From Little Things Big Things Grow: A Collection with Seed Studies for Medical Systematic Review Literature Search

Shuai Wang
The University of Queensland
St Lucia, Australia
shuai.wang2@uq.edu.au

Harrisen Scells
The University of Queensland
St Lucia, Australia
h.scells@uq.net.au

Justin Clark
Bond Institute for Evidence-Based Healthcare
Gold Coast, Australia
jclark@bond.edu.au

Bevan Koopman
CSIRO
Herston, Australia
bevan.koopman@csiro.au

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

ABSTRACT

Medical systematic review query formulation is a highly complex task done by trained information specialists. Complexity comes from the reliance on lengthy Boolean queries, which express a detailed research question. To aid query formulation, information specialists use a set of exemplar documents, called 'seed studies', prior to query formulation. Seed studies help verify the effectiveness of a query prior to the full assessment of retrieved studies. Beyond this use of seeds, specific IR methods can exploit seed studies for guiding both automatic query formulation and new retrieval models. One major limitation of work to date is that these methods exploit 'pseudo seed studies' through retrospective use of included studies (i.e., relevance assessments). However, we show pseudo seed studies are not representative of real seed studies used by information specialists. Hence, we provide a test collection with real world seed studies used to assist with the formulation of queries. To support our collection, we provide an analysis, previously not possible, on how seed studies impact retrieval and perform several experiments using seed study based methods to compare the effectiveness of using seed studies versus pseudo seed studies. We make our test collection and the results of all of our experiments and analysis available at http://github.com/ielab/sysrev-seed-collection.

CCS CONCEPTS

• **Information systems** → **Test collections**; *Retrieval effectiveness*; *Retrieval models and ranking*; **Expert search**.

KEYWORDS

Test Collection, Information Retrieval Evaluation, Systematic Reviews Creation, Seed Studies

1 INTRODUCTION

Systematic reviews are comprehensive reviews about a particular body of literature. Systematic reviews are used extensively in medicine for various reasons: by clinicians to make health decisions and by governmental and institutional policy and practice decisions about health topics. To ensure comprehensiveness, creators of systematic reviews employ a standardised approach to searching for literature in the form of Boolean queries. However, creating Boolean queries is often so complex that the authors of systematic reviews often engage with information specialists: highly specialised librarians skilled in searching for medical literature for systematic reviews. The Boolean query could be considered one of the most important aspects of a systematic review: it not only controls how many potentially relevant studies are retrieved but, perhaps more importantly, the total number of studies to screen. In other words, because systematic reviews aim to be comprehensive, the researchers will screen (i.e., assess for relevance to be included in the final systematic review) every study retrieved by the Boolean query. To ensure that no relevant studies are left out in the screening phase and to account for assessor biases, it is common for at least two assessors to screen studies. Naturally, since there may be thousands, if not tens of thousands of studies to screen, and since the screening process is duplicated by independent assessors, the Boolean query must be as effective as possible.

One technique that information specialists may use to assist in creating effective queries is to use 'seed studies': exemplar documents known to the information specialists a priori and often provided by the researchers of the systematic review. Information specialists use seed studies in various ways: identifying potential terms or phrases to use in the Boolean query; validating their search by analysing the precision and recall of their search against seed studies. Although, as we will demonstrate in the following sections, seed studies may not be a good indicator for query effectiveness.

The techniques that information specialists use to design and validate their search using seed studies have not gone unnoticed by the Information Retrieval community. There have been several studies that have exploited seed studies for automatic query formulation [35] and screening prioritisation (i.e., ranking the set of retrieved studies) [23, 42]. However, one major limitation of these previous studies is that the collections that they used *did not contain any seed studies*. Instead, a small portion of the final relevant studies were held out as a set of 'pseudo seed studies'. Given that these works use studies that are known to be relevant and included in a systematic review, we believe that the effectiveness of these methods is overestimated. However, methods that use pseudo seed studies are relying of documents that are known to be relevant and included in a systematic review; we posit and show that the effectiveness is thus overestimated. To address this issue, this resource paper presents a new Information Retrieval test collection for systematic review literature search; our novel contribution is that our collection includes real seed studies for each systematic review topic.

Our test collection enables the development of new Information Retrieval methods for systematic review literature search that depend on or utilise seed studies. To this end, we also investigate the impact of seed studies versus pseudo seed studies by reproducing two prominent methods from the literature (automatic query formulation and seed-driven document ranking). In addition, we demonstrate a new technique that can use seed studies which is often actually utilised when constructing a systematic review but overlooked in Information Retrieval methods: snowballing (otherwise known as citation chasing). This method is discussed in more detail in the sections that follow. We make our test collection and the results of all of our experiments and analysis available at github.com/ielab/sysrev-seed-collection.

2 BACKGROUND & RELATED WORK

2.1 Systematic Review Creation

Creating a systematic review is a highly time-consuming, costly and complex task, which involves numerous stages [4]. Information Retrieval methods have been investigated in an attempt to support and accelerate the creation of systematic reviews, especially for the tasks of query formulation [3, 33, 35–37, 39, 41] and screening prioritisation [11, 12, 23, 24, 30, 34, 40, 42, 44].

One of the initial stages in the systematic review process is searching and screening medical literature. This stage contributes the most to the costs of a systematic review [26], and it is at this stage that our collection is intended to be used. Figure 1 provides a high-level overview of the early stages of the systematic review process, and the possible uses cases or tasks that our collection enables. The figure shows that information specialists first use seed studies to create a Boolean query. The most common use for seed studies is identifying and extracting terms that are topically related to the systematic review. The Boolean query is then executed to retrieve a set of studies for manually screening by the researchers of the systematic review. In reality, there are two phases of screening: one at the abstract level, which we have described here; and one at the full text level, where the full-text of the abstract level studies are assessed for inclusion in the systematic review. In our collection, the data that we were provided did not contain the abstract level

relevance information of studies, nor the studies that were marked as not-relevant at an abstract level and full text level. We will discuss the implications of this in more detail in the sections that follow. The stages that follow in the systematic review creation process (e.g., data extraction or study synthesis) are not relevant to this collection, but are impacted by the methods that are created using our methods. For example, one of the use-cases of our collection, screening prioritisation (i.e., ranking studies), can be used to begin the following stages earlier.

2.2 Seed Studies

Seed studies are used extensively by information specialists for query formulation and validation in systematic review literature search [6, 15]. The use of seed studies is somewhat related to various explicit relevance feedback mechanisms, e.g. Rocchio [31]. In the Information Retrieval domain, explicit relevance feedback has been used in applications such as query expansion [8, 20, 22], active learning [28], and user modelling [16].

Active learning using explicit relevance feedback has been extensively studied in the area of technology assisted reviews [1, 9–11, 21, 23, 43, 44]: a catch-all phrase for high-recall tasks such as systematic review literature search, legal search, patent search, among others. The difference between explicit relevance feedback and seed studies is subtle but important. Seed studies cannot be considered relevant because they (1) have not necessarily been assessed as such when provided to an information specialist; and (2) may not be studies suitable for a systematic review (i.e., systematic review typically synthesise the results of several *randomised controlled trials*).[1] Therefore, although seed studies may be used in the place of explicit relevance feedback documents for certain applications such as query expansion, they should not be considered of equal weight in terms of relevance to a systematic review.
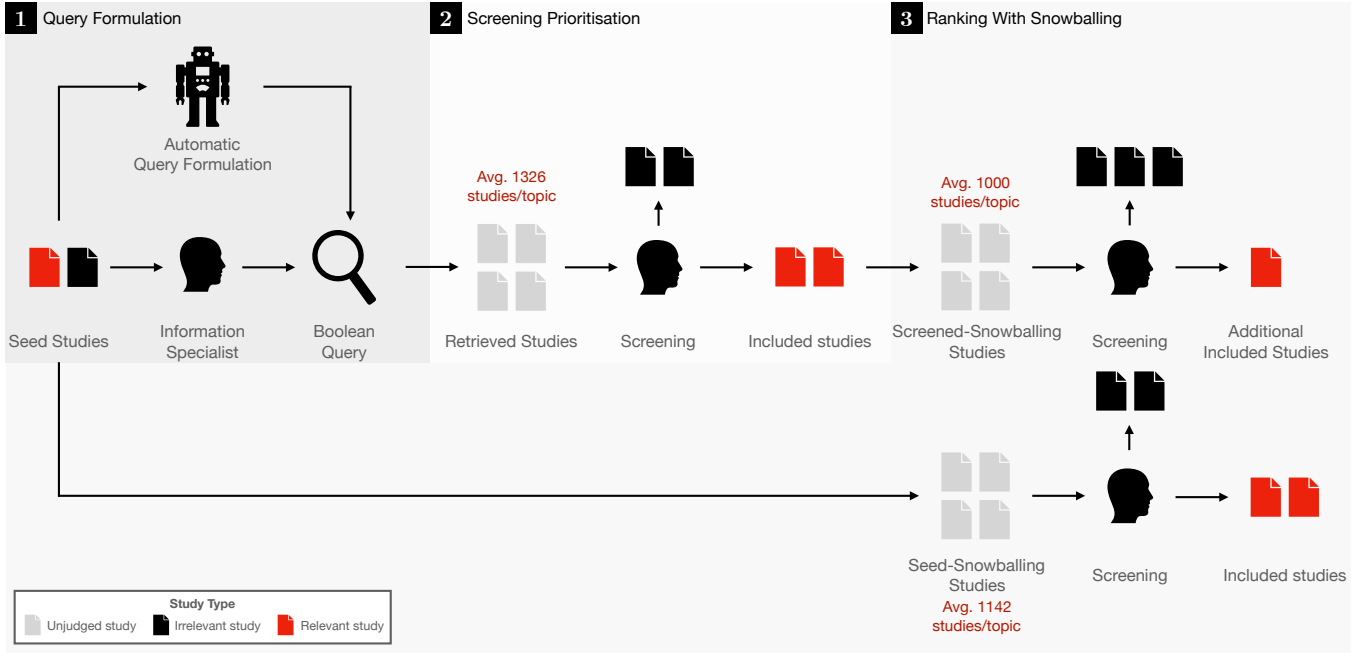
2.3 IR Systematic Review Collections

There are several test collections for research in systematic review literature search. These include i) the CLEF Technology Assisted Review collections [17–19]; ii) a similar test collection from Scells et al. [38]; iii) a test collection about systematic review updates [2]; and iv) collections for tasks such as data extraction and summarisation [27]. However, as we have raised earlier, none of these collections contain seed studies. This means that the true effectiveness is unknown of any existing methods that claim to use seed studies. Our collection enables the proper evaluation of existing and future methods that use seed studies.

2.4 Snowballing

In order to maximise comprehensiveness, one additional step in the systematic review creation process that is currently unexplored by current Information Retrieval methods is that of snowballing or citation chasing [5]. Snowballing identifies new studies to screen from the references or citations of other relevant studies. Snowballing is often performed 'backwards' and 'forwards' over the references to identify all studies that cite a given study and all studies cited by a given study. Although commonly performed in practise, there are very few studies in the literature that explore the topic of snowballing [29]. Nevertheless, snowballing has been

---

[1]Randomised controlled trials with many patients are considered a higher form of evidence compared to studies like case reports which may pertain to a single patient.

Figure 1: High-level overview of the systematic review creation processes that are relevant to our collection, and several use cases of our collection for automating these processes. Also shown are the three use-cases we will demonstrate for our collection: 1 query formulation, 2 screening prioritisation, and 3 ranking with snowballing.

shown to contribute 51% of the included studies for a systematic review [13]. The reason we investigate snowballing is to provide a more complete set of relevance assessments. Our preliminary experiments showed that not all included studies were retrieved by the Boolean query. We obtained total recall for many of our topics only when snowballing. Therefore, we perform snowballing to collect a more realistic set of studies that would be screened by researchers of a systematic review.

3 COLLECTION DETAILS

3.1 Topic Attributes

The basis of our collection comes from our co-author at the Bond Institute for Evidence-Based Healthcare, who is a senior information specialist. Each topic is a systematic reviews created using Boolean queries developed by our co-author over the past five years. We were provided unstructured data that we organised into 40 topics.

Each topic in our collection has several attributes, as shown in Table 1. Firstly, we assign a unique ID to each systematic review in our collection. As our collection contains completed and published systematic reviews, we also include the title and the URL of the published review. We also include a short description of the search provided by our information specialist co-author.

Apart from these metadata attributes, each topic also contains information about the query. Specifically, the PubMed Boolean query, the date restrictions applied to the search (i.e., used to restrict which studies are retrieved based on the publication date), and the

Name	Description
ID	Unique ID of a topic.
Link to Review	Link (URL) to the published review.
Title	Title of the systematic review.
Description	Topical summary of the search.
Date restrictions	Date restriction used to retrieve studies.
PubMed query	PubMed query used to retrieve studies.
Seed studies	PMIDs of seed studies.
Included studies	PMIDs of included studies.
Retrieved studies	PMIDs of studies retrieved by the query.
Snowballed studies	PMIDs of snowballed studies.

Table 1: Attributes of each topic in our collection. PMID refers to 'PubMed identifier', and is used to uniquely refer to a study or document in the PubMed database.

seed studies used to assist in the development of the query. Note that it is common to use multiple search queries across several medical literature databases to maximise recall. However, most of these databases require a monetary subscription to access. Therefore, we choose to only include the PubMed query in our collection. This is a limitation of all collections listed in Section 2.3, including ours.

The next set of attributes corresponds to the relevance assessments. The included studies represent studies that are relevant at the full-text level. This means that these studies were assessed as sufficient to be included in the final review. As such, it should be noted that evaluation of an Information Retrieval system using

our collection is more 'strict' (i.e., it is naturally more challenging to identify those studies that will be included in the final systematic review than those that are potentially relevant at an abstract level). Additionally, we provide the set of retrieved studies that were retrieved by the Boolean query.

The final attribute of each topic corresponds to snowballed studies. We include two sets of snowballed studies for each topic: one corresponding to snowballing the seed studies and another corresponding to snowballing the retrieved included studies. The second set of snowballed studies corresponds to simulating the process researchers of a systematic would do in order to identify additional relevant studies after assessing the set of retrieved studies. Note that this is an overlooked process in Information Retrieval methods in this space. We take this opportunity to do an initial investigation of the impact snowballing has on retrieval effectiveness and integrate the comparison of snowballing seed studies into the investigation.

3.2 Data Processing

Several attributes of topics required data processing beyond extraction. Specifically, the included studies, the retrieved studies, and the snowballed studies required additional processing to generate.

3.2.1 Included studies. The raw data that we were provided did not include any relevance assessment information. As such, we were required to create the relevance assessments for each topic. To do this we manually extracted the list of included studies that were used in the analysis of the systematic review. Note that this was not as simple as scraping the references of a systematic review, but involved manually matching the citations used in the analysis to references of published studies. Sometimes an included study is not available in PubMed. Here, we do not include these studies in our relevance assessments as we assume it is not retrievable.

3.2.2 Retrieved studies. The raw data that we were provided also did not include the studies that were retrieved by the Boolean query. As such, we reproduced the search for all topics to obtain a set of retrieved studies. We automated this step by using the Entrez API [25]. The date restrictions were applied to each Boolean query, therefore others should be capable of reproducing the set of retrieved documents for each topic if necessary.

3.2.3 Snowballed studies. Rather than manually snowballing studies, we use two tools: Citationchaser [14] and SpiderCite [7]. We first use CitationChaser to format a list of studies into the specific format required by SpiderCite.[2] Using SpiderCite, we obtain the list of cited and citing studies given the input set. Each study in SpiderCite has a DOI that we use to retrieve the study from PubMed.

3.3 Collection Statistics & Analysis

3.3.1 Seed and included studies analysis. On average, 15 seed studies are used for query construction, with a median number of 12. The average number of included studies per systematic review is 29, with a median number of 11.5. When comparing the overlap of seed studies and included studies, we found that, on average, only 36.3% of seed studies are contained in the included studies, and make up 26.1% of all included studies. This finding suggests that even though

seed studies are helpful for Boolean query construction, most are disregarded after the query construction phase. This demonstrates that treating included studies as pseudo seed studies as in prior work [23, 35] may lead to inaccurate results.

3.3.2 Searching analysis. When using the Boolean queries and date restrictions to retrieve studies, the mean number of retrieved documents per query is approximately 1,326, with median 709. Across several topics, we also found that not all included studies could be found in the retrieved studies. We found that only 75.5% of all included studies across all topics can be found inside the retrieved studies. There are two possible reasons why some included studies can not be found in candidate documents. The first is that some included studies may have been identified through snowballing. The second is that some studies may have been identified by searching in a database other than PubMed (although such studies may still exist in PubMed).

However, topic #18 does not contain any included studies in the retrieved studies. This particular systematic review only has a single included study. For topic #18, the reason that only one study is included is because many studies screened as relevant at the abstract level encountered a high risk of bias in the full-text.

3.3.3 Snowballing analysis. Our collection includes two snowballed sets of studies for each topic. The first set corresponds to snowballed seed studies (**seed-snowballing**) and the other corresponds to simulated snowballing of the included studies in the retrieved set (**screened-snowballing**).

For the seed-snowballing set, we find that 35 topics retrieved at least one included study. The topics that did not retrieve any included studies using this snowballing technique were 46, 52, 53, 66, and 96. The average number of documents snowballed per query is 1142, generally smaller than that retrieved from the searched documents list, as the number of seed studies used are smaller.

For the screened-snowballing set, only 34 topics retrieve at least one additional included study. The six topics that did not retrieve any additional included studies already contained all included studies prior to snowballing. These topics were 7, 10, 17, 39, 64, and 66. These topics represent an interesting research problem: applying snowballing to these topics would have resulted in wasted time and effort by the researchers of the systematic review. We leave investigations into determining whether or not to apply snowballing for future work. After removing already screened studies, the screened-snowballing sets contain, on average, 1000 studies.

3.3.4 Effectiveness comparison of retrieval methods. Finally, we aim to investigate the effectiveness of different search methods when used alone and when combined. The results of this analysis are presented in Figure 2. These plots show that using the seed studies alone (i.e., no searching and no snowballing) achieves the lowest recall but the highest precision. Snowballing the seed studies does increase recall but dramatically lowers precision.

Many topics retrieve almost all included studies with the Boolean query. Yet, combining the retrieved studies with other methods further improves recall while lowering precision. Combining both snowballing methods with the retrieved studies obtains the highest recall and lowest precision. However, total recall is still not achieved. Thus, we analysed the recall for different queries in Figure 3.

---

[2]The format is RIS; a standardised tag format that enables citation programs to exchange data, and the only format SpiderCite supports as input.
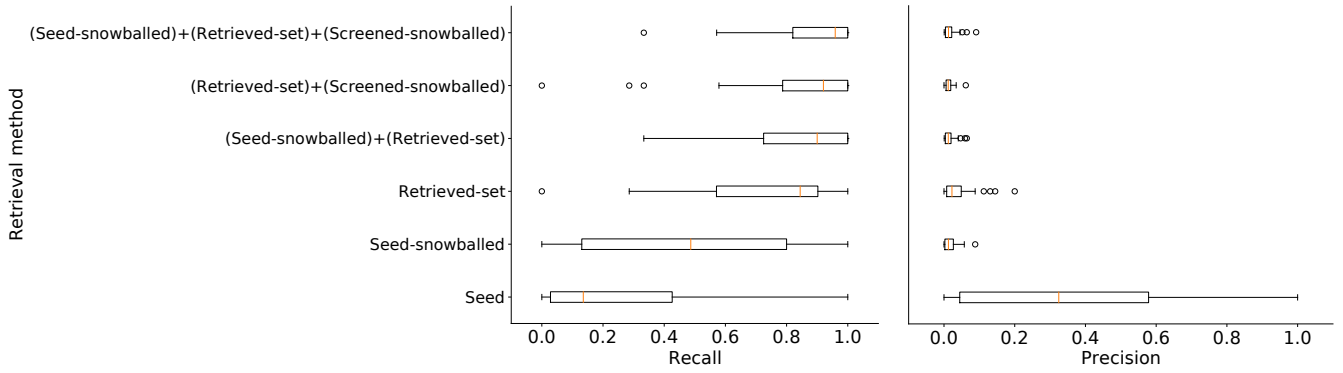
Figure 2: Recall and precision box plots of included studies for different retrieval methods. The methods listed include seed studies (seed), the Boolean query results (retrieved set), the two snowballing sets (seed-snowballed, i.e., snowballing applied to seed studies; and screened-snowballed, i.e., snowballing applied to retrieved included studies). These sets of studies are also combined in numerous ways, as indicated by '+'.
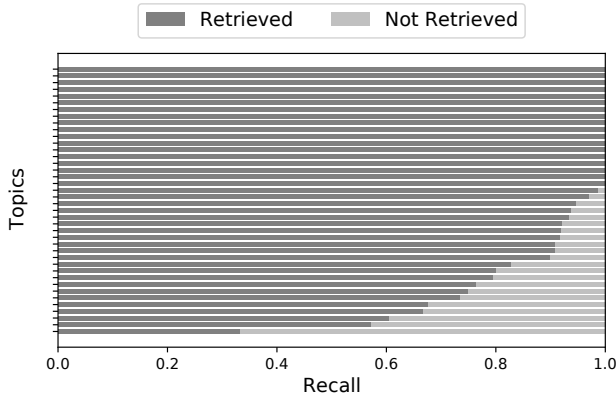
Figure 3: Recall distribution of all of the topics given the combined set of studies that includes retrieved studies, the seed-snowballing set, and the screened-snowballing set.

We found that for some topics, the recall is unusually low. We investigated these topics and found that these systematic reviews used several medical literature databases other than PubMed to retrieve studies. This results in the inclusion of studies that exist in the PubMed database, but are not retrieved by the PubMed query. We also randomly chose some topics reaching total recall and found that even though some of them still use a combination of multiple medical literature databases, they all used PubMed as one of the search sources. Thus, it is possible to create a more effective query for these topics. We leave such a problem for future work.

4 QUERY FORMULATION

We begin our demonstration of the use-cases of our collection with a reproduction of two automatic query formulation methods that use seed studies, see **1** in Figure 1. We use the implementations of Scells et al. [35]. We compare both automatic query formulation methods when using seed studies and pseudo seed studies.

4.1 Methods & Experimental Settings

The query formulation experiments are based on two existing methods from the literature [36, 37]. These methods are fully automated adaptations of manual or semi-automated procedures that information specialists use in practice. The first is called the *conceptual method* and is what the majority of information specialists use to formulate queries [6]. The automated conceptual method [36] takes as input a preliminary string for identifying salient terms (we use the title of the systematic review). The seed studies are then used to optimise the coverage of different combinations of terms expanded from those in the title. The second is called the *objective method* and is a more recent procedure that takes a statistical approach to query formulation [15]. At a high level, the automated objective method [37] first identifies and ranks salient terms from seed studies using term frequency statistics of the seed studies and a background collection. Next, terms are filtered and added to Boolean clauses by tuning these statistics to a held-out portion of seed studies. Given that the objective method relies on a held-out portion and the conceptual does not, we run both methods for three iterations using different arrangements of seed studies so that both methods use the same set of seed studies each iteration. We run these three iterations twice: once for the real seed studies and once for the pseudo seed studies.

One other aspect of the automatic versions of the two query formulation methods is the notion of an *instantiation*. In other words, the inclusion or exclusion of different aspects of Boolean queries (e.g., including or excluding MeSH, or including or excluding phrases). To this end, we perform experiments for only the most effective instantiation of each method (Conceptual/Phrase, and Objective/Phrase/Recall/MeSH). We refer the reader to the original study [37] for a comprehensive description of all experimental settings and implementation details that we have used.)

4.2 Results & Analysis

The results of our automatic query formulation reproduction study using our collection are presented in Table 2. We report precision, recall, and average number of studies retrieved.

	Method	Precision	Recall	Avg. Retrieved
	Original queries	0.01748	0.73659	1,326
Objective	Pseudo	0.00005	0.23457	746,193
	Seed	0.00024	0.31659	806,760
	Pseudo (oracle precision)	0.00015	0.22151	346,023
	Seed (oracle precision)	0.00573	0.34188	779,851
	Pseudo (oracle recall)	0.00014	0.50142	1,550,669
	Seed (oracle recall)	0.00572	0.51923	1,209,792
Conceptual	Pseudo	0.00664	0.27273	433,113
	Seed	0.00093	0.26781	362,968
	Pseudo (oracle precision)	0.00682	0.20940	30,961
	Seed (oracle precision)	0.00203	0.29202	360,549
	Pseudo (oracle recall)	0.00657	0.37334	667,398
	Seed (oracle recall)	0.00183	0.41382	848,223

Table 2: Effectiveness of queries formulated using pseudo seed studies (Pseudo) and seed studies (Seed). Also included are retrieval results of the queries for each topic (Original queries). Oracle experiments and the evaluation measures optimised can be seen in brackets.

4.2.1 Comparison of seed studies and pseudo seed studies. We first investigate the differences between using seed studies and pseudo seed studies for the objective method. Using real seed studies to formulate queries is more effective (precision and recall) than using pseudo seed studies for the objective method. One possible reason seed studies make better queries is that the pseudo seed studies may have been identified through snowballing (and therefore never originally retrieved by terms that exist in the Boolean query).

For the conceptual method, the result indicate that seed studies produce less effective queries than pseudo seed studies. We observe a noticeable decrease in precision and small decrease in recall between the seed studies and pseudo seed studies. However, while all queries using the objective method retrieve at least one study across all three iterations, queries from 17 topics constructed using the conceptual method do not retrieve any studies whatsoever. These results suggest that the automatic conceptual method generally produces less effective queries then the objective method (regardless of the use of seed or pseudo seed studies).

4.2.2 Comparison of oracle-setting based results. Given the possible issues with query formulation, and that certain (pseudo) seed studies may impact the effectiveness of resulting queries, we also investigate an oracle approach to identifying the most effective queries across the three iterations. The oracle process can be thought of as simulating the selection of an effective query by choosing from a list of three possible candidate queries.

When using the oracle query, the results show a similar trend as the results obtained from the objective method above: queries constructed using seed studies achieve a higher precision and recall than pseudo seed studies. The conceptual method shows that even though using seed studies still results in a lower precision, the

recall is higher than when pseudo seed studies are used. Again, as 17 topics across all three iterations do not retrieve any studies; this outcome may be biased to those topics that retrieved more than one study.

4.2.3 Comparison to original queries. Finally, we compare the results of the two automatic query formulation methods to the retrieval results of the queries for each topic in our collection. The original queries are highly effective, achieving dramatically higher precision and recall results. The average number of studies retrieved is also considerably lower than the automatic methods. We leave the development of more effective queries for systematic review literature search for future work.

4.3 Overall Findings

From the automatic query formulation experiments, we found that:

- For the objective method, higher effectiveness can be achieved when queries are constructed using seed studies. In fact, pseudo seed studies are detrimental to the query formulation procedure.
- For the conceptual method, queries may require manual modification to prevent biased results. We encountered similar results to Scells et al. [35] where several topics retrieved no studies. We leave further investixgations for future work.

5 SCREENING PRIORITISATION

We continue with another possible use-cases for this collection with a reproduction of a method that uses seed studies for screening prioritisation (i.e., ranking the set of retrieved studies), see 2 in Figure 1. The techniques was originally proposed by Lee and Sun [23] and we use the reproduced implementation provided by Wang et al. [42]. We investigate the effectiveness of seed-driven document ranking (SDR) methods, again comparing seed studies with pseudo seed studies. Note that as per previous studies, SDR refers to the specific process of screening prioritisation and a method of screening prioritisation called the SDR method. We make this distinction clear by referring to 'SDR' as the task and the 'SDR method' as the ranking function.

5.1 Methods & Experimental Setup

The SDR method proposed by Lee and Sun [23] is based on the observation that terms in relevant documents are more similar than terms of irrelevant documents. Thus, a ranking model is devised based on this observation which weights each term in the document based on the inter-study similarity. In the SDR method, a study's relevance score is calculated by the sum of pre-computed term weighting multiplied by the likelihood of the term to appear (calculated using the query likelihood model — QLM). In previous research, two study representation models have been explored:

Bag of words where all terms in a study are used.
Bag of clinical words where clinical terms in a study are used.

Bag of words representations are more effective than the bag of clinical words representation [42]. Therefore, we adopt the bag of words representation (as indicated by **BOW**) here. Adding to this we investigate a new aspect of SDR: seed studies versus pseudo seed studies.

	Method	MAP	P@10	P@100	P@1000	R@10	R@100	R@1000	nDCG@10	nDCG@100	nDCG@1000	LR%
Single Pseudo	BM25-BOW	0.1214†	0.1539†	0.0696†	0.0156	0.1226†	0.4049†	0.6855†	0.1846†	0.2760†	0.3764†	0.6915†
	QLM-BOW	0.1671	**0.2190**	0.0829	0.0168	**0.1646**	0.4572	0.7059	**0.2636**	0.3377	0.4291	0.5966
	SDR-BOW	0.1661	0.2165	0.0833	0.0168	0.1632	0.4638	0.7055	0.2599	0.3386	0.4286	0.6117
	AES-BOW	0.1501	0.1932	0.0807	0.0169	0.1361	0.4349	0.7022	0.2345	0.3141	0.4110	0.6365
	AES-BOW-P	0.1527	0.1965	0.0831	0.0170	0.1361	0.4455	0.7049	0.2349	0.3184	0.4134	0.6017
	SDR-BOW-AES	0.1698	0.2185	0.0864	0.0172	0.1573	0.4710	0.7060	0.2615	0.3437	0.4328	0.6007
	SDR-BOW-AES-P	**0.1709**	0.2165	**0.0877**	**0.0172**	0.1531	**0.4726**	**0.7114**	0.2596	**0.3446**	**0.4337**	**0.5820**
Multiple Pseudo	BM25-BOW	0.2045†	0.2253†	0.0712†	0.0154	0.1888†	0.4281†	0.6792	0.3145†	0.3817†	0.4758†	0.7073†
	QLM-BOW	0.3429	0.4629	0.1222	**0.0190**	0.2767	0.5531	**0.7251**	0.5934	0.5520	0.6174	**0.5255**
	SDR-BOW	**0.3457**	**0.4726**	**0.1231**	0.0189	**0.2798**	**0.5630**	0.7226	**0.6027**	**0.5574**	**0.6196**	0.5483
	AES-BOW	0.2731†	0.3415†	0.1059†	0.0183	0.2228†	0.5041†	0.7147	0.4466†	0.4727†	0.5529†	0.6156†
	AES-BOW-P	0.2810†	0.3488†	0.1094†	0.0185	0.2260†	0.5202†	0.7179	0.4546†	0.4834†	0.5598†	0.5743
	SDR-BOW-AES	0.3374	0.4464†	0.1201	0.0189	0.2666	0.5517	0.7211	0.5791†	0.5465†	0.6131	0.5672
	SDR-BOW-AES-P	0.3344	0.4341†	0.1200	0.0189	0.2634	0.5538	0.7235	0.5662†	0.5439	0.6103	0.5426
Seed Studies	BM25-BOW	0.1153†	0.1300†	0.0595†	0.0151	0.1067†	0.3404†	0.6333†	0.1516†	0.2375†	0.3439†	0.7591†
	QLM-BOW	**0.2294**	**0.2850**	0.0932	0.0169	0.2016	0.4614	0.6706	**0.3370**	**0.3867**	**0.4655**	**0.5749**
	SDR-BOW	0.2289	0.2800	**0.0945**	0.0169	**0.2028**	0.4695	0.6712	0.3253	0.3860	0.4619	0.6012
	AES-BOW	0.1784†	0.2250	0.0830†	0.0169	0.1644	0.4243	0.6657	0.2570	0.3276†	0.4184†	0.6390
	AES-BOW-P	0.1835†	0.2275	0.0860	0.0169	0.1614	0.4354	0.6667	0.2656	0.3381†	0.4248†	0.6008
	SDR-BOW-AES	0.2201	0.2650	0.0927	**0.0171**	0.1976	0.4615	**0.6724**	0.3151	0.3790	0.4581	0.5979
	SDR-BOW-AES-P	0.2198	0.2600	0.0912	0.0169	0.1972	0.4484	0.6680	0.3122	0.3732	0.4549	0.5750

Table 3: Results of baselines and SDR methods on our collection in three experimental settings: single pseudo seed studies, multiple pseudo seed studies, and seed studies. In the header, P refers to 'precision' and R refers to 'recall'. For AES methods, word2vec PubMed embeddings are denoted by '-P'. AES methods that do not have this demarcation correspond to word2vec embeddings, including PubMed and Wikipedia. Statistical significance (Student's two-tailed paired t-test with Bonferonni correction, $p < 0.05$) between SDR method and all other methods is indicated by †.

5.1.1 Single pseudo seed study. Firstly, we assume a single included study as an available seed study for each systematic review topic, as per Lee and Sun [23]. As there are multiple included studies corresponding to every systematic review topic, the overall effectiveness of retrieval methods on each topic is calculated using the average effectiveness of utilising every included study. Using this leave-one-out cross-validation strategy, the results gathered tend to be more reliable and unbiased across all included studies in a systematic review topic [32].[3]

5.1.2 Multiple pseudo seed studies. Using multiple pseudo seed studies is more effectiveness then using a single seed study for SDR [42]. For our collection, we also perform SDR with multiple pseudo seed studies. We adopt the experimental settings of Wang et al. [42] to evaluate the effectiveness of SDR when using multiple pseudo seed studies. We adopt the same seed study grouping strategy in which 20% of included studies are chosen using a sliding-window approach. The groups are then combined by concatenating their titles and abstracts to act as the input to the retrieval methods. The effectiveness on each topic is then calculated using the average effectiveness from all groups.

5.1.3 Seed studies. Using the seed studies in our collection, we can now realistically investigate the effectiveness of SDR. This experiment combines all the real seed studies by concatenating their titles and abstracts, as used in the multiple pseudo seed study

experiments. The combined studies then act as an input for the SDR method and all of the included studies are used for evaluation.

5.1.4 Retrieval Methods. Apart from the original SDR method proposed by Lee and Sun [23], we also perform experiments using several baselines: BM25, a query likelihood model (QLM), and a word embedding-based model (AES). As in the previous SDR papers, we used two pre-trained word embeddings for the AES method: one trained on PubMed and Wikipedia and one trained on only PubMed. Additionally, we also include fusion methods used in the original paper to interpolate the SDR method with AES using the same parameters in the original article ($\alpha = 0.3$).

5.1.5 Evaluation Measures. We evaluate the different arrangements of seed studies and methods with rank-based measures. We use the same evaluation measures as in our reproducibility study [42]. In addition to MAP, which measures the ranking effectiveness of the entire list of studies, we also measure precision, recall, and nDCG at different cut-offs — {10,100,1000}, and Last Relevant% (LR%), which reports the percentage of studies that must be screened in order to identify all included studies.

5.2 Results & Analysis

The results of using a single pseudo seed study, multiple pseudo seed studies, and seed studies are shown in Table 3.

5.2.1 Single and multiple pseudo seed studies. For single pseudo seed studies, the SDR method improves effectiveness for deep evaluation metrics (i.e., precision@{100,1000}, recall@{100,1000}, and

---

[3]Topic #18 only has one included study. We disregard this topic in the evaluation across all experiments.

nDCG@{100,1000}). SDR-BOW-AES-P is the most effective, apart from shallow evaluation measures (i.e., precision @10, recall@10 and nDCG@10.). For shallow evaluation measures, QLM achieves the highest effectiveness. This may be because the word embeddings based measures help alleviate vocabulary mismatch, improving recall but at the expense of early precision.

For multiple pseudo seed studies, SDR is the most effective on all evaluation measures except precision@1000, recall@1000, and LR%. Although, it is interesting to observe that the fused method is not able to achieve higher effectiveness than using a single retrieval method alone. This may be due to the poor performance of the AES method, and remains an interesting future challenge for automatically determining when to apply fusion. Using multiple pseudo seed studies is always more effective than single seed studies.

All the aforementioned results are in line with our previous reproduction study Wang et al. [42], with the exception of the finding here that result fusion did not increase effectiveness.

5.2.2 Comparison of pseudo-seed results and real-seed results. Finally, we found that using real seed studies verses multiple pseudo seed studies dramatically impacts effectiveness. Firstly, even though the effectiveness of using seed studies still outperforms the use of a single pseudo seed study in some methods, the effectiveness of seed studies is significantly worse than using multiple pseudo seed studies. One possible explanation is that seed studies are not always relevant to the systematic review topic (pseudo seed studies are by definition). Using a non-relevant seed study could degrade the quality of results. This effect is highlighted in Figure 4, where the effectiveness of using all seed studies is closer to using a single pseudo study than it is to using multiple pseudo seed studies.

Secondly, using multiple pseudo seed retrieval with the SDR and SDR fused methods consistently outperforms other retrieval methods. However, when using real seed studies, the QLM method is most effective. Our explanation is that while the SDR method still captures the semantic meaning of terms through term weighting, one common use of seed studies in systematic review creation is term extraction, which is better represented by methods like QLM.

In conclusion, pseudo seed studies are not representative of real seed studies and this impacts seed-drive retrieval (SDR). That is why it is important to have test collections with real seed studies such as the one provided in this paper.

5.3 Overall Findings

From the SDR experiment, we found that:

- Using pseudo seed studies produces unrealistic results compared to using seed studies (i.e., the results are higher than if one used seed studies). This is likely a result of the fact that the pseudo seed studies are an example of explicit relevance feedback, whereas seed studies are more akin to pseudo relevance feedback.
- The choice of ranking models in this task (e.g., BM25, QLM, SDR, AES) is dependent on the kind of seed studies used. This may be due to how terms are weighted (i.e., relevant terms are likely to appear in pseudo seed studies but are less likely to appear in seed studies).
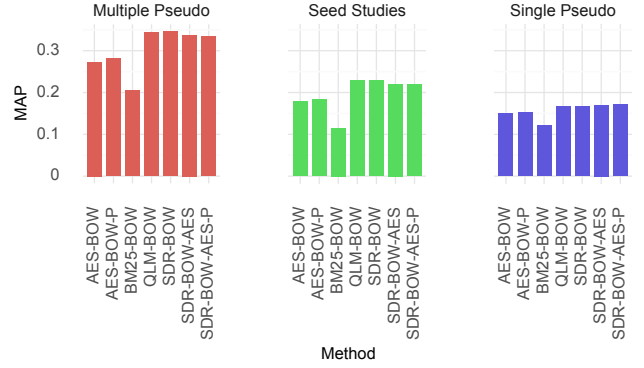
Figure 4: Comparison between the different arrangements of seed and pseudo-seed studies across the SDR methods. We have chosen to show only MAP, but all evaluation measures we chose showed the same trend.
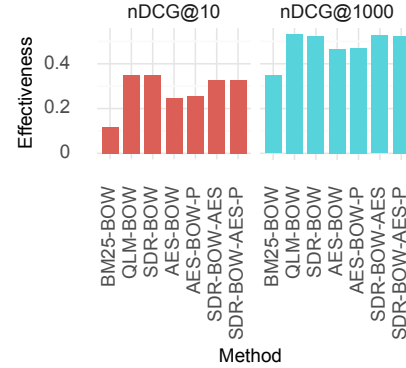
Figure 5: Comparison between nDCG@10 and nDCG@1000 for the results in Table 4.

6 RANKING WITH SNOWBALLING

Our last demonstration of the use-cases of our collection is a new technique that we have devised specifically for this paper: ranking with snowballing, see ③ in Figure 1. Our collection provides two snowballing sets for each topic. Using seed studies and the two snowballing sets, we investigate the impact of snowballing when using SDR. We investigate two use cases of snowballing within the context of SDR:

(1) The effectiveness of SDR on a combined set of seed-snowballing studies and retrieved studies.
(2) The effectiveness of SDR on the screened-snowballing set.

These two use cases demonstrate (1) the effectiveness of ranking combined seed-snowballing and retrieved studies sets (i.e., integrating the seed-snowballing set into the retrieved studies and then ranking); and (2) the effectiveness of ranking post-screening (i.e., simulating the ranking of the screened-snowballing set using the screened retrieved studies as input).

	Method	MAP	P@10	P@100	P@1000	R@10	R@100	R@1000	nDCG@10	nDCG@100	nDCG@1000	LR%
seed-snowball+search	BM25-BOW	$0.0936^{\dagger}$	$0.0949^{\dagger}$	$0.0421^{\dagger}$	0.0137	$0.0943^{\dagger}$	$0.3372^{\dagger}$	$0.7767^{\dagger}$	$0.1175^{\dagger}$	$0.2074^{\dagger}$	$0.3469^{\dagger}$	$0.7389^{\dagger}$
	QLM-BOW	**0.2672**	0.2744	0.0813	0.0163	0.2453	**0.5483**	0.8612	0.3490	**0.4269**	**0.5326**	$0.5236^{\dagger}$
	SDR-BOW	0.2663	**0.2897**	**0.0815**	0.0159	**0.2592**	0.5390	0.8529	**0.3503**	0.4212	0.5244	0.5672
	AES-BOW	$0.1886^{\dagger}$	$0.1949^{\dagger}$	0.0690	**0.0165**	$0.1815^{\dagger}$	$0.4619^{\dagger}$	0.8595	$0.2466^{\dagger}$	$0.3366^{\dagger}$	$0.4668^{\dagger}$	0.5921
	AES-BOW-P	$0.1950^{\dagger}$	$0.2026^{\dagger}$	0.0700	0.0162	$0.1906^{\dagger}$	0.4755	0.8530	$0.2548^{\dagger}$	$0.3459^{\dagger}$	$0.4700^{\dagger}$	0.5514
	SDR-BOW-AES	0.2572	$0.2513^{\dagger}$	0.0787	0.0164	$0.2297^{\dagger}$	0.5210	**0.8675**	$0.3271^{\dagger}$	0.4136	0.5271	0.5521
	SDR-BOW-AES-P	0.2546	$0.2487^{\dagger}$	0.0779	0.0164	0.2343	0.5224	0.8666	0.3266	0.4121	0.5245	0.5298

Table 4: Results of SDR methods on our collection when the seed-snowballing set and retrieved studies are combined. Denotations are identical to those in the caption of Table 3.

	Method	MAP	Prec @100	Recall @100	nDCG @100	LR%
screened-snowball	BM25-BOW	$0.0208^{\dagger}$	0.0100	0.2045	$0.0707^{\dagger}$	$0.4606^{\dagger}$
	QLM-BOW	**0.1279**	0.0212	0.3090	**0.2104**	0.3317
	SDR-BOW	0.0921	0.0218	0.3055	0.1829	0.3514
	AES-BOW	0.0474	0.0218	0.3167	0.1436	0.3412
	AES-BOW-P	0.0602	0.0206	0.3111	0.1530	0.3329
	SDR-BOW-AES	0.0898	**0.0226**	**0.3199**	0.1858	0.3257
	SDR-BOW-AES-P	0.0871	0.0218	0.3149	0.1815	**0.3214**

Table 5: Results of SDR methods on our collection when using the screened-snowballing set. Denotations are identical to those in the caption of Table 3.
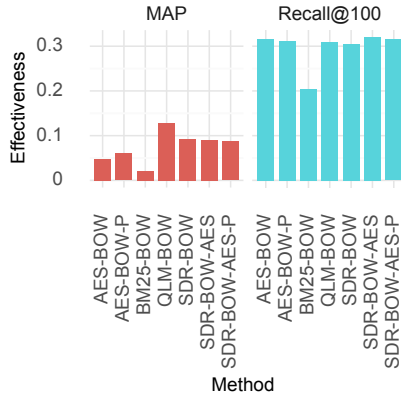
Figure 6: Comparison between MAP and recall@100 for the results in Table 5.

6.1 Methods & Experimental Setup

6.1.1 Combined seed-snowballing and retrieved studies ranking. In Section 5 we investigated the effectiveness of SDR using different arrangements of seed studies. We found that when real seed studies are used, the QLM method outperforms the SDR method in almost all evaluation measures. In this experiment, we investigate if similar results arise when the seed-snowballing set is combined with retrieved studies. As the first experiment, we used the same experimental setting as in Section 5.1.3. We first combine the retrieved studies and the seed-snowballing set. Next, we concatenate titles and abstracts of seed studies as input to the ranking methods. We report the same evaluation measures as described in Section 5.1.5.

6.1.2 Screened snowballing document ranking. As the second experiment, we simulate the process of screening prioritisation for the screened-snowballing set. Included studies in the retrieved set plus the seed studies are used as input for SDR. Given the difference in performance of the two sets of studies (i.e., retrieved included studies and seed studies), one could consider different term weighting functions depending on the study. However, we leave such investigation for future work. We perform our evaluation on the included studies that did not appear in the retrieved set of studies. This means that the results for this experiment are not comparable to the results of the other screening prioritisation experiments that appear earlier in this paper. We report a subset of the evaluation measures as described in Section 5.1.5.

6.2 Results & Analysis

6.2.1 Combined seed-snowballing and retrieved studies ranking. Results for this experiment are shown in Table 4. The most effective methods are QLM and SDR, with SDR more effectiveness on shallow measures and more effectiveness on deeper measures. These results are highlighted further in Figure 5. When comparing these combined results with the seed studies result from Table 3, we found that all methods can achieve higher recall{@100,@1000} and LR%, which suggests that adding the seed snowballing set can significantly increase the number of relevant studies retrieved.

Despite the fact that there are more studies overall when combining the seed-snowballing set with retrieved studies, it is worth it in terms of the overall improvement in effectiveness. In practise, it is beneficial to combine the results of seed-snowballing with retrieved studies for screening prioritisation.

6.2.2 Screened snowballing document ranking. Results for this experiment are shown in Table 5. The two best performing methods here were QLM and the AES methods. QLM is the most effective for MAP and nDCG@100. Meanwhile, fusion methods are the most effective for precision@100, recall@100, and LR%. These results are highlighted in Figure 6. These findings demonstrate that QLM is effective for the SDR task when ranking screened-snowballing studies.

6.3 Overall Findings

From the ranking with snowballing experiments, we found that:

- Combining the seed-snowballing set and retrieved studies is beneficial for SDR.
- QLM and AES are best for screened-snowballing document ranking. However, there is room for future work in determining the optimal combination of studies to use for SDR.

7 CONCLUSION

We present a new test collection to properly evaluate systematic review literature search methods which uses seed studies. Along with the test collection, we provided a detailed analysis of this resource. Firstly, we found that only a small portion of seed studies are in effect included in a systematic review. This is because not all seed studies provided to an information specialist satisfy the inclusion criteria for the systematic review: some seed studies are provided for a broad understanding of the topic of the review. Thus, the use of pseudo seed studies drawn from the included studies does not resemble all the properties of actual seed studies used in practice. Secondly, we also investigated the impact of seed studies by reproducing two existing methods that use seed studies [23, 36], but that were only evaluated using pseudo seed studies (a sample from the included studies). Our experiments show that using pseudo seed studies overestimates the effectiveness of these methods.

The test collection also enables an investigation of the difference between snowballing seed studies and screened studies. Here, we found that ranking a combined lists of studies from seed-snowballing documents and searched candidate documents may further boost the effectiveness of ranking models for the task of screening prioritisation.

The test collection makes two important contributions:

(1) it enables a considerably more realistic evaluation of methods that use seed studies, and now one can use all included studies for relevance assessments, instead of requiring a held-out portion as pseudo seed studies; and

(2) it provides realistic data to develop or train new methods that use seed studies.

Seed studies are vital for effective query formulation for information specialists and are commonly used. Despite this, many existing Information Retrieval test collections for systematic review literature search do not contain seed studies. Our test collection will promote the development and realistic evaluation of methods that seed studies can be used to improve systematic review literature search. This includes methods we already explored in this paper like screening prioritisation [23, 42] and query formulation [36], and some we leave for future works such as active learning [10] or MeSH term suggestion [41]. Such methods can have considerable real-world impacts, as systematic reviews are highly time consuming and costly. Cheaper and faster systematic reviews can have dramatic implications for patient outcomes and institutional policy decisions that affect the health decisions of entire countries.

8 ACKNOWLEDGMENT

REFERENCES

[1] Mustafa Abualsaud, Nimesh Ghelani, Haotian Zhang, Mark D Smucker, Gordon V Cormack, and Maura R Grossman. 2018. A system for efficient high-recall retrieval. In *The 41st international ACM SIGIR conference on research & development in information retrieval*. 1317–1320.

[2] Amal Alharbi and Mark Stevenson. 2019. A dataset of systematic review updates. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1257–1260.

[3] Amal Alharbi and Mark Stevenson. 2020. Refining Boolean queries to identify relevant studies for systematic review updates. *Journal of the American Medical Informatics Association* 27, 11 (2020), 1658–1666.

[4] Krystal Bullers, Allison M Howard, Ardis Hanson, William D Kearns, John J Orriola, Randall L Polo, and Kristen A Sakmar. 2018. It Takes Longer than You Think: Librarian Time Spent on Systematic Review Tasks. *Journal of the Medical Library Association* 106, 2 (2018), 198.

[5] Miew Keen Choong, Filippo Galgani, Adam G Dunn, and Guy Tsafnat. 2014. Automatic evidence retrieval for systematic reviews. *Journal of medical Internet research* 16, 10 (2014), e3369.

[6] Justin Clark. 2013. Systematic Reviewing. In *Methods of Clinical Epidemiology*, Gail M. Williams Suhail A. R. Doi (Ed.).

[7] Justin Clark, Paul Glasziou, Chris Del Mar, Alexandra Bannach-Brown, Paulina Stehlik, and Anna Mae Scott. 2020. A full systematic review was completed in 2 weeks using automation tools: a case study. *Journal of Chronic Diseases* 121 (May 2020), 81–90. https://doi.org/10.1016/j.jclinepi.2020.01.008

[8] Francesco Colace, Massimo De Santo, Luca Greco, and Paolo Napoletano. 2011. Improving text retrieval accuracy by using a minimal relevance feedback. In *International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management*. Springer, 126–140.

[9] Gordon V Cormack and Maura R Grossman. 2014. Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 153–162.

[10] Gordon V Cormack and Maura R Grossman. 2015. Autonomy and reliability of continuous active learning for technology-assisted review. *arXiv preprint arXiv:1504.06868* (2015).

[11] Gordon V Cormack and Maura R Grossman. 2016. Scalability of continuous active learning for reliable high-recall text classification. In *Proceedings of the 25th ACM international on conference on information and knowledge management*. 1039–1048.

[12] Giorgio Maria Di Nunzio. 2020. A Study on a Stopping Strategy for Systematic Reviews Based on a Distributed Effort Approach. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 112–123.

[13] Trisha Greenhalgh and Richard Peacock. 2005. Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. *Bmj* 331, 7524 (2005), 1064–1065.

[14] Neal Haddaway, Matt Grainger, and Charles Gray. 2021. *citationchaser: An R package and Shiny app for forward and backward citations chasing in academic searching*. https://doi.org/10.5281/zenodo.4543513

[15] Elke Hausner, Siw Waffenschmidt, Thomas Kaiser, and Michael Simon. 2012. Routine Development of Objectively Derived Search Strategies. *Systematic reviews* 1, 1 (2012), 19.

[16] Sampath Jayarathna, Atish Patra, and Frank Shipman. 2015. Unified relevance feedback for multi-application user interest modeling. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*. 129–138.

[17] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. 2017. CLEF 2017 technologically assisted reviews in empirical medicine overview. In *CEUR workshop proceedings*, Vol. 1866. 1–29.

[18] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and René Spijker. 2018. CLEF 2018 Technologically Assisted Reviews in Empirical Medicine Overview. In *CLEF*.

[19] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. 2019. CLEF 2019 technology assisted reviews in empirical medicine overview. In *CEUR workshop proceedings*, Vol. 2380.

[20] Rianne Kaptein, Jaap Kamps, and Djoerd Hiemstra. 2008. *The impact of positive, negative and topical relevance feedback*. Technical Report. AMSTERDAM UNIV (NETHERLANDS).

[21] Athanasios Lagopoulos, Antonios Anagnostou, Adamantios Minas, and Grigorios Tsoumakas. 2018. Learning-to-rank and relevance feedback for literature appraisal in empirical medicine. In *International conference of the cross-language evaluation forum for European languages*. Springer, 52–63.

[22] Victor Lavrenko and W Bruce Croft. 2017. Relevance-based language models. In *ACM SIGIR Forum*, Vol. 51. ACM New York, NY, USA, 260–267.

[23] Grace E Lee and Aixin Sun. 2018. Seed-driven document ranking for systematic reviews in evidence-based medicine. In *The 41st international ACM SIGIR conference on research & development in information retrieval*. 455–464.

[24] Dan Li and Evangelos Kanoulas. 2020. When to stop reviewing in technology-assisted reviews: Sampling from an adaptive distribution to estimate residual

relevant documents. *ACM Transactions on Information Systems (TOIS)* 38, 4 (2020), 1–36.

[25] Donna Maglott, Jim Ostell, Kim D Pruitt, and Tatiana Tatusova. 2005. Entrez Gene: gene-centered information at NCBI. *Nucleic acids research* 33, suppl_1 (2005), D54–D58.

[26] Jessie McGowan and Margaret Sampson. 2005. Systematic Reviews Need Systematic Searchers (IRP). *Journal of the Medical Library Association* 93, 1 (2005), 74.

[27] Christopher Norman, Mariska Leeflang, and Aurélie Névéol. 2018. Data extraction and synthesis in systematic reviews of diagnostic test accuracy: a corpus for automating and evaluating the process. In *AMIA Annual Symposium Proceedings*, Vol. 2018. American Medical Informatics Association, 817.

[28] Mateus Pereira, Elham Etemad, and Fernando Paulovich. 2020. Iterative learning to rank from explicit relevance feedback. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*. 698–705.

[29] Animesh Prasad, Manpreet Kaur, and Min-Yen Kan. 2018. Neural ParsCit: A Deep Learning Based Reference String Parser. *Journal on Digital Libraries* 19 (2018), 323–337.

[30] Piotr Przybyła, Austin J Brockmeier, Georgios Kontonatsios, Marie-Annick Le Pogam, John McNaught, Erik von Elm, Kay Nolan, and Sophia Ananiadou. 2018. Prioritising references for systematic reviews with RobotAnalyst: a user study. *Research synthesis methods* 9, 3 (2018), 470–488.

[31] JJ Rocchio and Gerard Salton. 1965. Information search optimization and interactive retrieval techniques. In *Proceedings of the November 30–December 1, 1965, fall joint computer conference, part I*. 293–305.

[32] Claude Sammut and Geoffrey I. Webb (Eds.). 2010. *Leave-One-Out Cross-Validation*. Springer US, Boston, MA, 600–601. https://doi.org/10.1007/978-0-387-30164-8_469

[33] Harrisen Scells, Guido Zuccon, and Bevan Koopman. 2019. Automatic Boolean Query Refinement for Systematic Review Literature Search. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 1646–1656. https://doi.org/10.1145/3308558.3313544

[34] 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 *European conference on information retrieval*. Springer, 399–414.

[35] 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.

[36] Harrisen Scells, Guido Zuccon, Bevan Koopman, and Justin Clark. 2020. Automatic Boolean query formulation for systematic review literature search. In *Proceedings of The Web Conference 2020*. 1071–1081.

[37] Harrisen Scells, Guido Zuccon, Bevan Koopman, and Justin Clark. 2020. A computational approach for objectively derived systematic review search strategies. In *European conference on information retrieval*. Springer, 385–398.

[38] 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. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1237–1240.

[39] Harrisen Scells, Guido Zuccon, Mohamed A. Sharaf, and Bevan Koopman. 2020. *Sampling Query Variations for Learning to Rank to Improve Automatic Boolean Query Generation in Systematic Reviews*. Association for Computing Machinery, New York, NY, USA, 3041–3048. https://doi.org/10.1145/3366423.3380075

[40] Alison Sneyd and Mark Stevenson. 2021. Stopping Criteria for Technology Assisted Reviews based on Counting Processes. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2293–2297.

[41] Shuai Wang, Hang Li, Harrisen Scells, Daniel Locke, and Guido Zuccon. 2021. MeSH Term Suggestion for Systematic Review Literature Search. In *Proceedings of the 25th Australasian Document Computing Symposium* (Virtual Event, Australia) *(ADCS '21)*. Association for Computing Machinery, New York, NY, USA, Article 8, 8 pages. https://doi.org/10.1145/3503516.3503530

[42] Shuai Wang, Harrisen Scells, Ahmed Mourad, and Guido Zuccon. 2022. Seed-Driven Document Ranking for Systematic Reviews: A Reproducibility Study. In *European Conference on Information Retrieval*. Springer, 686–700.

[43] Haotian Zhang, Mustafa Abualsaud, Nimesh Ghelani, Mark D Smucker, Gordon V Cormack, and Maura R Grossman. 2018. Effective user interaction for high-recall retrieval: Less is more. In *Proceedings of the 27th ACM international conference on information and knowledge management*. 187–196.

[44] Jie Zou, Dan Li, and Evangelos Kanoulas. 2018. Technology assisted reviews: Finding the last few relevant documents by asking yes/no questions to reviewers. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 949–952.