Testing a filtering strategy for systematic reviews: evaluating work savings and recall

Randi Proescholdt[1], Tzu-Kun Hsiao[1], Jodi Schneider[1], Aaron M. Cohen[2], Marian S. McDonagh[2], Neil R. Smalheiser [3]
[1]University of Illinois at Urbana-Champaign, Champaign, IL; [2]Oregon Health & Science University, Portland, OR; [3]University of Illinois at Chicago, Chicago, IL

Abstract

Systematic reviews are extremely time-consuming. The goal of this work is to assess work savings and recall for a publication type filtering strategy that uses the output of two machine learning models, Multi-Tagger and web RCT Tagger, applied retrospectively to 10 systematic reviews on drug effectiveness. Our filtering strategy resulted in mean work savings of 33.6% and recall of 98.3%. Of 363 articles finally included in any of the systematic reviews, 7 were filtered out by our strategy, but 1 "error" was actually an article using a publication type that the SR team had not pre-specified as relevant for inclusion. Our analysis suggests that automated publication type filtering can potentially provide substantial work savings with minimal loss of included articles. Publication type filtering should be personalized for each systematic review and might be combined with other filtering or ranking methods to provide additional work savings for manual triage.

Introduction

Systematic reviews (SRs) are extremely time-consuming; an average SR takes 67 weeks[1] and costs about $141,000[2] in staff time. A variety of machine learning approaches are being examined to assist SR teams, often focused on prioritizing the records retrieved[3,4] or reducing the need or extent of dual screening[5,6]. The time required is correlated with the number of records requiring manual triage of titles and abstracts for apparent relevance[7]. Hence, a key goal is to reduce the number of records that a SR team initially needs to examine while preserving recall, as close to 100% as possible. The goal of this work is to evaluate the potential of a particular strategy, using publication type and study design filters for automatic filtering of articles for contributing to automation of SRs.

In this paper, we tested our filtering strategy retrospectively against 10 previously completed SRs about comparative drug effectiveness. Our strategy uses two machine learning models, Multi-Tagger[8] and web RCT Tagger,[9] in combination with National Library of Medicine (NLM)'s MeSH terms and publication types in order to retain as many relevant articles as possible, while reducing the number of articles needing manual screening. The models have previously been evaluated using information retrieval measures, but need further evaluation in order to gain the trust of systematic reviewers[10] and to estimate the potential work savings in real-life situations. The 10 SRs used to evaluate this strategy came from the Drug Effectiveness Review Project (DERP). DERP is a collaboration of state Medicaid agencies that commission SRs aimed to help inform decisions about the drugs that would be available to Medicaid recipients in each state.

Methods

We included a series of SRs from 2003-2018 conducted for DERP by the Pacific Northwest Evidence-based Practice Center at Oregon Health & Science University (OHSU)[11]. For each of the 10 DERP reports used in this analysis, we received information, such as the study designs each review planned to include and reference libraries containing records for the citations screened, including decisions on inclusion in the SR. For each of the 10 SRs being studied, our evaluation calculated the **work savings** (i.e., the number of articles in the initial retrieval set that were filtered out by our strategy, divided by the number of articles in the initial retrieval set) and the **recall** (the number of articles finally included in the SR that passed by our filtering strategy, divided by the total number finally included in the SR). The filtering strategy is shown in Table 1. We retained abstracts if any of rules 1-3 applied. For rules 1 and 2, we checked automated publication type predictive scores from the Multi-Tagger[8] and web RCT Tagger[9] against designated thresholds which optimally balanced precision and recall (i.e., rather than optimizing recall alone, we chose the threshold which gave the highest F1; any article receiving a score below the threshold was filtered out, and any article equal to or above the threshold was retained). For rule 3, we retrieved NLM's MeSH indexing. If an included design was found in the MeSH terms or publication types, the article was retained.

Table 1: Abstract filtering rules. Items were retained if any of rules 1-3 applied, and filtered out otherwise.

Rule Number	Source of Publication Type or Study Design Information	Taggers and MeSH terms relevant to study designs over the 10 reviews	Condition
Rule 1	Multi-Tagger	Case-Control Studies Clinical Study Cohort Studies Meta-analysis Practice Guideline Prospective Studies Randomized Controlled Trial Retrospective Studies Systematic Review	Above threshold that gave the optimal F1[8]; or item not processed by Multi-Tagger (i.e., article not in English or lacking abstract)
Rule 2	Web RCT Tagger	Randomized Controlled Trial	Above 0.01 threshold for RCTs[9]; or item not processed by Web RCT Tagger (i.e., article not in English or lacking abstract)
Rule 3	MeSH Terms and Publication Types	Case-Control Studies Clinical Study Clinical Trial Cohort Studies Meta-Analysis Observational Study Practice Guideline Prospective Studies Randomized Controlled Trial Retrospective Studies Systematic Review	One or more relevant study design terms were applied in NLM MeSH indexing.

From each SR (summarized in Table 2), we used the following data: the list of study designs that the review stated as relevant for inclusion; PMIDs titles and abstracts screened manually (i.e., triage), PMIDs full-text screened, and PMIDs included in the final review. We customized the list of relevant study designs from Table 1 to each review, as shown in Table 2.

Not all of the study designs each SR listed as relevant for inclusion had direct matches to tags in Multi-Tagger. In such cases, we applied related tags we deemed likely to be relevant. Since there is no single Multi-Tagger score encompassing all comparative observational studies, we applied the following tags and MeSH terms for SRs that listed observational studies as relevant for inclusion: Cohort Studies, Case-Control Studies, Retrospective Studies, Prospective Studies, and Clinical Study. To ensure high recall, we also applied Observational Study as a MeSH term for all reviews that included observational studies and Clinical Trial as a MeSH term for reviews that included randomized controlled trials. Additionally, we applied Systematic Review, Meta-Analysis, and Practice Guideline as tags and MeSH terms for all 10 reviews, since the DERP team sought and reviewed the full-text of articles with these designs in order to help identify any articles missed by the original search.

Table 2: The 10 reviews from DERP, their included study designs, and the corresponding Multi-Tagger study designs.

#	SR Name	Study Designs Eligible for Inclusion in SR	Multi-Tagger Tags Applied
1	Anticoagulants-Original-Report	1. Head-to-head or active-controlled randomized trials 2. Systematic reviews 3. Cohort or case-control observational studies	-Case-Control Studies -Clinical Study -Cohort Studies -Meta-analysis -Practice Guideline -Prospective Studies -Randomized Controlled Trial -Retrospective Studies -Systematic Review
2	Asthma-COPD	1. Head-to-head randomized controlled clinical trials 2. Comparative systematic reviews 3. Comparative observational studies	-Case-Control Studies -Clinical Study -Cohort Studies -Meta-analysis -Practice Guideline -Prospective Studies -Randomized Controlled Trial -Retrospective Studies -Systematic Review
3	Benzodiazepines-Summary-Review	1. Systematic reviews	-Meta-analysis -Practice Guideline -Systematic Review
4	Hepatitis-C-Update-2	Best evidence available from: 1. Head-to-head randomized controlled trials 2. Observational studies 3. Systematic reviews 4. Other designs (e.g., pooled analyses)	-Case-Control Studies -Clinical Study -Cohort Studies -Meta-analysis -Practice Guideline -Prospective Studies -Randomized Controlled Trial -Retrospective Studies -Systematic Review
5	Long-Acting-Insulins	1. Head-to-head randomized controlled trials 2. Comparative observational studies 3. Systematic reviews	-Case-Control Studies -Clinical Study -Cohort Studies -Meta-analysis -Practice Guideline -Prospective Studies -Randomized Controlled Trial -Retrospective Studies -Systematic Review
6	Long-Acting-Opioids-Update-7	1. Head-to-head controlled clinical trials 2. Comparative systematic reviews 3. Comparative observational studies	-Case-Control Studies -Clinical Study -Cohort Studies -Meta-analysis -Practice Guideline -Prospective Studies -Randomized Controlled Trial -Retrospective Studies -Systematic Review
7	MS-Drugs-Update-3	1. Head-to-head controlled clinical trials 2. Placebo-controlled trials 3. Comparative observational studies 4. Comparative systematic reviews	-Case-Control Studies -Clinical Study -Cohort Studies -Meta-analysis

			-Practice Guideline -Prospective Studies -Randomized Controlled Trial -Retrospective Studies -Systematic Review
8	Newer-Diabetes-Meds-Update-2	1. Head-to-head randomized controlled trials 2. Head-to-head prospective cohort studies 3. Case-control studies	-Case-Control Studies -Clinical Study -Cohort Studies -Meta-analysis -Practice Guideline -Prospective Studies -Randomized Controlled Trial -Retrospective Studies -Systematic Review
9	PCSK9	1. Controlled clinical trials 2. Systematic reviews 3. Comparative observational studies	-Case-Control Studies -Clinical Study -Cohort Studies -Meta-analysis -Practice Guideline -Prospective Studies -Randomized Controlled Trial -Retrospective Studies -Systematic Review
10	Second-Generation-Antipsychotics-Update-5	1. Head-to-head randomized controlled trials 2. Placebo-controlled trials 3. Comparative systematic reviews 3. Comparative observational studies with a concurrent control group	-Case-Control Studies -Clinical Study -Cohort Studies -Meta-analysis -Practice Guideline -Prospective Studies -Randomized Controlled Trial -Retrospective Studies -Systematic Review

For each SR, we tabulated: a) the number of articles in the initial retrieval set (i.e., the actual search retrieved by the SR team in preparing their report); b) the number of articles filtered out using the strategy just described (i.e., the number of articles that the SR team actually screened but would not have if they had used our strategy); c) the percent work savings; d) the number of articles that DERP finally included in their final SR report (i.e., the number of articles actually included in the SR, based on our assumption that the SR team's actual results included the ideal set of articles); e) the number of finally included articles that were lost using the strategy just described; f) the percentage recall. These statistics are shown in Table 3. (In our analysis, we did not analyze whether our filtering strategy could have resulted in additional relevant articles for final inclusion in the SR reports.)

Our error analysis examined each article that was filtered out by our PT strategy but included in the final SR. We examined the model predictive scores and MeSH terms to understand why the article was filtered out, as well as an assessment of its publication type based on documentation in the DERP reference library, the article's metadata, and the article's full-text. We also assessed whether it met the SR's original inclusion criteria in terms of study design.

Results

Table 3. Summary statistics.

DERP Report	# in initial retrieval set	# filtered out by our strategy	% work savings	# of included articles	# of included articles removed by our strategy	% recall
Anticoagulants-Original-Report	1766	659	37.32	82	0	100
Asthma-COPD	1964	497	25.31	28	0	100
Benzodiazepines-Summary-Review	581	302	51.98	12	0	100
Hepatitis-C-Update-2	4917	1417	28.82	75	2	97.33
Long-Acting-Insulins	1086	301	27.72	37	1	97.3
Long-Acting-Opioids-Update-7	503	60	11.93	13	0	100
MS-Drugs-Update-3	1849	825	44.62	45	3	93.33
Newer-Diabetes-Meds-Update-2	1065	400	37.56	21	1	95.24
PCSK9	75	32	42.67	13	0	100
Second-Generation-Antipsychotics-Update-5	1110	314	28.29	37	0	100

Table 4. List of included articles filtered out.

DERP Report Name	PMID	Title	Actual study design	Reason filtered out	Error or exception?
Hepatitis-C-Update-2	16267758	Risk factors for perinatal transmission of hepatitis C virus (HCV) and the natural history of HCV infection acquired in infancy.	Cohort Study	Cohort studies predictive score below threshold	**Error**
Hepatitis-C-Update-2	22813094	Comparison of current US risk strategy to screen for hepatitis C virus with a hypothetical targeted birth cohort strategy.	Comparative study; birth cohort strategy	Cohort studies predictive score below threshold	**Error**
Long-Acting-Insulins	22966091	Does insulin glargine increase the risk of cancer compared with other basal insulins?: A French nationwide cohort study based on national administrative databases.	Cohort Study	Cohort studies predictive score below threshold	**Error**
MS-Drugs-Update-3	19936821	Parenthood and immunomodulation in patients with multiple sclerosis.	Cohort Study	Cohort studies predictive score below threshold	**Error**

MS-Drugs-Update-3	24131589	Prevalence of cutaneous adverse events associated with long-term disease-modifying therapy and their impact on health-related quality of life in patients with multiple sclerosis: a cross-sectional study.	Cross-Sectional Study	Cross-Sectional Study tagger not used	**Exception**
MS-Drugs-Update-3	24463630	Pregnancy outcomes in the clinical development program of fingolimod in multiple sclerosis.	Clinical Study	Clinical Study predictive score below threshold	**Error**
Newer-Diabetes-Meds-Update-2	25300980	Drug utilization, safety, and effectiveness of exenatide, sitagliptin, and vildagliptin for type 2 diabetes in the real world: data from the Italian AIFA Anti-diabetics Monitoring Registry.	Retrospective cohort study	Cohort studies predictive score below threshold	**Error**

Acting on these initial assumptions, we found that our filtering strategy resulted in work savings ranging from 11.9% to 52.0% (mean 33.6%) and recall of 93.3% to 100% (mean 98.3%). We examined the 7 articles finally included in any of the SRs but filtered out by our strategy, and found that 1 of these "errors" was actually an article whose publication type that the SR team had not pre-specified as relevant for inclusion (Table 6). Because this exclusion is not a true error, we recalculated recall by dividing the number of true errors in each SR by the number of total included articles. The recalculated recall of our filtering strategy ranges from 95.24% to 100% (mean 98.5%). Five of the remaining six errors occurred because the score for one particular article type, Cohort Studies, was below our chosen threshold. Had we adjusted the threshold for Cohort Studies down to 0.02, we would have achieved slightly better recall (i.e., range of 93.3% to 100%; mean 99.1%), with minimal loss of work savings (range of 11.1% to 52.0%; mean 31.6%). The recalculated recall using only the true errors AND using the lowered threshold for Cohort Studies results in recall ranging from 95.6% to 100% (mean 99.3%).

Discussion

In the present study, we tested the hypothesis that one can achieve substantial work savings and near-perfect recall using a publication type filtering strategy for automated triage that was applied retrospectively to 10 Drug Effectiveness Review Project SRs. Using predictive scores from Multi-Tagger[8], we initially set thresholds for filtering out articles based on an optimal balance between precision and recall; future work could also consider personalized thresholds to optimize for high recall for particular study designs. The results are very encouraging and will inform our plans to implement publication type filtering prospectively during the creation of new SRs by a variety of teams.

Another opportunity for future work is to analyze whether this filtering strategy may potentially identify additional articles not found by the SR team. Because our focus was work savings and recall based on the actual SR results, we did not consider whether this filtering strategy may be better in some ways than human searching. For example, the strategy could enable an SR team to start from a larger initial set of articles that they would not have had the resources to screen manually. Additional analysis is needed to understand whether such a strategy would result in additional articles relevant for inclusion while still resulting in work savings. In the present study, we assumed that the actual SR results were the ideal set of articles.

Our study has several limitations. First, we only included articles that have PMIDs in our analysis due to the availability of MeSH terms and publication types in PubMed metadata. Additionally, because the study was retrospective, we were limited in our understanding of the SR process, such as the context of the inclusion of some articles. Past research on reducing workload in reviews[12] has noted that different topics exhibit different work savings and recall. We did not examine the role of topic other than to note that different SRs varied in the number and kind of article types that they deemed relevant, which could certainly impact on the performance of our strategy. Additional

SRs conducted by a variety of teams need to be analyzed in order to ascertain the most appropriate predictive score thresholds that will ensure maximal recall while still providing substantial work savings. Prioritizing screening based on the tagger score, rather than pre-specifying a threshold, should also be tested in future work. Some studies designs often include a variety of types, and those with less common designs can be missed by the tagger. For example, two database studies (PMID 22966091; PMID 25300980) and a birth cohort study (PMID 22813094) that were filtered out lacked the MeSH Cohort Studies term and may have received low Cohort Study Multi-Tagger predictions because their characteristics are not typical of Cohort Studies. One of the DERP reports used a “Best Evidence” approach, in which some designs (e.g., Randomized Controlled Trials) were prioritized over others, and some designs (including some not explicitly stated in the inclusion criteria) were considered if and when articles using the prioritized designs were not found. Our initial strategy did not account for this approach. Our choices described here, regarding which observational study designs to include, were somewhat arbitrary and across-the-board; however, our findings suggest that the list of included study designs should be expanded or refined to optimize results for individual SRs.

Conclusion

In order to apply the Multi-Tagger tool realistically in the workflow of a SR team, we suggest careful consideration of what article types might potentially be relevant but are often omitted from explicit inclusion. Automated publication type filtering may also be useful for other types of evidence syntheses such as rapid reviews[15] and scoping reviews. As well, publication type filtering should optimally be combined with other filtering or ranking methods[13–15] that may provide additional work savings at the manual triage stage. In the future, we plan to provide web-based tools for anyone to obtain predictive publication type scores for articles not indexed by PubMed (i.e., indexed in databases such as EMBASE or PsycINFO).

Data Availability

Data is publicly available at http://doi.org/10.13012/B2IDB-9257002_V1

Acknowledgements

Supported by NIH grant 5R01LM010817. Thanks to Halil Kilicoglu for comments on the manuscript.

References

1. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open [Internet]. 2017 Feb 1 [cited 2019 Jun 4];7(2):e012545. Available from: http://doi.org/10.1136/bmjopen-2016-012545

2. Michelson M, Reuter K. The significant cost of systematic reviews and meta-analyses: A call for greater involvement of machine learning to assess the promise of clinical trials. Contemp Clin Trials Commun [Internet]. 2019 Dec 1 [cited 2021 Aug 2];16:100443. Available from: http://doi.org/10.1016/j.conctc.2019.100443

3. Hamel C, Kelly SE, Thavorn K, Rice DB, Wells GA, Hutton B. An evaluation of DistillerSR’s machine learning-based prioritization tool for title/abstract screening – impact on reviewer-relevant outcomes. BMC Med Res Methodol [Internet]. 2020 Oct 15 [cited 2020 Nov 10];20(1):256. Available from: https://doi.org/10.1186/s12874-020-01129-1

4. Tsou AY, Treadwell JR, Erinoff E, Schoelles K. Machine learning for screening prioritization in systematic reviews: comparative performance of Abstrackr and EPPI-Reviewer. Syst Rev [Internet]. 2020 Apr 2 [cited 2021 Jan 2];9(1):73. Available from: https://doi.org/10.1186/s13643-020-01324-7

5. Giummarra MJ, Lau G, Gabbe BJ. Evaluation of text mining to reduce screening workload for injury-focused systematic reviews. Inj Prev J Int Soc Child Adolesc Inj Prev [Internet]. 2019 Aug 26; Available from: http://doi.org/10.1136/injuryprev-2019-043247

6. Gates A, Guitard S, Pillay J, Elliott SA, Dyson MP, Newton AS, et al. Performance and usability of machine learning for screening in systematic reviews: a comparative evaluation of three tools. Syst Rev [Internet]. 2019 Nov 15 [cited 2019 Dec 1];8(1):278. Available from: https://doi.org/10.1186/s13643-019-1222-2

7. Allen IE, Olkin I. Estimating time to conduct a meta-analysis from number of citations retrieved. JAMA [Internet]. 1999 Aug 18 [cited 2020 Aug 11];282(7):634–5. Available from: http://doi.org/10.1001/jama.282.7.634

8. Cohen AM, Schneider J, Fu Y, McDonagh MS, Das P, Holt AW, et al. Fifty ways to tag your PubTypes: multi-tagger, a set of probabilistic publication type and study design taggers to support biomedical indexing and evidence-based medicine [Internet]. 2021 Jul [cited 2021 Jul 29]. Available from: http://doi.org/10.1101/2021.07.13.21260468

9. Cohen AM, Smalheiser NR, McDonagh MS, Yu C, Adams CE, Davis JM, et al. Automated confidence ranked classification of randomized controlled trial articles: an aid to evidence-based medicine. J Am Med Inform Assoc [Internet]. 2015 May [cited 2019 Apr 9];22(3):707–17. Available from: http://doi.org/10.1093/jamia/ocu025

10. O'Connor AM, Tsafnat G, Thomas J, Glasziou P, Gilbert SB, Hutton B. A question of trust: can we build an evidence base to gain trust in systematic review automation technologies? Syst Rev [Internet]. 2019 Jun 18 [cited 2019 Nov 6];8(1):143. Available from: https://doi.org/10.1186/s13643-019-1062-0

11. DERP Reports [Internet]. Pacific Northwest Evidence-based Practice Center. 2020 [cited 2021 Jul 30]. Available from: https://www.ohsu.edu/evidence-based-practice-center/derp-reports

12. Pham MT, Waddell L, Rajić A, Sargeant JM, Papadopoulos A, McEwen SA. Implications of applying methodological shortcuts to expedite systematic reviews: three case studies using systematic reviews from agri-food public health. Res Synth Methods [Internet]. 2016 Dec [cited 2021 Aug 9];7(4):433–46. Available from: http://doi.org/10.1002/jrsm.1215

13. Kontonatsios G, Spencer S, Matthew P, Korkontzelos I. Using a neural network-based feature extraction method to facilitate citation screening for systematic reviews. Expert Syst Appl X [Internet]. 2020 Jul 1 [cited 2020 Dec 8];6:100030. Available from: http://doi.org/10.1016/j.eswax.2020.100030

14. Weißer T, Saßmannshausen T, Ohrndorf D, Burggräf P, Wagner J. A clustering approach for topic filtering within systematic literature reviews. MethodsX [Internet]. 2020 Jan 1 [cited 2021 Aug 12];7:100831. Available from: http://doi.org/10.1016/j.mex.2020.100831

15. Thomas J, Noel-Storr A, Marshall I, Wallace B, McDonald S, Mavergames C, et al. Living systematic reviews: 2. Combining human and machine effort. J Clin Epidemiol [Internet]. 2017 Nov 1 [cited 2019 Aug 21];91:31–7. Available from: http://doi.org/10.1016/j.jclinepi.2017.08.011