Outcome-based Evaluation of Systematic Review Automation

Wojciech Kusa
wojciech.kusa@tuwien.ac.at
TU Wien
Vienna, Austria

Guido Zuccon
g.zuccon@uq.edu.au
The University of Queensland
Brisbane, Australia

Petr Knoth
petr.knoth@open.ac.uk
The Open University
Milton Keynes, The United Kingdom

Allan Hanbury
allan.hanbury@tuwien.ac.at
TU Wien
Vienna, Austria

ABSTRACT

Current methods of evaluating search strategies and automated citation screening for systematic literature reviews typically rely on counting the number of relevant publications (i.e. those to be included in the review) and not relevant publications (i.e. those to be excluded). Significant importance is put into promoting the retrieval of all relevant publications through great attention to recall-oriented measures, and demoting the retrieval of non-relevant publications through precision-oriented or cost metrics. This established practice, however, does not accurately reflect the reality of conducting a systematic review, because not all included publications have the same influence on the final outcome of the systematic review. More specifically, if an important publication gets excluded or included, this might significantly change the overall review outcome, while not including or excluding less influential studies may only have a limited impact. However, in terms of evaluation measures, all inclusion and exclusion decisions are treated equally and, therefore, failing to retrieve publications with little to no impact on the review outcome leads to the same decrease in recall as failing to retrieve crucial publications.

We propose a new evaluation framework that takes into account the impact of the reported study on the overall systematic review outcome. We demonstrate the framework by extracting review meta-analysis data and estimating outcome effects using predictions from ranking runs on systematic reviews of interventions from CLEF TAR 2019 shared task. We further measure how closely the obtained outcomes are to the outcomes of the original review if the arbitrary rankings were used. We evaluate 74 runs using the proposed framework and compare the results with those obtained using standard IR measures. We find that accounting for the difference in review outcomes leads to a different assessment of the quality of a system than if traditional evaluation measures were used. Our analysis provides new insights into the evaluation of retrieval results in the context of systematic review automation, emphasising the importance of assessing the usefulness of each document beyond binary relevance.

ICTIR '23, July 23, 2023, Taipei, Taiwan
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0073-6/23/07.
https://doi.org/10.1145/3578337.3605135

CCS CONCEPTS

• **Information systems** → **Evaluation of retrieval results**; **Information retrieval**; *Retrieval effectiveness*; Specialized information retrieval.

KEYWORDS

systematic reviews, citation screening, evaluation, study outcomes, effect based evaluation, information retrieval

ACM Reference Format:
Wojciech Kusa, Guido Zuccon, Petr Knoth, and Allan Hanbury. 2023. Outcome-based Evaluation of Systematic Review Automation. In *Proceedings of the 2023 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '23), July 23, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3578337.3605135

1 INTRODUCTION

A systematic literature review is a well-established and rigorous methodology for synthesising and evaluating the evidence on a specific research question, which is particularly important in the field of medicine [13]. However, it is also gaining importance in other areas such as social sciences and engineering [2, 3, 17, 28]. The process involves a systematic search, critical appraisal, and synthesis of the available literature on a topic. During the critical appraisal step, every included publication has its weight and effect calculated based on the outcomes reported by that publication. This information influences the final outcome of the review.

One of the essential steps in conducting a systematic review is the process of citation screening, in which a large number of publications are initially identified through a literature search and then screened to determine those relevant to the review [1, 33]. This process can be time-consuming and labour-intensive, involving making thousands of eligibility decisions. Given the importance of citation screening in systematic literature reviews, there have been numerous attempts to automate the process [27]. Previous studies have investigated the use of automated citation screening methods for systematic literature reviews by utilising various natural language processing (NLP), machine learning (ML), and information retrieval (IR) methods to retrieve, rank, or classify references [11, 16, 19, 27, 29–31, 34, 38–40].

To understand the effectiveness of automated citation screening methods, practitioners have relied on metrics based on the notions of recall, precision and cost – and of a binary assessment of relevance[1] [19, 27, 34]. This practice assigns to every publication to be included in the review the same importance. So, for example, if method M_1 identifies as potentially relevant publications $\{A, B, C\}$ while method M_2 identifies publications $\{A, D, E\}$, and the ground truth is that the relevant publications are $\{A, B, D\}$, then M_1 and M_2 achieve the same recall, precision and cost. However, we argue, that the two sets $\{A, B, C\}$ and $\{A, D, E\}$ may not be equally important, and thus identifying either of B or D may not be equivalent if the outcomes of the review were considered. In fact, if excluded, some publications can significantly change a review's conclusion to the extent that the conclusion might be the opposite (e.g., from favouring a drug to favouring a placebo) [25, 26]. On the other hand, not including other publications might have only a small quantitative impact on the outcomes of the review.

We argue that a holistic evaluation of retrieval and automated citation screening methods for systematic review creation should not only consider the concepts of recall, precision and cost, but also the quality of the outcomes generated from the analysis of the automatically included publications. Following this direction, we propose a new evaluation framework that considers inclusion and exclusion information and meta-analysis data from reviews created by Cochrane – the largest organisation responsible for creating systematic literature reviews in medicine,[2] to estimate outcomes and weights of included publications. This information can be used to assess the quality of ranking and classification methods. This framework allows for assessing automatic approaches from the angle of how closely their *outcomes* – not just their set of included publications – are to the outcomes of the original review. By comparing the outcomes of the automated model to those of the original review, we can gain a better understanding of the quality of the automated approach and its effect on the final outcome of the review.

We propose five aspects of analysis focusing on different features of review outcomes. We explore initial experiments on the CLEF TAR 2019 dataset [16]. Our simulation results show that by randomly removing one publication per review (average recall of 92% publications), 95% of outcomes remain unchanged. However, after removing five publications (average recall of 63%), 76% of the outcomes are still the same, showing that the relationship between recall and achieved outcomes is not linear. We also show that the outcome-based evaluation emphasises different aspects of the models' performance than the traditional IR evaluation measures. We finally propose multi-objective optimisation to handle the problem of non-estimable outcomes.

We believe this new evaluation approach will provide a better understanding of the impact of automatic literature screening methods on the outcome of systematic literature reviews and help identify areas in which these methods can be improved.

2 RELATED WORK

The effectiveness of automatic approaches for search strategy creation and systematic review screening has been traditionally evaluated using binary relevance ratings [16, 27, 34], often sourced at the title and abstract screening level, rather than at the full-text level.

When the screening problem is treated as a ranking task (e.g., for the sub-task of screening prioritisation or stopping prediction), then rank-based metrics and metrics at a fixed cut-off are commonly used, e.g., *nDCG@n*, *Precision@n*, *Recall@n*, last relevant found [12, 32]. Cost-based and economic-based metrics are also used, especially in the context of the query formulation task in the CLEF TAR shared task [14–16], e.g., total cost (TC) or total cost with a weighted penalty (TCW). The TREC Total Recall track [9] also used a cut-off based metric, $recall@aR+b$, which is defined as the recall achieved when $aR+b$ documents have been identified, where R is the number of relevant documents in the collection and a and b are parameters. When $a = 1$ and $b = 0$, $recall@aR + b$ is equivalent to R-precision. In the patent domain, the PRES score has been proposed which takes into account achieved recall and the user's search effort [23].

When the screening problem is treated as a classification task, metrics based on the confusion matrix and the notion of Precision and Recall are commonly used [27, 34]: aside from Precision and Recall, metrics include variations of the harmonised mean between the two, i.e. F_β–score, utility, U19 [35–37], sensitivity-maximising thresholds [6], and AUC [4]. Another metric, Work Saved over Sampling (WSS), measures the amount of work saved when using machine learning models to screen irrelevant publications [5, 18, 19, 24]. The True Negative Rate (TNR) and nreTNR (normalised rectified TNR) were proposed as an alternative as it addresses some of the limitations of WSS regarding averaging scores from multiple datasets [20, 21].

Nussbaumer-Streit et al. [26] compared repeated literature searches using 14 abbreviated approaches (combinations of various databases with and without searches of reference lists) on a sample of 60 Cochrane systematic reviews of clinical interventions. They recalculated the main summary-of-findings table of each Cochrane review and asked original review authors whether the conclusions changed compared to the original review. They found that in only 2% of cases (95% CI: 0%–9%), combining one database with another or with searches of reference lists was falsely reaching an opposite conclusion compared to comprehensive searches. This outcome shows that identifying *all* relevant studies is not always crucial for obtaining the same review results.

Automated citation screening has become increasingly popular in systematic literature reviews due to its potential to reduce the time and cost required. However, current evaluation methods for these methods are limited to binary relevance assessment, where each publication is considered either relevant or irrelevant, and do not account for the impact of each publication on the review outcome. This is a vital issue, as the assumption that all relevant publications are equally important to the final outcome of the systematic review is not necessarily valid. Without an accurate assessment of the importance of each document, the conclusions of a systematic review may be biased or incomplete. To address this issue, in this paper, we propose a novel methodology for assessing citation screening based on evaluating outcome differences, which enables us to determine the impact of each publication on the systematic review.

[1] Every publication to be included in the review is labelled as relevant, while every excluded publication is non-relevant.
[2] https://www.cochrane.org

3 EVALUATION FRAMEWORK

This paper proposes a new evaluation framework for automated citation screening. Our framework includes three steps which are detailed in the following subsections. The first step is data extraction, where we extract statistics of the studies included in the review and match studies to publications. The second step is model evaluation, where we use the extracted data to estimate the review's outcomes for rankings or classifications of the citation list. The third step is the analysis of the results, where we compare the outcomes obtained from the alternative rankings to the outcomes of the original review. Our proposed framework allows for a more nuanced evaluation of automated citation screening methods. By considering the impact of each publication on the review's outcomes, we can identify which publications are most important to retrieve and prioritise them accordingly. Next we describe each step in detail.

3.1 Data Extraction

Cochrane systematic reviews distinguish between *study* and *publication.* A study is a distinct piece of research conducted to answer a specific research question or investigate a particular hypothesis. It typically involves a group of participants, data collection methods, and specific objectives. Publications, on the other hand, are the atomic units which reviewers screen. Each study can be reported by several publications, such as journal articles, conference proceedings, or research reports. Each publication may present different aspects or findings of the same study, but they are all derived from the same underlying research. We assume that a study has been found if at least one publication reporting it was successfully retrieved.

For every review, based on its Cochrane review ID, we identify its corresponding RevMan file and list of included publications. A RevMan file is the format used by Cochrane containing all statistical data about studies and outcomes included in the review. Outcomes of Cochrane reviews are reported in the following hierarchy: one comparison can have several outcomes, and one outcome can consist of a few subgroups. We extract all metadata from the RevMan files, such as the comparisons, outcomes and subgroups and the results of every included study. Note that the use of RevMan files is for experimental convenience, but is not a requirement from the framework: the required data could be provided in other formats. Furthermore, Cochrane recently announced that future systematic literature reviews would contain statistical data in more common CSV and RIS formats.[3]

Cochrane reports a list of included publications and studies which correspond to them. Traditionally, retrieval was conducted at the level of publications [14–16]. In order to be able to re-use previous relevance judgments, we need to assign PubMed IDs to these publications. Our process for matching PubMed IDs to publications is based on four steps in the following order:

- We check if the PubMed ID information is provided on the Cochrane references webpage.
- We conduct search in PubMed using ENTREZ[4] by searching for the same title and authors.
- We search for the PubMed ID in SemanticScholar[5] using publication DOI from Cochrane references webpage.
- We search again in PubMed, this time with a relaxed requirement by searching for an exact match in the title only.

3.2 Model Evaluation

When conducting a meta-analysis, for every outcome, each study has its weight and effect size calculated first (respectively columns 6 and 7 on example forest plots in Figure 1). Effect size is an essential statistical concept in the analysis of research data [10]. It is a measure that quantifies the magnitude of difference between two groups in a study. Researchers use a variety of effect measures to compare outcome data between two intervention groups, including odds ratios and mean differences.

For instance, in ratio effect measures, a value of 1 represents no difference between the groups [7, 8]. On the other hand, in difference measures, a value of 0 represents no difference between the groups. Values higher or lower than these "null" values may indicate either benefit or harm of an experimental intervention, depending on the order of the interventions in the comparison and the nature of the outcome. Every estimate is expressed with a measure of uncertainty, such as a confidence interval (CI) or standard error (SE).

Effects depend on the number of events reported by that study, whereas weights assigned to each study are influenced by other studies included in this outcome. So when removing one study from the meta-analysis, only the weights of the remaining studies will change, but their effect sizes will stay the same (compare Figures 1a and 1c). There are several types of outcomes reported by Cochrane, in our study, we focus on the dichotomous and continuous outcomes only and calculate them following the approach by Deeks and Higgins [8].

Our framework takes arbitrary ranking or classification runs and calculates the final outcomes of the review based on publications included in the run. When evaluating a classification run or a search result, we take all publications predicted as relevant. When evaluating ranking runs, we need to assume a cut-off point. Previous studies working on systematic review automation used either cut-off at r% of recall [5, 20], or at d% of total dataset size [14, 15].

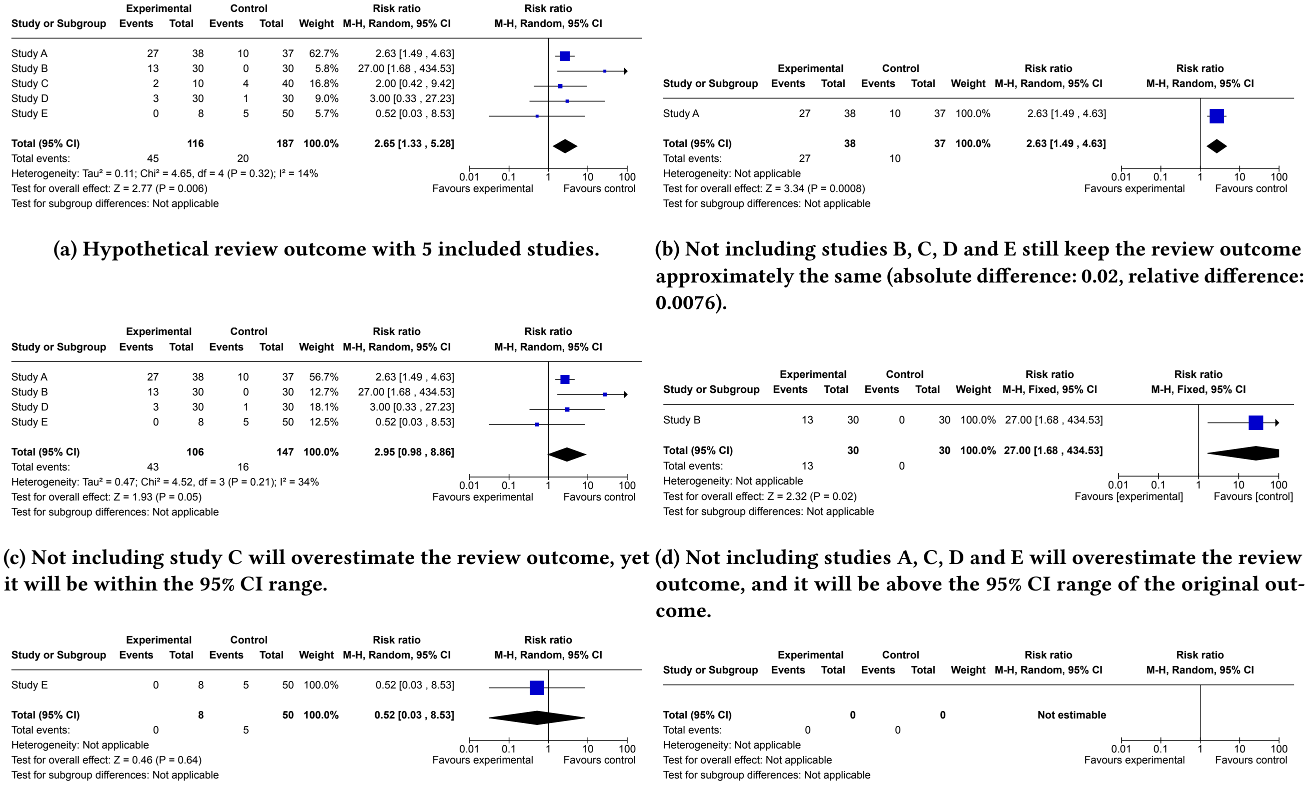
3.3 Results Analysis

We analyse the results by examining the outcomes generated by the run and compare them with the outcomes obtained by the original review (Figure 1). We extend the analysis done by Nussbaumer-Streit et al. [26], who proposed two categories of "changed conclusions": (1) if the new review drew the opposite conclusion, (2) if it is not possible to draw a conclusion or the new conclusion has less certainty. We distinguish five aspects of analysis for review outcomes against the original review (Figure 1a). The first two of these aspects are real-valued, whereas the remaining three are categorical:

[3] https://www.cochrane.org/news/cochrane-improving-way-we-manage-and-share-data-associated-our-reviews

[4] https://www.ncbi.nlm.nih.gov/search/

[5] https://www.semanticscholar.org



(a) Hypothetical review outcome with 5 included studies.

(b) Not including studies B, C, D and E still keep the review outcome approximately the same (absolute difference: 0.02, relative difference: 0.0076).

(c) Not including study C will overestimate the review outcome, yet it will be within the 95% CI range.

(d) Not including studies A, C, D and E will overestimate the review outcome, and it will be above the 95% CI range of the original outcome.

(e) Not including studies A, B, C and D will change the study outcome – from 'favours control' to 'favours experimental'.

(f) Not including any study makes the outcome non-estimable.

Figure 1: Different versions of review outcomes represented as forest plots. Each row is a single study. Columns from the right represent, respectively: (1) the study identifier, (2) number of events in the experimental group (e.g., patients with specific symptoms or adverse events), (3) experimental group size, (4) number of events in the control group, (5) control group size, (6) the weight of a study, and (7) effect size of a study: a difference (e.g., risk ratio or standardised mean difference) in events between experimental or control group. Simulations and figures done using RevMan Web, available at http://revman.cochrane.org.

(1) *Magnitude of difference* — By how much are the outcomes different in their effect size (Figure 1a versus 1b)? In other words, what is the numerical *impact* on the review outcome when certain studies are not included? This is measured by calculating the relative difference in effect size between the original outcome O_o and predicted outcome O_p: $MoD = \frac{\|O_o - O_p\|}{\|O_o\|}$. When $O_o = 0$ and $O_p \neq 0$, we assume $MoD = 100\%$; otherwise, when $O_o = O_p = 0$, we set $MoD = 0\%$. Similarly, when the predicted outcome cannot be estimated, we assume $MoD = 100\%$.

(2) *Distance from CI* — Is the new outcome within the Confidence Interval (CI) of the original outcome (Figure 1c)? The answer is a distance between the predicted outcome O_p and the closest of the pair (CI_{lower}, CI_{upper}):

$$\Delta_{CI} = \begin{cases} \|O_p - CI_{lower}\| & \text{if } O_p < CI_{lower}, \\ \|O_p - CI_{upper}\| & \text{if } O_p > CI_{upper}, \\ 0 & \text{otherwise.} \end{cases}$$

(3) *Overestimation/underestimation* — Is the outcome overestimated or underestimated compared to the original one (Figure 1d)? We first check if the calculated outcome is equal (due to the limits of precision of data reported in RevMan files, we use the relative and absolute tolerance of 10^{-5} and 10^{-6} respectively). Then, if the outcome is different, we check if the result is higher than the original (overestimation) or lower (underestimation). The answer has three options: "overestimated", "underestimated", and "equal".

(4) *Sign* — Does the outcome have the same sign as the original one (Figure 1e)? In other words, are the new conclusions opposite to the original ones? The answer is binary: "yes"/"no". This aspect corresponds to the first category from Nussbaumer-Streit et al. [26].

(5) *Estimability* — Is it possible to calculate the outcome (Figure 1f)? An outcome cannot be calculated if there are no included studies concerning it. The answer is binary: "yes"/"no".

Table 1: Statistics of the considered dataset.

		CLEF TAR 2019	
Dataset split		Training	Test
Reviews' type		— Interventional —	
# Reviews		17	15
Outcomes per review	Min	2	3
	Median	9	15
	Max	41	128
Studies per outcome	Min	1	1
	Median	2	2
	Max	55	40

4 EXPERIMENT SETUP

Contrary to the traditional evaluation based on retrieving relevant publications, with our framework we envision the evaluation in an outcome-based approach. Specifically, we do not treat a dataset as a collection of systematic reviews but rather a collection of outcomes. The problem of conducting a systematic review is multi-dimensional. One can think of it as having several outcomes reporting different dimensions of the review, and the evaluation of the user's needs is conducted independently from each outcome's perspective. We do not want to average across reviews, each containing a different number of outcomes. We add or average these outcome-level results instead.

Before we present the results, we first discuss the used dataset and models.

4.1 Dataset and Models

We used a collection of 38 systematic reviews of interventions from the CLEF TAR 2019 training and test datasets [16]. Each review consists of a Cochrane ID, a protocol, and a list of publications described by their PubMedIDs with qrels both on the title and abstract level and a full-text level. We enhanced the dataset by collecting RevMan files and information about the data and analysis as described in Section 3.1.

Out of 38 reviews in CLEF TAR 2019, our script found studies and outcomes for 32 reviews (17 in the training subset and 15 in the test subset). We summarise the statistics of the 32 reviews we consider in Table 1. There is a significant discrepancy in the number of outcomes reported by the reviews, ranging from as few as 2 or 3 outcomes in small reviews to 128 outcomes in the largest one. Moreover, the majority of these outcomes come from just one or two studies, which presents an additional challenge.

These 32 reviews report 1640 included publications, out of which we managed to find PubMed IDs for 1175 of them (71.6%). Next, we wanted to match publications identified with our script to the CLEF TAR 2019 qrels based on the PubMed ID. There were, in total, 778 relevant documents on the full-text level identified in the CLEF TAR for these 32 reviews. We successfully merged 741 publications (95.2% of the total in CLEF TAR); there are only 37 publications in CLEF TAR 2019 qrels which we do not have in our records.

We use 34 official CLEF TAR 2019 runs from three teams. The teams used a variety of ranking methods, including traditional BM25, interactive BM25, continuous active learning, relevance feedback, and various stopping criteria. Additionally, we included 40 runs based on the reproducibility of the active learning method by Yang et al. [41]. In total, we evaluate 74 runs, but for the sake of brevity, in this paper, we present the results on a subset of 28 runs, as some of the runs were very similar to each other. Our model requires full-text assessments, and thus, we use qrels from the full-text level, despite the fact that runs have been trained on titles and abstracts. While this might not be fair towards the evaluated systems, our experiments aim not to establish which systems are better but to provide an example of the operationalisation of our framework and its implications.

5 OUTCOME-BASED EVALUATION

We first run a simulation study to understand the results of our evaluation framework better in a controlled manner. Then, we discuss the usage of the evaluation framework with retrieval and classification runs on CLEF TAR 2019 collection.

5.1 Preliminary Simulation

We are interested in executing a preliminary study to understand the effect our outcome-oriented evaluation has on the analysis of systematic review automation methods.

We simulate the evaluation framework by taking the set of included *publications* for each review and randomly removing [1, 2, 3, 4, 5, 10, 15, 20, 30, 50, 100] publications from the set and then recalculating the outcomes. In other words, we are interested in exploring the impact of false negatives on the final review outcome. We compare the outcomes with the 'gold' outcomes from the original review. Results from all 32 systematic reviews are reported in Table 2. In our analysis, we consider the metrics from all five analysis aspects (Section 3.3), as well as the Recall.

Figure 2 presents box plots of averaged relative difference (aspect (1)) values from our simulation at a cut-off at 20% of the total number of documents. These results validate our expectations regarding the behaviour of this aspect of analysis as the relative difference grows with the number of removed publications. On the other hand, the distance to confidence intervals (aspect (2), Figure 3) does not show any specific trend on the CLEF 2019 reviews.

Out of all the metrics, the one that changes the most when varying the number of removed publications is estimability (5).

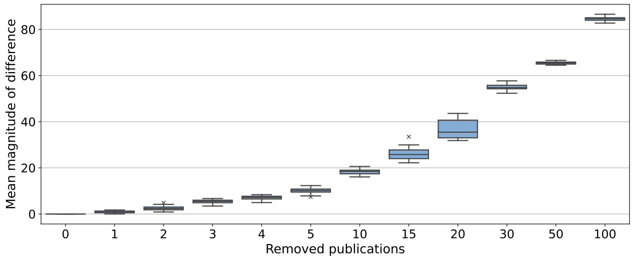


Figure 2: Box plots presenting relative difference values from 20 simulations on the publication level. Note that the x-axis does not preserve the linear step.

Table 2: Initial results of the simulation on the publication level. Outcomes are aggregated across 32 systematic reviews and are averaged from 20 different random seeds.

			N relevant **publications** removed from the review										
Analysis Aspect		gold	1	2	3	4	5	10	15	20	30	50	100
1	Mean relative difference	0.0	0.9	2.5	5.3	7.1	10.0	18.3	26.2	36.5	54.9	65.5	84.5
2	Mean distance from CI	0.000	0.002	0.003	0.004	0.007	0.008	0.013	0.042	0.102	0.018	0.008	0.083
3	Equal outcome	824	786	750	706	657	623	496	410	340	256	164	80
	Different	0	38	73	117	167	200	328	413	483	567	659	743
	- Underestimated	0	17	27	38	57	66	98	103	90	55	58	23
	- Overestimated	0	20	45	79	109	134	229	309	393	512	601	720
4	Have same sign	824	815	800	774	756	735	663	597	516	365	277	121
	Have different sign	0	9	24	49	67	88	160	227	307	458	546	702
5	Reported outcomes	824	816	804	781	767	743	675	610	529	371	284	128
	Missing outcomes	0	7	20	43	56	80	148	213	294	452	539	695
Average *Recall* for publications		1.00	0.92	0.84	0.75	0.70	0.63	0.45	0.35	0.28	0.22	0.14	0.05
Average *Recall* for studies		1.00	0.97	0.91	0.80	0.77	0.68	0.53	0.43	0.37	0.31	0.22	0.12

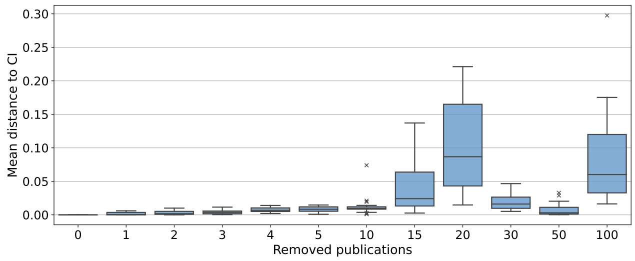


Figure 3: Box plots presenting distance to confidence intervals values from 20 simulations on the publication level. Note that the x-axis does not preserve the linear step.

As more publications are removed, it becomes more and more challenging to calculate outcomes, predominantly because half of the original outcomes relied on one or two studies. At the very extreme, when 100 publications are removed from every review, only 15% of outcomes are still estimable.

The measure of overestimation and underestimation (3) is showing growing trends with more publications being removed. Already not including one publication per review (achieving an average recall of 92% for publications and 97% for studies) changed 38 outcomes (4.6% of the total number of outcomes). This shows that the commonly used threshold of 95% Recall does not enforce preserving the same outcomes of the review. We also notice that the sign (4) aspect is not very descriptive across the simulations as it is mainly influenced by non-estimable outcomes.

5.2 Evaluation with actual runs

In this section, we use the prediction on the test subset of the dataset from runs described in Section 4.1 and evaluate them using our framework. We further consider two baselines:

gold – the best possible run which returns all relevant studies from the original review first.

max-with-qrels – this run takes into account the limitations of the CLEF TAR collection and our PubMed articles matching process. It uses all relevant studies identified in the CLEF TAR 2019 qrels as relevant and places them first.

We follow the evaluation procedure of CLEF TAR and calculate the following traditional evaluation measures: Mean Average Precision (*MAP*), last relevant found, Recall@k% of top-ranked publications, with k in [5, 10, 20, 30, 50], Work Saved over Sampling at r% of recall with r in [95%, 100%] (*WSS*@95%, *WSS*@100%), *nDCG*@20% of top-ranked publications and Area Under Recall Curve (*AURC*). CLEF TAR as their primary reporting measure used *MAP*; therefore, we will treat *MAP* as the reference measure when sorting runs. We do not evaluate baselines with traditional measures, yet for the purpose of sorting, we assume that they achieved the highest MAP score.

We calculate the relative difference in study outcomes (analysis aspect (1) in Section 3.3) for every outcome in all reviews. The lower the average score is, the better the runs, as their effect differs less from the original review effect. As considered runs were rankings, we follow the same procedure as for Recall and nDCG, namely we calculate the relative difference at k% of top-ranked publications with k in [5, 10, 20, 30, 50].

Figure 4 presents a box plot of relative difference per outcome calculated at 30% cut-off of dataset size for 15 test CLEF TAR reviews. Except for the best run, all other runs changed their rank when ordered using their mean relative difference score compared to the MAP-based ranking. While top runs, according to MAP scores, have low variability, there are runs among the top 10 which show considerable fluctuation. This means there are specific reviews for which these runs will lead to significantly different decisions about the outcome. This behaviour is comparable for relative difference at other cut-offs *k*.

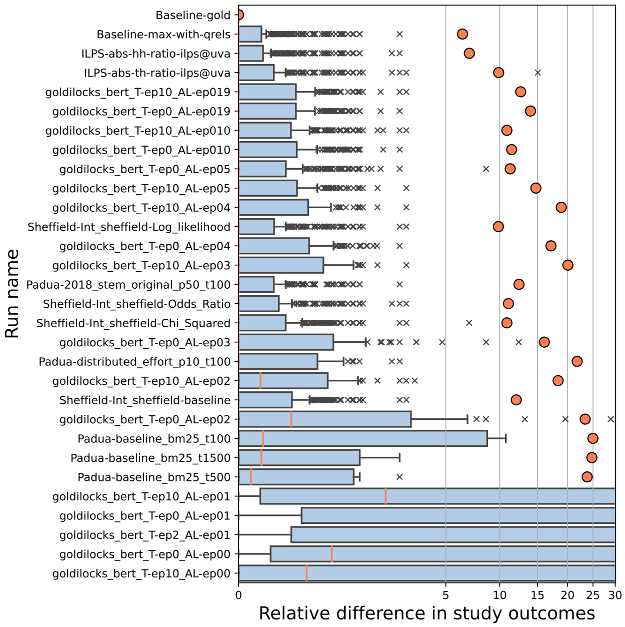


Figure 4: Box plot presenting runs with their relative difference in study outcomes for an evaluation with a cut-off at 30% of the total number of documents for each review. Runs are sorted by their MAP score. The orange circle denotes the mean relative difference @30%. The X-axis is cut at 30, while the outliers exist up to the value of 100; we cut for visualisation purposes.

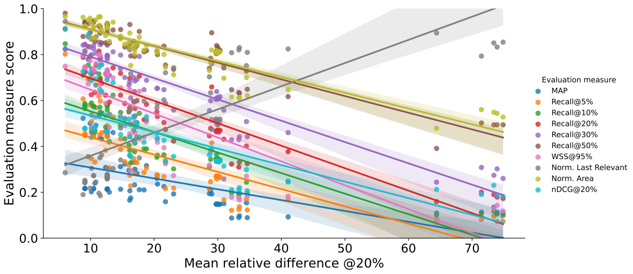


Figure 5: Linear regression fits between relative difference at 20% cut-off of documents and other evaluation measures scores. Correlations for relative difference at other cut-offs follow similar trends.

What is also interesting is that the mean relative difference at 30% cut-off for the *max-with-qrels* baseline run equals 6.24. Furthermore, for the relative difference score calculated at 100% of documents, this baseline score is also not equal to 0. This means that the limitations of the CLEF TAR collection and qrels establish a lower bound for the best achievable value of relative difference.

Figure 5 presents correlation between relative difference calculated at 20% cut-off of dataset size and evaluation measures used at CLEF TAR 2019. The score correlates positively with the last relevant found, but there is a negative correlation with all other measures. This confirms our intuition that a higher average relative difference score across outcomes means a worse model effectiveness, as the ideal 'best' model should achieve a difference of 0.

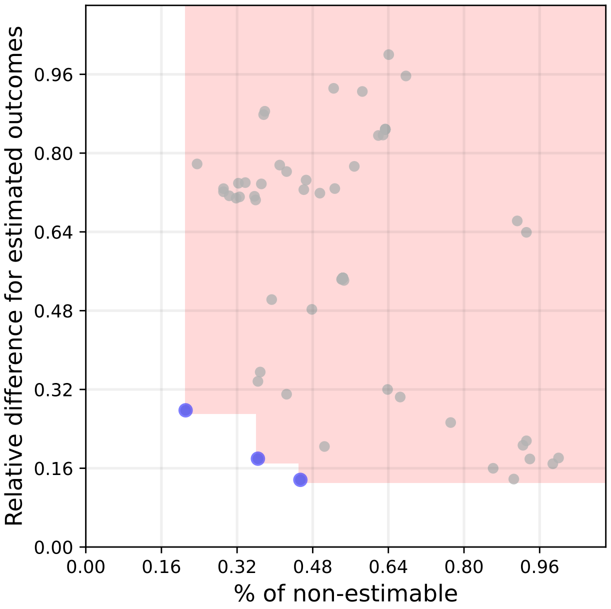


Figure 6: Visualisation of the Pareto frontier for two objectives: (1) number of non-estimable outcomes on the x-axis and (2) sum of relative difference for estimable outcomes on the y-axis. Both objectives are to be minimised. Runs are evaluated at a cut-off at 5% of the total number of documents for each review. Non-dominated runs are marked with a blue colour.

5.3 Pareto Frontier Optimisation

Based on the simulation results, we note a problem with non-estimable outcomes. Should these outcomes be assigned a zero score or maybe an infinite value? This raises the issue of handling these values in the evaluation process for calculating relative difference scores. In our study, we assigned a zero value to non-estimable outcomes, which allowed us to assume that the relative difference equals 100%. Nevertheless, this yields the problem of when the actual outcome is equal to the zero value (i.e., the study does not favour the experimental nor the control group), as the difference, in this case, would also be zero. One way to overcome the issue of non-estimable outcomes would be to evaluate both estimability and relative difference implemented, for instance, using the Pareto frontier [22].

Figure 6 presents the Pareto frontier evaluated at a cut-off at 5% of the total number of documents. On the x-axis, we show the number of non-estimable outcomes for each run. On the y-axis, there is a sum of relative difference for estimable outcomes. We min-max normalise the sums including the gold baseline run (gold represents the best achievable score of $(0, 0)$). Both objectives should be minimised, i.e., we want to have as few non-estimable outcomes as possible and for all estimated outcomes, the difference would be as close to zero as possible. Contrary to the previous evaluations, we can notice that no single run would dominate on both dimensions.

6 LIMITATIONS

The primary objective of this paper was to introduce the concept of evaluating automated methods for systematic reviews based on their impact on review outcomes, rather than relying on binary qrels. In this section, we reflect on the potential limitations that arise when attempting to fully operationalise our proposed framework.

Do not optimise models using this measure. A practice that can be observed across the field is treating evaluation measures as an optimisation objective. We believe that our evaluation approach should not be used for optimising models. The notion of difference in study outcomes is only known a-posteriori when the review is completed. Using absolute differences in study outcomes as an optimisation objective might lead to over-fitting to biases in data.

Other types of systematic reviews. We focus only on systematic reviews of interventions which have a clear structure and evaluate the effectiveness of specific treatments, programs, or policies by comparing experimental setups with control groups. However, there are several other types of systematic reviews, such as diagnostic test accuracy reviews, prognostic reviews, and qualitative research reviews, each of which presents unique challenges for automation and evaluation [16]. Future work should investigate how this outcome-based evaluation framework can be extended to these other types of reviews.

Different outcome types. While our proposed evaluation framework focuses on continuous and dichotomous outcomes, other types of outcomes may be reported in systematic reviews, including ordinal, count, and time-to-event data. In our analysis, however, we found that continuous and dichotomous outcomes comprised most of the outcomes in the dataset we studied, accounting for 92% of all reported outcomes across 32 CLEF TAR 2019 reviews.

We believe that our evaluation framework could be generalised to incorporate other types of outcomes. Additionally, while we attempted to closely follow the evaluation protocols from the Cochrane handbook, some shortcuts were taken during the implementation process (for 2.4% of outcomes our effect calculations yielded marginally different results). In future work, ideally, access to RevMan or another official program for calculating study outcomes would be needed to make sure that all outcome types are covered.

Title and abstract screening. We work on the outcomes extracted from the full-text screening and use relevance judgments from full-text screening to judge the runs. However, most models are trained on titles and abstracts, which might make this an unfair comparison.

7 CONCLUSION

This paper puts forward a novel, outcome-based evaluation framework for assessing the effectiveness of automatic search strategies and citation screening methods in the context of systematic literature reviews. Our proposed framework evaluates the quality of these methods based on how closely the outcomes of their included publications match the actual review outcomes. We believe that this approach offers a more accurate reflection of real-world scenarios where not all included publications have the same impact on the final review outcome.

In addition to proposing the framework, we explore five analysis aspects that it enables, including measuring the numerical difference in predicted systematic review outcomes. We run initial experiments to simulate the impact of false negatives on reviews' outcomes showing that five missing publications per review can change 24% of outcomes. We also compare the evaluation results obtained using our framework with those obtained using traditional evaluation methods on CLEF TAR 2019 runs, highlighting the differences in focus between the two approaches.

Overall, we believe this framework represents a step forward in developing more effective and realistic methods for evaluating automation methods in the context of systematic literature reviews in medicine and in other domains in which the importance of systematic reviews is increasing.

ACKNOWLEDGMENTS

This work was supported by the EU Horizon 2020 ITN/ETN on Domain Specific Systems for Information Extraction and Retrieval – DoSSIER (H2020-EU.1.3.1., ID: 860721). This research is partially funded by the Australian Research Council, Discovery Project DP210104043.

REFERENCES

[1] Alexandra Bannach-Brown, Piotr Przybyła, James Thomas, Andrew S. C. Rice, Sophia Ananiadou, Jing Liao, and Malcolm Robert Macleod. 2019. Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. *Systematic Reviews 2019 8:1* 8, 1 (1 2019), 1–12. https://doi.org/10.1186/S13643-019-0942-7

[2] Gary S Bilotta, Alice M Milner, and Ian Boyd. 2014. On the use of systematic reviews to inform environmental policies. *Environmental Science & Policy* 42 (2014), 67–77.

[3] Stefanie Castillo and Petar Grbovic. 2022. The APISSER Methodology for Systematic Literature Reviews in Engineering. *IEEE Access* 10 (2022), 23700–23707.

[4] Aaron M Cohen, Kyle Ambert, and Marian McDonagh. 2010. A prospective evaluation of an automated classification system to support evidence-based medicine and systematic review. In *AMIA annual symposium proceedings*, Vol. 2010. American Medical Informatics Association, 121.

[5] A. M. Cohen, W. R. Hersh, K. Peterson, and Po Yin Yen. 2006. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association* 13, 2 (3 2006), 206–219. https://doi.org/10.1197/jamia.M1929

[6] Siddhartha R Dalal, Paul G Shekelle, Susanne Hempel, Sydne J Newberry, Aneesa Motala, and Kanaka D Shetty. 2013. A pilot study using machine learning and domain knowledge to facilitate comparative effectiveness review updating. *Medical Decision Making* 33, 3 (2013), 343–355.

[7] Jonathan J Deeks, Julian PT Higgins, Douglas G Altman, and Cochrane Statistical Methods Group. 2019. Analysing data and undertaking meta-analyses. *Cochrane handbook for systematic reviews of interventions* (2019), 241–284.

[8] Jonathan J. Deeks and Julian P. T. Higgins. 2010. Statistical algorithms in Review Manager 5. *Statistical Methods Group of The Cochrane Collaboration* 1, 11 (2010).

[9] Maura R Grossman, Gordon V Cormack, and Adam Roegiest. 2016. TREC 2016 Total Recall Track Overview.. In *TREC*.

[10] Julian PT Higgins, Tianjing Li, and Jonathan J Deeks. 2019. Choosing effect measures and computing estimates of effect. *Cochrane handbook for systematic reviews of interventions* (2019), 143–176.

[11] Brian E. Howard, Jason Phillips, Kyle Miller, Arpit Tandon, Deepak Mav, Mihir R. Shah, Stephanie Holmgren, Katherine E. Pelch, Vickie Walker, Andrew A. Rooney, Malcolm Macleod, Ruchir R. Shah, and Kristina Thayer. 2016. SWIFT-Review: A text-mining workbench for systematic review. *Systematic Reviews* 5, 1 (5 2016), 1–16. https://doi.org/10.1186/s13643-016-0263-z

[12] Brian E. Howard, Jason Phillips, Arpit Tandon, Adyasha Maharana, Rebecca Elmore, Deepak Mav, Alex Sedykh, Kristina Thayer, B. Alex Merrick, Vickie Walker, Andrew Rooney, and Ruchir R. Shah. 2020. SWIFT-Active Screener: Accelerated document screening through active learning and integrated recall estimation. *Environment International* 138 (5 2020), 105623. https://doi.org/10.1016/J.ENVINT.2020.105623

[13] Akers Jo, Aguiar-Ibáñez Raquel, Burch Jane, Chambers Duncan, Eastwood Alison, Fayter Debra, Hempel Susanne, Light Kate, Rice Stephen, Rithalia Amber, Stewart Lesley, Stock Christian, Wilson Paul, and Woolacott Nerys. 2009. *Systematic Reviews: CRD's guidance for undertaking reviews in health care.* CRD, University of York, York. www.york.ac.uk/inst/crd

[14] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. 2017. CLEF 2017 technologically assisted reviews in empirical medicine overview. *CEUR Workshop Proceedings* 1866 (9 2017), 1–29. https://pureportal.strath.ac.uk/en/publications/clef-2017-technologically-assisted-reviews-in-empirical-medicine-

[15] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. 2018. CLEF 2018 technologically assisted reviews in empirical medicine overview. *CEUR Workshop Proceedings* 2125 (7 2018). https://pureportal.strath.ac.uk/en/publications/clef-2018-technologically-assisted-reviews-in-empirical-medicine-

[16] E. Kanoulas, Dan Li, Leif Azzopardi, and René Spijker. 2019. CLEF 2019 Technology Assisted Reviews in Empirical Medicine Overview. In *CLEF.*

[17] Staffs Keele et al. 2007. *Guidelines for performing systematic literature reviews in software engineering.* Technical Report. Technical report, ver. 2.3 ebse technical report. ebse.

[18] Georgios Kontonatsios, Sally Spencer, Peter Matthew, and Ioannis Korkontzelos. 2020. Using a neural network-based feature extraction method to facilitate citation screening for systematic reviews. *Expert Systems with Applications: X* 6 (7 2020), 100030. https://doi.org/10.1016/j.eswax.2020.100030

[19] Wojciech Kusa, Allan Hanbury, and Petr Knoth. 2022. Automation of Citation Screening for Systematic Literature Reviews Using Neural Networks: A Replicability Study. In *Advances in Information Retrieval*, Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (Eds.). Springer International Publishing, Cham, 584–598. https://arxiv.org/abs/2201.07534v1

[20] Wojciech Kusa, Aldo Lipani, Petr Knoth, and Allan Hanbury. 2023. An Analysis of Work Saved over Sampling in the Evaluation of Automated Citation Screening in Systematic Literature Reviews. *Intelligent Systems with Applications* 18 (2023), 200193. https://doi.org/10.1016/j.iswa.2023.200193

[21] Wojciech Kusa, Aldo Lipani, Petr Knoth, and Allan Hanbury. 2023. VoMBaT: A Tool for Visualising Evaluation Measure Behaviour in High-Recall Search Tasks. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23).* ACM, Taipei, Taiwan, 5. https://doi.org/10.1145/3539618.3591802

[22] Alexander V Lotov and Kaisa Miettinen. 2008. Visualizing the Pareto Frontier. *Multiobjective optimization* 5252 (2008), 213–243.

[23] Walid Magdy and Gareth JF Jones. 2010. PRES: a score metric for evaluating recall-oriented information retrieval applications. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval.* 611–618.

[24] Stan Matwin, Alexandre Kouznetsov, Diana Inkpen, Oana Frunza, and Peter O'Blenis. 2010. A new algorithm for reducing the workload of experts in performing systematic reviews. *Journal of the American Medical Informatics Association* 17, 4 (7 2010), 446–453. https://doi.org/10.1136/JAMIA.2010.004325

[25] B Nussbaumer-Streit, I Klerings, AI Dobrescu, E Persad, A Stevens, C Garritty, C Kamel, L Affengruber, VJ King, and G Gartlehner. 2020. Excluding non-English publications from evidence-syntheses did not change conclusions: a meta-epidemiological study. *Journal of clinical epidemiology* 118 (2020), 42–54.

[26] Barbara Nussbaumer-Streit, Irma Klerings, Gernot Wagner, Thomas L. Heise, Andreea I. Dobrescu, Susan Armijo-Olivo, Jan M. Stratil, Emma Persad, Stefan K. Lhachimi, Megan G. Van Noord, Tarquin Mittermayr, Hajo Zeeb, Lars Hemkens, and Gerald Gartlehner. 2018. Abbreviated literature searches were viable alternatives to comprehensive searches: a meta-epidemiological study. *Journal of Clinical Epidemiology* 102 (2018), 1–11. https://doi.org/10.1016/j.jclinepi.2018.05.022

[27] Alison O'Mara-Eves, James Thomas, John McNaught, Makoto Miwa, and Sophia Ananiadou. 2015. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews* 4, 1 (2015), 1–22.

[28] Mark Petticrew and Helen Roberts. 2008. *Systematic reviews in the social sciences: A practical guide.* John Wiley & Sons.

[29] Harrisen Scells, Guido Zuccon, and Bevan Koopman. 2020. You can teach an old dog new tricks: Rank fusion applied to coordination level matching for ranking in systematic reviews. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part I 42.* Springer, 399–414.

[30] Harrisen Scells, Guido Zuccon, and Bevan Koopman. 2021. A comparison of automatic Boolean query formulation for systematic reviews. *Information Retrieval Journal* 24, 1 (2021), 3–28.

[31] Harrisen Scells, Guido Zuccon, Bevan Koopman, and Justin Clark. 2020. Automatic Boolean Query Formulation for Systematic Review Literature Search. In *The Web Conference 2020 - Proceedings of the World Wide Web Conference, WWW 2020.* ACM, New York, NY, USA, 1071–1081. https://doi.org/10.1145/3366423.3380185

[32] Harrisen Scells, Guido Zuccon, Bevan Koopman, Anthony Deacon, Leif Azzopardi, and Shlomo Geva. 2017. A test collection for evaluating retrieval of studies for inclusion in systematic reviews. *SIGIR 2017 - Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (8 2017), 1237–1240. https://doi.org/10.1145/3077136.3080707

[33] Guy Tsafnat, Paul Glasziou, George Karystianis, and Enrico Coiera. 2018. Automated screening of research studies for systematic reviews using study characteristics. *Systematic Reviews 2018 7:1* 7, 1 (4 2018), 1–9. https://doi.org/10.1186/S13643-018-0724-7

[34] Raymon van Dinter, Bedir Tekinerdogan, and Cagatay Catal. 2021. Automation of systematic literature reviews: A systematic literature review. *Information and Software Technology* 136 (8 2021), 106589. https://doi.org/10.1016/j.infsof.2021.106589

[35] Byron C. Wallace, Kevin Small, Carla E. Brodley, and Thomas A. Trikalinos. 2010. Active learning for biomedical citation screening. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2010), 173–181. https://doi.org/10.1145/1835804.1835829

[36] Byron C Wallace, Kevin Small, Carla E Brodley, and Thomas A Trikalinos. 2011. Who should label what? instance allocation in multiple expert active learning. In *Proceedings of the 2011 SIAM international conference on data mining.* SIAM, 176–187.

[37] Byron C Wallace, Thomas A Trikalinos, Joseph Lau, Carla Brodley, and Christopher H Schmid. 2010. Semi-automated screening of biomedical citations for systematic reviews. *BMC bioinformatics* 11, 1 (2010), 1–11.

[38] Shuai Wang, Harrisen Scells, Bevan Koopman, and Guido Zuccon. 2022. Automated MeSH Term Suggestion for Effective Query Formulation in Systematic Reviews Literature Search. *Intelligent Systems with Applications* (2022), 200141.

[39] Shuai Wang, Harrisen Scells, Bevan Koopman, and Guido Zuccon. 2023. Can ChatGPT write a good boolean query for systematic review literature search? *arXiv preprint arXiv:2302.03495* (2023).

[40] Shuai Wang, Harrisen Scells, Bevan Koopman, and Guido Zuccon. 2023. Neural Rankers for Effective Screening Prioritisation in Medical Systematic Review Literature Search. In *Proceedings of the 26th Australasian Document Computing Symposium* (Adelaide, SA, Australia) *(ADCS '22).* Association for Computing Machinery, New York, NY, USA, Article 4, 10 pages. https://doi.org/10.1145/3572960.3572980

[41] Eugene Yang, Sean MacAvaney, David D Lewis, and Ophir Frieder. 2022. Goldilocks: Just-right tuning of BERT for technology-assisted review. In *European Conference on Information Retrieval.* Springer, 502–517.