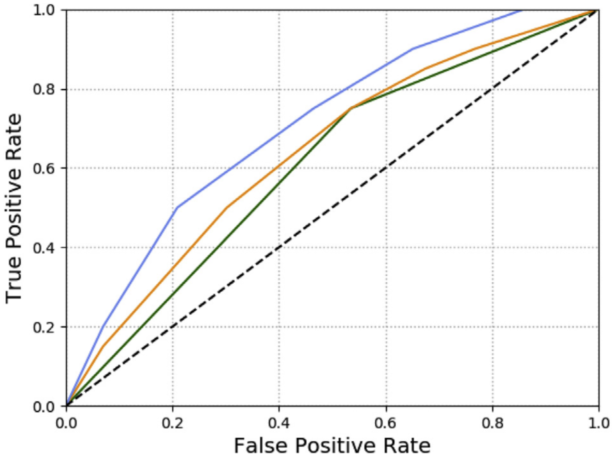

Fig. 3. Area under the receiver operating characteristic curve for classifiers using all features including the coverage score (blue, AUC: 0.71); only features extracted from the systematic review (orange, AUC: 0.64) and only using time elapsed since the search date (green, AUC: 0.61). The dotted line corresponds to the random classifier (AUC: 0.5). Higher AUC indicates better classifier performance. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

update was August 2016 (it was published in March 2018) [29,30]. The first classifier (all features) produced a signal 1.8 years after the systematic review was published, estimating the risk at 0.72 (Fig. 4b). The second classifier never produced a signal of a risk of conclusion change and reached a maximum estimated risk of 0.42 during the period of analysis. The third classifier produced the second signal after 4.2 years. If systematic reviewers were using the tool, they would have seen a signal of a risk in conclusion change 4.8 years earlier than the search was performed for the update.

4. Discussion

We found that accessing information about the presence and size of new and potentially relevant trials made the greatest improvement to our ability to estimate the risk of a conclusion change. The results showed that access to information that was more time-consuming to collect was the most useful in improving the classification tree's performance in estimating risk. As part of a broader consideration of factors that might influence the decision to update a systematic review, tools like the one we propose here could help systematic reviewers, journals, and funders avoid potentially unnecessary updates and focus on systematic reviews with conclusions that do not reflect currently available evidence. A tool that learns to estimate in advance whether a systematic review update would produce a change in conclusion could be constructed by learning from a large, general set of already published systematic review updates.

Previous studies have proposed or used different types of information to support the decision to update a systematic review [22,31]. Prospective evaluation of different approaches in this space is challenging because of the resource-intensive nature of undertaking systematic reviews. Takwoingi et al. [24] constructed a multicomponent tool for deciding whether to update a systematic review, using a set of nine features extracted from the primary meta-analyses of systematic reviews published in the Cochrane Database of Systematic Reviews. This tool was used as part of a process for prioritizing systematic review updates [23], but its application is limited to systematic reviews with meta-analyses of at least two trials and may only apply to systematic reviews published in that journal. Our approach differed in that we examined which features could be extracted to potentially predict the risk of conclusion change from a broader set of systematic reviews. The assumption is that with a broad and large set of published examples of systematic review updates, we would be able to determine which general characteristics would be most useful in a more general tool.

There are several implications to the work we presented here. The results suggest that it is feasible to develop tools that can be applied quickly to signal when a systematic review conclusion is no longer an accurate reflection of currently available evidence and focus resources on their update. Conversely, we might also be able to use these types of tools to avoid allocating funding and resources to systematic reviews that are unnecessary [5]. The results of the research here also indicate the value of knowing which trials are likely to be relevant to a systematic review in advance, beyond avoiding the time taken to search and screen for trials when undertaking a systematic review. The results suggest that a tool for quickly estimating the risk of conclusion change might rely on knowing in advance the set of trials that are most likely to be included in an update. This means that existing tools for automating or supporting trial screening could be coupled with tools for estimating risks of conclusion change. Automated surveillance of ongoing and completed trials relevant to a published systematic review are likely to improve with time [26,32–34], and these could be used to reduce or eliminate the need for screening and may eventually be used to automate signals for prioritizing systematic review updates.

Future work in this area would benefit from deeper integration with trial registries. Given that trial registrations represent an early indication that new trial evidence will become available, surveillance of relevant clinical trials could help us to estimate when a systematic review may be at risk of a conclusion change. Surveillance of ongoing and completed trials may make it possible to prepare in advance for systematic reviews that are likely to be at risk of a conclusion change as soon as new results are made available to the public. Although the structured, machine-readable, and connected reporting of trial results is improving [35], less effort has been spent on establishing public access to information connecting systematic reviews to their updates and the sets of studies they include. Improvements in structured reporting and transparency in this

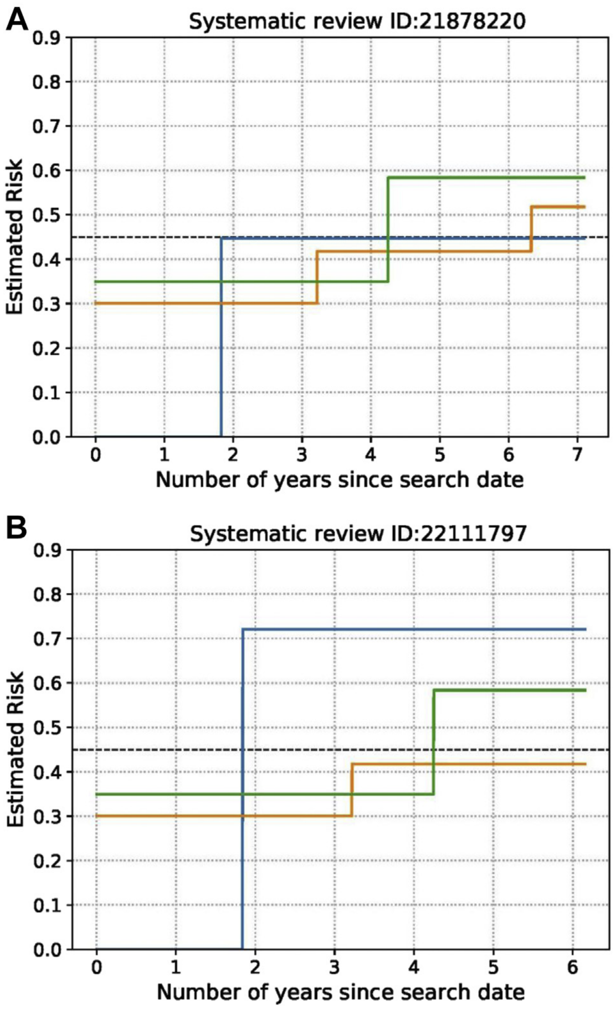

Fig. 4. Estimated risk for (A) a systematic review exhibiting no change in conclusion; and (B) a systematic review exhibiting a change in conclusion. We compare three classifiers: using all features (blue), using only features extracted from the systematic review (orange), and using only the time elapsed since the search date (green). A threshold at 0.45 (dotted line) is used to determine the presence of a signal. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

space would make it easier to track how evidence coverage degrades over time for published systematic reviews and may also help to reduce the number of redundant and unnecessary systematic reviews [36]. There are likely to be a range of other characteristics of systematic reviews that can be extracted and may be useful as features in predictive models of conclusion change. For example, information about the specialty or class of interventions might be indicative of the fragility of the conclusions in a particular area and could be integrated.

There were limitations to this study. First, we considered only English language systematic reviews that included clinical trials for our analysis, and the models may not generalize to other languages or systematic reviews of other study designs. Second, identifying systematic review updates that ask and answer the same clinical questions is a challenge, and our data set included a substantial number of systematic reviews from the Cochrane Database of Systematic Reviews, where updates are clearly demarcated. Third, we did not include systematic reviews that did not have updates because this would have required manual screening of new and relevant trials, which was not feasible. Fourth, the data set we used to construct the classification tree models was relatively small and we did not test the resulting models on unseen systematic review updates. Once a tool based on a larger data set has been constructed, a prospective evaluation of its ability to predict conclusion changes in advance of a systematic review update would be needed. Future work in the area would benefit from a structured database of systematic reviews with information about included trials [32].

5. Conclusion

We built three classifiers to determine which characteristics of systematic reviews are useful for estimating the risk of conclusion change in systematic review updates. The aim is to improve decisions about when to update systematic reviews, and we suggest that it may be useful to systematic reviewers who want to know in advance whether the conclusion of a review is likely to change if they incorporate newly available evidence. The tool is different from previous approaches because it uses a set of existing systematic review updates to learn how characteristics of systematic reviews and the trials that meet their inclusion criteria correspond to the risk of a conclusion change. The results show that access to information about the presence and size of new and potentially relevant trials is most useful for estimating risk. Given the potential value that these tools may have in improving the efficiency of systematic reviews, we think that further work building a database of systematic review updates from which to learn is warranted.

CRediT authorship contribution statement

Rabia Bashir: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing - original draft, Writing - review & editing. **Didi Surian:** Methodology, Software, Validation, Investigation, Formal analysis, Writing - review & editing, Visualization. **Adam G. Dunn:** Conceptualization, Methodology, Validation, Writing - review & editing, Formal analysis, Writing - original draft, Supervision.

References

[1] Cohen AM, Ambert K, McDonagh M. Studying the potential impact of automated document classification on scheduling a systematic review update. BMC Med Inform Decis Mak 2012;12:33.

[2] García LM, Pardo-Hernandez H, Superchi C, de Guzman EN, Ballesteros M, Roteta NI, et al. Methodological systematic review identifies major limitations in prioritisation processes for updating. J Clin Epidemiol 2017;86:11–24.

[3] Sutton D, Qureshi R, Martin J. Evidence reversal—When new evidence contradicts current claims: a systematic overview review. J Clin Epidemiol 2017;94:76–84.

[4] 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;147:224–33.

[5] Ioannidis J. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. Milbank Q 2016; 94:485–514.

[6] Beller E, Clark J, Tsafnat G, Adams C, Diehl H, Lund H, et al. Making progress with the automation of systematic reviews: principles of the International Collaboration for the automation of systematic reviews (ICASR). Syst Rev 2018;7:77.

[7] Boudin F, Nie J-Y, Bartlett JC, Grad R, Pluye P, Dawes M. Combining classifiers for robust PICO element detection. BMC Med Inform Decis Mak 2010;10:29.

[8] Choong MK, Galgani F, Dunn AG, Tsafnat G. Automatic evidence retrieval for systematic reviews. J Med Internet Res 2014;16:e223.

[9] Huang K-C, Chiang I-J, Xiao F, Liao C-C, Liu CC-H, Wong J-M. PICO element detection in medical text without metadata: are first sentences enough? J Biomed Inform 2013;46:940–6.

[10] Kim SN, Martinez D, Cavedon L, Yencken L. Automatic classification of sentences to support evidence based medicine. BMC Bioinformatics 2011;12(Suppl 2):S5.

[11] Miwa M, Thomas J, O'Mara-Eves A, Ananiadou S. Reducing systematic review workload through certainty-based screening. J Biomed Inform 2014;51:242–53.

[12] Shemilt I, Simon A, Hollands GJ, Marteau TM, Ogilvie D, O'Mara-Eves A, et al. Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. Res Synth Methods 2014;5:31–49.

[13] Tsafnat G, Glasziou P, Choong MK, Dunn A, Galgani F, Coiera E. Systematic review automation technologies. Syst Rev 2014;3:74.

[14] Wallace BC, Trikalinos TA, Lau J, Brodley C, Schmid CH. Semi-automated screening of biomedical citations for systematic reviews. BMC Bioinformatics 2010;11:55.

[15] 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;4:5.

[16] Bagheri E, Rios P, Pourmasoumi A, Robson RC, Hwee J, Isaranuwatchai W, et al. Improving the conduct of systematic reviews: a process mining perspective. J Clin Epidemiol 2018;103:101–11.

[17] Bashir R, Surian D, Dunn AG. Time-to-update of systematic reviews relative to the availability of new evidence. Syst Rev 2018;7:195.

[18] Barrowman NJ, Fang M, Sampson M, Moher D. Identifying null meta-analyses that are ripe for updating. BMC Med Res Methodol 2003;3:13.

[19] Sampson M, de Bruijn B, Urquhart C, Shojania K. Complementary approaches to searching MEDLINE may be sufficient for updating systematic reviews. J Clin Epidemiol 2016;78:108–15.

[20] Shekelle PG, Motala A, Johnsen B, Newberry SJ. Assessment of a method to detect signals for updating systematic reviews. Syst Rev 2014;3:13.

[21] Sutton AJ, Cooper NJ, Jones DR, Lambert PC, Thompson JR, Abrams KR. Evidence-based sample size calculations based upon updated meta-analysis. Stat Med 2007;26:2479–500.

[22] Sutton AJ, Donegan S, Takwoingi Y, Garner P, Gamble C, Donald A. An encouraging assessment of methods to inform priorities for updating systematic reviews. J Clin Epidemiol 2009;62:241–51.

[23] Welsh E, Stovold E, Karner C, Cates C. Cochrane Airways Group reviews were prioritized for updating using a pragmatic approach. J Clin Epidemiol 2015;68:341–6.

[24] Takwoingi Y, Hopewell S, Tovey D, Sutton AJ. A multicomponent decision tool for prioritising the updating of systematic reviews. BMJ 2013;347:f7191.

[25] Shao W, Adams CE, Cohen AM, Davis JM, McDonagh MS, Thakurta S, et al. Aggregator: a machine learning approach to identifying MEDLINE articles that derive from the same underlying clinical trial. Methods 2015;74:65–70.

[26] Surian D, Dunn AG, Orenstein L, Bashir R, Coiera E, Bourgeois FT. A shared latent space matrix factorisation method for recommending new trial evidence for systematic review updates. J Biomed Inform 2018;79:32–40.

[27] Peiris CL, Shields N, Brusco NK, Watts JJ, Taylor NF. Additional physical therapy services reduce length of stay and improve health outcomes in people with acute and subacute conditions: an updated systematic review and meta-analysis. Arch Phys Med Rehabil 2018;99:2299–312.

[28] Peiris CL, Taylor NF, Shields N. Extra physical therapy reduces patient length of stay and improves functional outcomes and quality of life in people with acute or subacute conditions: a systematic review. Arch Phys Med Rehabil 2011;92:1490–500.

[29] Sharma VK, Ng KW, Venketasubramanian N, Saqqur M, Teoh HL, Kaul S, et al. Current status of intravenous thrombolysis for acute ischemic stroke in Asia. Int J Stroke 2011;6:523–30.

[30] Wang X, You S, Sato S, Yang J, Carcel C, Zheng D, et al. Current status of intravenous tissue plasminogen activator dosage for acute ischaemic stroke: an updated systematic review. Stroke Vasc Neurol 2018;3:28–33.

[31] Langan D, Higgins JP, Gregory W, Sutton AJ. Graphical augmentations to the funnel plot assess the impact of additional evidence on a meta-analysis. J Clin Epidemiol 2012;65:511–9.

[32] Martin P, Surian D, Bashir R, Bourgeois FT, Dunn AG. Trial2rev: combining machine learning and crowd-sourcing to create a shared space for updating systematic reviews. JAMIA Open 2019, ooy062.

[33] Dalal SR, Shekelle PG, Hempel S, Newberry SJ, Motala A, Shetty KD. A pilot study using machine learning and domain knowledge to facilitate comparative effectiveness review updating. Med Decis Making 2013;33:343–55.

[34] Shekelle PG, Shetty K, Newberry S, Maglione M, Motala A. Machine learning versus standard techniques for updating searches for systematic reviews: a diagnostic accuracy study. Ann Intern Med 2017;167:213–5.

[35] Zarin DA, Tse T. Sharing individual participant data (IPD) within the context of the trial reporting system (TRS). PLoS Med 2016;13: e1001946.

[36] Bashir R, Dunn AG. Software engineering principles address current problems in the systematic review ecosystem. J Clin Epidemiol 2019; 109:136–41.