A Test Collection for Evaluating Retrieval of Studies for Inclusion in Systematic Reviews

Harrisen Scells
Queensland University of Technology
Brisbane, Australia

Guido Zuccon
Queensland University of Technology
Brisbane, Australia

Bevan Koopman
Australian e-Health Research Centre, CSIRO
Brisbane, Australia

Anthony Deacon
Queensland University of Technology
Brisbane, Australia

Leif Azzopardi
Strathclyde University
Glasgow, Scotland (UK)

Shlomo Geva
Queensland University of Technology
Brisbane, Australia

ABSTRACT

This paper introduces a test collection for evaluating the effectiveness of different methods used to retrieve research studies for inclusion in systematic reviews. Systematic reviews appraise and synthesise studies that meet specific inclusion criteria. Systematic reviews intended for a biomedical science audience use boolean queries with many, often complex, search clauses to retrieve studies; these are then manually screened to determine eligibility for inclusion in the review. This process is expensive and time consuming. The development of systems that improve retrieval effectiveness will have an immediate impact by reducing the complexity and resources required for this process. Our test collection consists of approximately 26 million research studies extracted from the freely available MEDLINE database, 94 review (query) topics extracted from Cochrane systematic reviews, and corresponding relevance assessments. Tasks for which the collection can be used for information retrieval system evaluation are described and the use of the collection to evaluate common baselines within one such task is demonstrated. The test collection is available at https://github.com/ielab/SIGIR2017-PICO-Collection.

CCS CONCEPTS

•**Information systems → Test collections;**

1 INTRODUCTION

A systematic review is a type of literature review that appraises and synthesises the work of primary research studies to answer one or more research questions. Most authors follow the *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) method for conducting and reporting these reviews. This includes the definition of a formal search strategy to retrieve studies which are to be considered for inclusion in the review. Because of the use of a formal search strategy, systematic reviews are classified as the highest level of evidence by the NHMRC for answering diagnostic, prognostic, treatment and other clinical questions in medicine.[1] Hence, clinical medicine, medical research, and increasingly sectors outside of medicine, rely heavily on systematic reviews.

```
1. MMSE*.ti,ab.                  2. sMMSE.ti,ab.
3. Folstein*.ti,ab.              4. MiniMental.ti,ab.
5. "mini mental stat*".ti,ab.    6. or/1-5
```

Figure 1: Sample search strategy for a systematic review showing five boolean clauses (1–5) and the final query (6), which combines the previous clauses with a conjunctive `or`. The symbols `.ti,ab.` indicate that the title and abstract should be searched. The ⋆ symbols indicates wild card matching and terms in quotes indicate phrase matching.

Given a research question and a set of inclusion/exclusion criteria, researchers undertaking a systematic review define a search strategy (the query) to be issued to one or more search engines that index published literature (e.g. PubMed). In medical and biomedical research, search strategies are commonly expressed as (large) boolean queries[2]. Figure 1 shows an example search strategy for PubMed. After the search strategy has been executed, the title, and then abstract, of studies retrieved by it are reviewed in a process known as screening. Where the study appears relevant the full-text is then retrieved for more detailed examination.

The compilation of systematic reviews can take significant time and resources, hampering their effectiveness. Tsafnat et al. report that it can take several years to complete and publish a systematic review [15]. When systematic reviews take such significant time to complete, they can become out-of-date even at time of publishing. While the compilation of a systematic review involves several steps, one of the most time-consuming is screening. Previous work has reported that it can take experienced reviewers between 30 seconds and several minutes to screen a single study (title, abstract and metadata). The effect this has on the timeliness of reviews is highlighted by some notable examples; for example in Shemilt et al.'s *scoping* review, 1.8 million studies were screened, of which only about 4,000 were found to be potentially eligible [13]. Thus, the development of IR methods that have a high specificity (precision),

SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan

978-1-4503-5022-8/17/08…$15.00
DOI: http://dx.doi.org/10.1145/3077136.3080707

[1] National Health and Medical Research Council, Australia

[2] The use of boolean retrieval systems, rather than best-match/rank-based systems, is often motivated by the need for a deterministic result set that ensures reproducibility of the search over time. Rank-based methods, especially those using ever changing collection statistics (e.g., relevance feedback) do not allow for reproducible result sets.

while maintaining a high sensitivity (recall), would have a major impact on the time and resources required to undertake systematic reviews.

It is thus unsurprising that the challenge of compiling systematic reviews can be fertile ground for information retrieval (IR) research. As such, this paper presents a test collection for evaluating different IR methods aimed at improving the retrieval of primary studies for systematic reviews. The test collection consists of 94 search strategies (queries) for research questions with associated relevance assessments. The documents of the test collection are freely available MEDLINE studies, comprising title, abstract and metadata.

2 RELATED WORK

Automation using IR techniques may be beneficial to many of the steps for constructing a systematic review, including: the development of search strategies, screening of retrieved studies, and the clustering and thematic analysis of large sets of screened studies (also known as 'mapping') [11, 14]. We do not consider mapping below because our collection does not explicitly provide resources to evaluate mapping methods.

Development of search strategies: Currently, search strategies are developed by information specialists in an iterative process that requires preliminary searches on the collection to determine suitability of query keywords and approximate retrieval effectiveness. This process could be better supported using statistical indicators of search quality and suggesting related query keywords likely to improve retrieval effectiveness. For example, Karimi et al. [6] examined providing early indications of the quality of search strategies by returning ranked boolean results during the process of formulating search strategies.

Screening: This is the process by which the title and abstract of retrieved studies are reviewed to determine if it is worth examining the full text of the study in detail. Automation can be used to improve screening in a number of ways.

Better retrieval systems which decrease the number of false positives retrieved; i.e., increase precision while maintaining the same level of recall have been widely cited as an area for future development [3]. This would reduce the time spent screening. It is important to note that systematic reviews aim to achieve a high level of recall, thus guaranteeing that the review is comprehensive and minimising the impact of publication bias, among other reasons. Examples of prior work in this area include the work of Karimi et al. who applied query expansion in an attempt to improve recall [7].

Screening prioritisation could be introduced by ranking documents by likelihood of satisfying the inclusion criteria of the systematic reviews. This would allow relevant studies to be identified early on in the screening process, thus providing a feedback loop to improve the development of search strategies. (This method was successfully applied to a large scoping review by Shemilt et al.[13]) Screening prioritisation is typically done as a two-stage process. An initial set of studies are retrieved using a boolean retrieval process; these are then ranked according to some relevance measure.

Beyond screening prioritisation, automated classification methods can be applied at the second stage to filter, rather than simply rank, retrieved studies. A number of different classification methods have been developed, including support vector machine classification [4, 9], voting perceptron [5] and random forest [8]. More sophisticated systems combine both prioritisation via ranking and filtering via classification and provide significant work savings [9].

To minimise study selection bias and reduce human error, most systematic reviews have multiple reviewers screen each record. This is, however, resource intensive and time consuming, especially for large reviews. The use of classifiers to replace the need of a second screener has shown some promise [1].

The test collection described in this study provides the much needed resource to evaluate the aforementioned screening and prioritisation methods. Evaluation of semi-automated classification and automated double screening can also be supported by our test collection — although we do not explicitly consider these aspects.

Existing resources for evaluating retrieval systems aimed at improving the screening process are limited. Most resources are characterised by a small set of queries (search strategies), thus rendering findings about methods uncertain and subjected to experimental bias. Martinez et al. [9] evaluated their system using search strategies extracted from 17 systematic reviews and by searching MEDLINE. Cohen et al. [5] developed an evaluation collection containing 15 systematic drug class reviews. This collection was used also in subsequent work [4, 8]. Our test collection provides 94 search strategies. In addition, it relies on a large document corpus of 26 million studies, for which PICO and UMLS annotations have been extracted (see Section 3).

3 CREATION OF THE COLLECTION

3.1 Document Collection

The collection provides links to the identifiers of 26 million studies indexed in MEDLINE as of December 13th, 2016. Studies are comprised of a title and abstract; in addition, metadata fields including author(s), date of publication and type of study are available. We annotated all documents with PICO (population, intervention, control, outcome) tags using RobotReviewer [16]. The PICO framework is commonly used to help formulate search strategies for systematic reviews [12]. In addition, UMLS annotations for the collection are distributed by the NLM[3]. These annotations allow for the development and evaluation of new methods that exploit structured information — either PICO or UMLS — to improve retrieval, e.g. [2].

3.2 Query Topics

Query topics are represented by search strategies extracted from a set of published systematic reviews. We randomly selected a set of 93 systematic reviews published in the Cochrane initiative[4] between January 2014 and January 2016. We then manually examined the search strategies of each systematic review and extracted the search strategy that the authors devised for searching MEDLINE. (If a review contained more than one search strategy, we only extracted the first.) These search strategies are not expressed in a standard format; an example of a search strategy is given in Figure 1. Search strategies often separately define date-range restrictions to be applied to the publication dates of matching studies. Search strategies were extracted independently by two authors (one medical professional and one IR researcher) and disagreements were

[3] https://ii.nlm.nih.gov/MMBaseline/
[4] http://www.cochrane.org

```
{"query": {"bool": {"must_not": [
      {"range": {"pubdate": {
            "format": "YYYY-MM-DD",
            "gt": "2014-05-19"}}},
      {"range": {"pubdate": {
            "format": "YYYY-MM-DD",
            "lt": "1949-12-31"}}}],
    "should": [
      {"multi_match": {
          "fields": ["text.stemmed", "title.stemmed"],
          "query": "MMSE*"}},
      { "multi_match": {
          "fields": ["text.stemmed", "title.stemmed" ],
          "query": "sMMSE"}},
      {"multi_match": {
          "fields": ["text.stemmed", "title.stemmed" ],
          "query": "Folstein*"}},
      {"multi_match": {
          "fields": ["text.stemmed", "title.stemmed"],
          "query": "MiniMental"}},
      {"multi_match": {
          "fields": ["text.stemmed", "title.stemmed"],
          "query": "mini mental stat*",
          "type": "phrase"}}]}}}
```

Figure 2: Elasticsearch query for the search strategy of Figure 1; ⋆ is the wildcard operator in Elasticsearch.

resolved via an adjudication process. Each search strategy was manually converted into a boolean query in Elasticsearch, including date ranges restrictions. Figure 2 reports the Elasticsearch query for the search strategy of Figure 1.

3.3 Relevance Assessments

Relevance assessments were extracted from the systematic reviews themselves. Each systematic review contained a listing of the studies included and excluded as references. Of these studies, not all may have been retrieved by the queries we acquired: for example a study may have been retrieved by a search on a different database. We thus further differentiate between studies that are retrievable by the considered boolean queries, and studies that are not retrievable. We thus classified studies on four levels of relevance: excluded and not retrieved (*l1*), included and not retrieved (*l2*), excluded and retrieved (*l3*), and included and retrieved (*l4*). On average, 14 studies were identified as relevant per search strategy ($min = 1$, $max = 93$, $std = 16$). Figure 3(a) reports the distribution of relevant studies per search strategy at the lowest level of relevance (all studies are considered – l1).

3.4 Tasks and Evaluation Measures

Next we analyse the tasks modelled in our test collection and outline the evaluation measures most suited for each task.

Task 1 — retrieval for screening: The aim of this task is to retrieve all relevant studies, minimising non relevant studies, thereby minimising the time researchers need to spend in reviewing the full text of studies. As such, appropriate evaluation measures are precision-recall curves, F_β-measure, and work saved over sampling (WSS) [3]. In the F_β-measure, the parameter β controls the preference towards recall over precision; studies in systematic reviews automation used $\beta = 1$, $\beta = 3$ (recall three times more important than precision) and $\beta = 0.5$ [11]. WSS measures the work saved (wrt. the number of studies required to be screened) by comparing the number of not relevant studies that have not been retrieved (true negatives), those that have been retrieved, and recall.

Task 2 — screening prioritisation (ranking): The aim of this task is, given a set of retrieved studies, to rank studies according to their likelihood of being eligible for the systematic review (relevance). Achieving improvements in this task would have a significant effect on large reviews where researchers need to assess hundreds of thousands of studies; reviews for which only hundreds of studies need to be screened will unlikely benefit significantly from screening prioritisation. Unlike Task 1, recall does not play a key role in this task, as compared systems will have the same level of recall (they re-rank the same set of documents). Instead, precision-oriented measures are to be used. The use of precision-recall curves models cases where researchers are confident to examine only a subset of retrieved studies because enough representative samples have been already encountered (thus stopping at specific recall levels). The use of average precision (AP) covers cases where researchers are interested in examining the full ranking, but prefer to encounter relevant studies early on so that further processing of relevant studies may be finished before they finish examine the full list of results. Other gain-discount measures like nDCG, RBP and ERR could be used instead of MAP; however, it is yet unclear how to correctly model the discount curve to be associated with late retrieval of relevant studies within the task of screening prioritisation.

Task 3 — stopping point: The aim of this task is for the system to determine at which rank position researchers should stop screening the retrieved results to minimise the number of not relevant studies examined, while maintaining recall. This task builds upon the ranked results of Task 2. Task 3 attempts to predict the optimal point at which screening should be stopped. Stopping point determination has received increasing attention in IR recently, e.g., [10]. In this task researchers will screen every study up to the stopping point; evaluation is performed using F_β-measure and WSS.

4 ANALYSIS OF THE COLLECTION

4.1 Collection Statistics

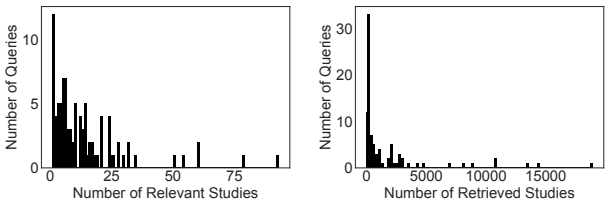
Search strategies varied in length from a minimum of 2 boolean clauses[5] to a maximum of 405 (mean=42, SD=49). All search strategies contained a date range constraining the search to studies published in that period. Special operators (field-based search, wildcard match, etc.) appeared in 91 of the 93 search strategies.

An IR system based on Elasticsearch (v5.2), was setup to index the studies in the collection (using field-based indexing to separate title, text, metadata information). Boolean retrieval was used to retrieve studies that satisfied each of the boolean search strategies in the collection. For each search strategy, the retrieved results formed the set of studies to be screened. The number of studies that the strategies retrieved varied largely: the median number of studies returned per search strategy was 1,159 ($min = 1$, $max = 1,271,362$, $std = 174,020$). Figure 3(b) summarises the distribution of the number of studies retrieved by the search strategies.

4.2 Screening Prioritisation Experiments

To demonstrate the use of the test collection we perform a number of experiments for Task 2 — screening prioritisation (Section 3.4). This task is to rank studies retrieved by the boolean query. We

[5]For example, the strategy in Figure 1 has 5 boolean clauses.



(a) Number of relevant studies per query.

(b) Number of retrieved studies per query (16 outliers excluded.)

Figure 3: Study retrieval statistics.

Table 1: Comparison of MAP between BM25 and Dirichlet LM and tf-idf. Differences between BM25 and Dirichlet LM are statistically significant (one-tailed paired t-test, $p < 0.05$). Other differences are not significant. *l0*: all levels of relevance considered as relevant; *l3*: only studies that are retrieved by the boolean query are considered as relevant.

	BM25	Dirichlet LM	tf-idf
MAP (*l0*)	0.0430	0.0311	0.0399
MAP (*l3*)	0.1062	0.0707	0.0985

compared the effectiveness of three common IR baselines over four levels of relevance: LM with Dirichlet smoothing, tf-idf and BM25[6]. Precision-recall curves and MAP were used to evaluate screening prioritisation effectiveness; note that all systems achieved the same level of recall. Results are reported in Figure 4 and Table 1 and show that both BM25 and tf-idf are equivalent baselines for the task of screening prioritisation, while Dirichlet LM is the poorest performing method.

5 CONCLUSION

We provide a test collection designed for the evaluation of retrieval systems used to identify research studies for inclusion in systematic reviews. The collection contains approximately 26 million studies, 93 query topics extracted from Cochrane systematic reviews and associated relevance assessments.

This collection represents a valuable resource in supporting the development of systems to reduce the effort and cost of compiling systematic reviews. This topic is attracting increasing interest within the IR community, as shown from the newly established CLEF 2017 eHealth Task 2[7] which focuses on the evaluation of IR systems for technologically assisted reviews in empirical medicine (with similar tasks as those identified here). Our collection could be used separately to develop and train systems for this shared task, or as an alternative validation resource. The collection along with associated resources is made available at https://github.com/ielab/SIGIR2017-PICO-Collection.

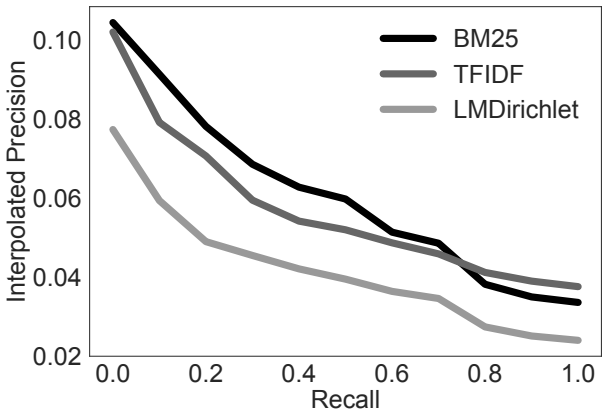


Figure 4: Precision-recall curves comparing the effectiveness of tf-idf, BM25, and Dirichlet LM (excluding studies that were not retrieved, i.e. *l3*) for screening prioritisation (Task 2) on our collection.

REFERENCES

[1] Tanja Bekhuis, Eugene Tseytlin, Kevin J Mitchell, and Dina Demner-Fushman. 2014. Feature engineering and a proposed decision-support system for systematic reviewers of medical evidence. *PloS one* 9, 1 (2014), e86277.

[2] Florian Boudin, Jian-Yun Nie, and Martin Dawes. 2010. Clinical information retrieval using document and PICO structure. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 822–830.

[3] AM Cohen, WR Hersh, K Peterson, and PY Yen. 2005. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association: JAMIA* 13, 2 (2005), 206–219.

[4] Aaron M Cohen. 2011. Performance of support-vector-machine-based classification on 15 systematic review topics evaluated with the WSS@ 95 measure. *Journal of the American Medical Informatics Association* 18, 1 (2011), 104–104.

[5] Aaron M Cohen, William 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 (2006), 206–219.

[6] Sarvnaz Karimi, Stefan Pohl, Falk Scholer, Lawrence Cavedon, and Justin Zobel. 2010. Boolean versus ranked querying for biomedical systematic reviews. *BMC medical informatics and decision making* 10, 1 (2010), 58.

[7] Sarvnaz Karimi, Stefan Pohl, Falk Scholer, Lawrence Cavedon, and Justin Zobel. 2010. Boolean versus ranked querying for biomedical systematic reviews. *BMC Medical Informatics and Decision Making* 10, 1 (2010), 1.

[8] Madian Khabsa, Ahmed Elmagarmid, Ihab Ilyas, Hossam Hammady, and Mourad Ouzzani. 2016. Learning to identify relevant studies for systematic reviews using random forest and external information. *Machine Learning* 102, 3 (2016), 465–482.

[9] David Martinez, Sarvnaz Karimi, Lawrence Cavedon, and Timothy Baldwin. 2008. Facilitating biomedical systematic reviews using ranked text retrieval and classification. In *Australasian Document Computing Symposium (ADCS)*. 53–60.

[10] David Maxwell, Leif Azzopardi, Kalervo Järvelin, and Heikki Keskustalo. 2015. Searching and stopping: An analysis of stopping rules and strategies. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 313–322.

[11] 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), 5.

[12] Connie Schardt, Martha B Adams, Thomas Owens, Sheri Keitz, and Paul Fontelo. 2007. Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC medical informatics and decision making* 7, 1 (2007), 16.

[13] Ian Shemilt, Antonia Simon, Gareth J Hollands, Theresa M Marteau, David Ogilvie, Alison O'Mara-Eves, Michael P Kelly, and James Thomas. 2014. Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. *Research Synthesis Methods* 5, 1 (2014), 31–49.

[14] Claire Stansfield, Alison O'Mara-Eves, and James Thomas. 2015. Reducing systematic review workload using text mining: opportunities and pitfalls. *Journal of EAHIL* 11, 3 (2015), 8–10.

[15] Guy Tsafnat, Paul Glasziou, Miew Keen Choong, Adam Dunn, Filippo Galgani, and Enrico Coiera. 2014. Systematic review automation technologies. *Systematic reviews* 3, 1 (2014), 74.

[16] Byron C Wallace, Joël Kuiper, Aakash Sharma, Mingxi Brian Zhu, and Iain J Marshall. 2016. Extracting PICO sentences from clinical trial reports using supervised distant supervision. *Journal of Machine Learning Research* 17, 132 (2016), 1–25.

[6] Parameter values were set to default values in Elasticsearch: $\mu = 2,000$, $k_1 = 1.2$, $b = 0.75$.

[7] https://sites.google.com/site/clefehealth2017/task-2