Technologically assisted systematic reviews in empirical medicine

Ties van Rozendaal, *University of Amsterdam*

***Abstract*—Finding all relevant documents out of a large collection is an essential process in evidence based medicine. This is a costly task that is likely to benefit from automation. We investigate three techniques from the information retrieval (IR) domain, using a custom build search engine. We search through PubMed Central and use the Cochrane review library as a golden standard. Improving the search results by expert feedback seems especially promising, as it is an easy process that increases recall. Yet this is an exploratory paper, and there is a high need for further research.**

***Keywords*—Evidence based medicine, systematic review, total recall problem, information retrieval, relevance feedback, user search**

I. Introduction

Evidence-based medicine has become an important strategy in health care and policy making [8]. In order to practice evidence-based medicine, it is important to have a clear overview over the current scientific consensus. These overviews are provided in systematic review articles, that summarise all evidence that is published regarding a certain topic (e.g., a treatment or diagnostic test).

In order to write a systematic review, researchers have to conduct a search that will retrieve all the documents that are relevant. This is a difficult task, known in the Information Retrieval (IR) domain as the total recall problem. The current approach to this problem in the field of systematic reviews can roughly be devised into two phases: a search phase and a filtering phase.

The goal of the search phase is to obtain all relevant documents. Although this is practically impossible, the search is optimised to return as many relevant documents as possible. This is done at the cost of obtaining many additional irrelevant documents. There are many medical libraries that can be searched, MEDLINE and EMBASE are the two most important libraries.

Because the search phase yields many irrelevant results, the relevant documents have to be filtered out. The selection is done manually, and is usually divided in multiple stages (e.g., a quick scan based on abstract followed by a full-text scan). As there can by many thousands of search results [7], this is a very resource-demanding task.

Medical libraries are expanding rapidly [5], complicating the already very costly task of finding the relevant documents. The need for automation in this process is high.

Supervised by Dr. Evangelos Kanoulas, *ILPS, University of Amsterdam*

In cooperation with Rene Spijker and Mariska Leeflang Renee Spijker, *Cochrane Netherlands*

Thanks to IEEE for their format

The two main concerns in applying automation techniques are transparency and high recall. Systematic Reviews need to have a clear and traceable method, and the search needs to be reproducible. New technologies should not affect this transparency. Furthermore, it is important to retrieve as many documents as possible, automation should not reduce the number of relevant documents found too much.

A. Optimizing document retrieval

Many papers have stressed the need for automation in document retrieval for systematic reviews ([12], [10], [4] and [9]). Some improvements have been implemented and are widely used today. Despite these tools, document retrieval is still a very time consuming task. This paragraph will provide a rough overview of the types of techniques that are currently being used or investigated. Read [9] or [10] for a more comprehensive overview.

1) Optimising Search

The goal for optimisation in the search phase is two fold: Most importantly the search needs to be optimised to contain as many relevant documents as possible. Secondly and subservient to the main goal, the number of irrelevant documents returned should be tried to kept low.

The search is currently being conducted as a boolean search. Search terms are combined using boolean operators (like AND, OR) and only documents that satisfy all requirements are returned. Wildcards (allowing any sequence of characters) and 'near' conditions (requiring certain terms to be near each other) are used to make the query more specific. Next to this, many libraries are including medical ontologies in their search engines. These allow researchers to search for documents containing very specific medical terms. The main search engines only search through the abstracts of the articles.

Nowadays, all these techniques are combined resulting in a very complex boolean query (often consisting of over 80 terms) [9]. There are many tools available to assist in the process, which help the user expand the query. Nonetheless, building a query is a very specialistic task which requires a lot of knowledge about the specific search engine that is being used.

2) Optimising document selection

Filtering the documents is essence a binary classification problem (relevant vs irrelevant class). As such, several publications investigated the performance of different classifiers on this task [4], [7], [11].

In order to train a classifier, a training set is needed consisting of documents of which the actual relevance is known.

This can be problematic, as new searches may require a new classifier, tweaked to that search. In order to train this classifier, the true relevance of the documents needs to be known.

Active learning has been proposed as a solution for this problem. In active learning a classifier is trained on-the-go, using expert feedback. As more documents are inspected, the classifier is retrained, aiding the manual filtering process.

Only a few studies investigated the use of classifiers (both with and without active learning) [7], [11].

B. Research question

Even though automation applications in the filtering phase may be helpful, they are still constrained by the documents found in the search phase. In turn, the search phase is limited by the users knowledge of the search engine, and boolean queries get increasingly more complex.

We suggest a complete revision of the search process. Many discoveries in IR have resulted in enormous improvements in search engines. The current method of boolean search on the abstract is an outdated method, and building a good query is difficult.

We propose to improve the document retrieval by refining the search engine. The goal is to make search more intuitive, and improve recall (leading to a less costly filtering phase).

We will investigate three improvements to the current use of search engines. Each improvement will be outlined in the following sections.

1) Abstract vs Full text

The EMBASE and MEDLINE Search engines only search through an articles abstract and meta-data. All the information in the body of the article is ignored. When search engines would take the body into account, it should be possible to search more precisely.

2) Boolean vs Best Match

A boolean query seems to be simple, as one can reason about why a certain document was, or was not found. However, the long boolean queries that are being used today are far from understandable. The reason is that many terms are added to the query to include or exclude specific results. In the IR domain, so called best-match search (BM) has become the standard. In best-match search, keywords are thought of as a continuum. A match-score is calculated for each document-query pair, and documents are ranked based on their score. Various measures exist, controlling for confounding factors such as document length or rarity of the keyword. Best match searches may eliminate the need for all the extra clauses that are used in a boolean search.

3) Relevance feedback

Even though best match may make query formulation easier, the results will still need fine tuning. Scientist in the systematic review domain have specialistic medical knowledge that should be used in the search process. Relevance feedback is a mechanism that can incorporate the expertise knowledge to finetune the search results. Much like active learning, relevance feedback adjusts the search as new documents are labeled. The positive and negative examples are used to modify the query so that the remaining results are reranked (with the goal to improve performance). Relevance feedback provides an intuitive way to include domain knowledge and increase efficiency.

C. Experiments

This paper aims to explore the effectiveness of the three improvements described above. In order to do so, a search engine was set up on the PubMed central library. The Cochrane library of reviews is used as a golden standard to evaluate the various search strategies.

II. Method

A. General Setup

1) The library

The PubMed Central library was used. This library was chosen because the full text of all articles is publicly available [3]. A large part of PubMed Central (1,227,716 out of 1,317,348 articles) is also published in the important MEDLINE library. All documents were downloaded, and the meta fields (title, publish date, keywords), body, and abstract were extracted from the raw XML.

2) The search engine

The Elasticsearch 2.3 engine [2] (running on Ubuntu 14.04) was used to index and search through the documents.

B. Evaluation

1) Gold standard

In order to evaluate a search engine, it needs to be evaluated against a search for which the desired output is known. This evaluation set is known as the gold standard. We used the Cochrane review library as a gold standard. The Cochrane library consists of many review articles that are clearly structured. For each article, the references are divided into several categories (see Fig 1) stating clearly which articles were used for the review and which articles were filtered out.

We focussed on Diagnostic Test Accuracy (DTA) review articles. Search in this area is generally considered the hardest, and a breakthrough in this field would likely be applicable to other areas as well [6]. All 58 DTA-review articles were downloaded from the Cochrane library [1]. For each review article, the references were extracted and sorted by reference type and duplicate references were removed.

Combined, the 58 review articles cited a total of 8,209 PubMed publications. Of these citations, 374 articles (5%) occurred in the PubMed central database (and were thus accessible by our search engine). The distribution of these references across articles is shown in Figure 1. The number of accessible references per review article is rather low. In order to perform an accurate evaluation, it is desirable to have 10 or more positive references. Therefore, the categories *included, awaiting, additional* and *excluded* were grouped together in our analysis. A document in either of these categories was marked as relevant to the topic of that review. All other documents were assumed to be irrelevant.

2) Measures

A search returns a (ranked) list of documents, and there are several metrics to quantify the accuracy of the fit. The measures used in this article are described and motivated below.

recall Recall expresses the proportion of documents that are correctly retrieved. This measure is also known as **sensitivity** in the systematic review domain.
It is defined by
$$\text{recall} = \frac{\text{\# relevant documents retrieved}}{\text{\# all relevant documents}}.$$

precision Precision expresses the proportion of the retrieved documents that are correct.
It is defined by
$$\text{precision} = \frac{\text{\# relevant documents retrieved}}{\text{\# all documents retrieved}}.$$

F_1 F_1 defines the geometric mean between recall and precision.
It is defined as $F_1 = 2 \cdot \frac{\text{recall} \cdot \text{precision}}{\text{recall} + \text{precision}}$

F_β In areas like systematic reviewing, recall may be more important than precision. The F_β measure allows to put more weight on one of the two. It is defined so that a β value of 10 means that recall is 10 times as important as precision.
It is defined by:
$$F_\beta = (1 + \beta^2) \cdot \frac{\text{recall} \cdot \text{precision}}{\text{recall} + (\beta^2 \cdot \text{precision})}$$
In this report a value of $\beta = 10$ will be used, as suggested by [4] for systematic review articles.

MAP The measures above define the performance of the search engine at a specific rank. As a result, each query has as many precision-, or F_β-values as the number of documents that it retrieves. This makes it difficult to combine the performance for different queries. The average precision is a measure of performance over the entire rank-list. It equals to the area under the precision-recall plot. The average precision can in turn be averaged over queries resulting in the Mean Average Precision (MAP).

All the measures above were implemented. For each query, precision/recall plots and effort/recall plots were constructed and manually inspected. Precision, recall and F scores are measured at the rank at which the last document was retrieved. F_1 seemed to be proportional to F_β, and to be concise only F_β values are reported. MAP seemed to be a good descriptor of search performance.

3) Query mining

Each review article in the Cochrane library contained a detailed description of the query used to search through PubMed. Because the features of the PubMed search engine do not match completely with their implementation in Elasticsearch, some changes had to be made. The general approach was to adjust the query in such a way that recall was maintained, at the possible cost of decreasing precision.

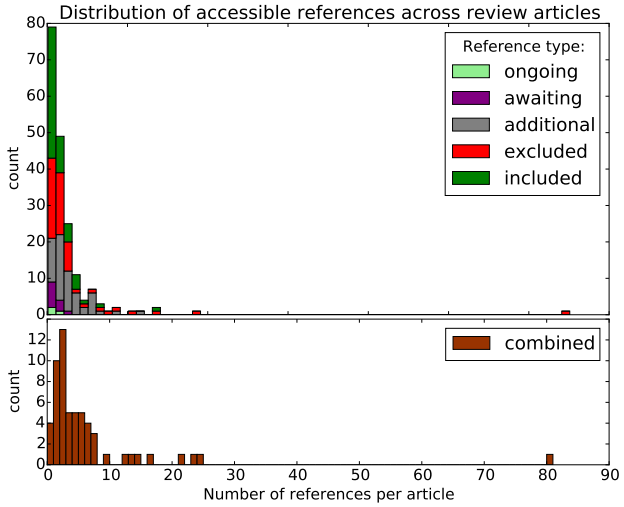

Fig. 1. **Histogram of accessible references per article** Only the subset of references that are available on PubMed Central are considered. References of the reviews are divided into the sections described in the top figure. Because the number of references per review article was low, all categories were grouped, resulting in the distribution in the bottom figure.

- *MeSH terms* (a popular medical ontology) were omitted from the queries.
- wildcards within literal strings were replaced with singe terms (e.g. "retinal nerve fib*" → "retinal nerve" AND fib)
- near-clauses were converted into AND clauses (e.g. fib* adj2 retinal → fib* AND nerve)

Next to this, search results were limited by date. The field *Assesed as up to date* in the review articles was used. Articles that were published after this date were not retrieved by our search engine.

All queries with 7 references or more were extracted manually (n=15).

III. Experiments

1) Abstract vs. Full text

The first experiment investigated the advantage of including the full text in the search. For both conditions, the processed boolean query used in the review article was used. In the *abstract* condition, the query was applied to the abstract, title, and meta fields of the article. In the *full-text* condition, the body of the article was also searched through.

2) Boolean vs. best match

In the second experiment, we tried investigating the use of Best Match (BM) queries. As explained in the introduction, using BM queries can simplify the search process considerably. However, because this is a new domain, it is not clear how the best match query should be formulated. We cannot use the original boolean queries, because of their complexity. We introduced two best match categories, based on the information provided in the review articles. The boolean query was used as a baseline condition (*boolean*).

For the first BM condition, we used the review title as a query (*title*). As opposed to other sections of the review article, the title contains few words and is supposed to represent the most essential keywords.

For the second condition, we used two example documents to build the best match query (*examples*). The first two relevant documents were provided to the search engine. The search engine selected words that co-occured most in both text to build a BM query. There are many parameters to this process, which are reported in appendix A.

In the *examples* condition, the examples that are given to the search engine will be trivial to find. Therefore, these two documents are removed from the evaluation set. In order to make a fair comparison between all three conditions, all of them are evaluated on the set without the example documents.

The entire experiment was repeated with and without accessing the full text.

3) Relevance Feedback

The effect of relevance feedback was investigated in the third experiment. It is not straightforward to use relevance feedback in a boolean query, for that reason the relevance feedback was implemented based on the best match query from the previous experiment. The *examples* condition was used as the initial query. The query would return a ranked list of documents, the true relevance-label of the first document on this list was returned to the search engine (simulating real user feedback). In the search engine, relevant documents were added to the examples, whereas irrelevant documents were added to the list of counter examples.

As there are only a few relevant documents per review article, negative examples are likely to dominate the result. To investigate this problem, a *positive feedback only* condition was introduced, only taking into account positive feedback. Because the feedback process was a lot slower, the search evaluation was automatically stopped after 2500 documents had been retrieved.

Again, the set without the initial examples was used for evaluation.

4) Relevance Feedback Tuning

As shown in appendix A, the relevance feedback was heavily parameterised. The effect of some parameters was investigated informally, yielding the optimal settings in appendix A. Next to this, two experiments were run investigating several values of *irrelevant que size* and *max query terms*.

The same evaluation set as in the previous two experiments was used.

IV. Results

1) Abstract vs. Full text

Figure 2 shows the performance for the boolean query on abstract and on full text. As expected, searching on full text improves recall as compared with search on abstract and meta only. However, the cost of this is decreased precision, meaning that more documents have to be examined in order to find the same number of relevant documents. The MAP is also lower for full-text search, and according to the F_β measure, searching on abstract only is more desirable.

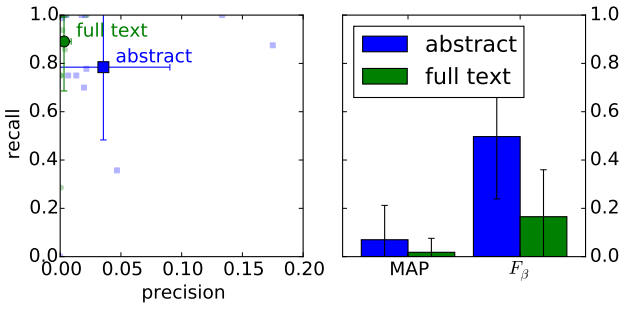

Fig. 2. **Performance using a boolean query on full text and abstract.** The left plot shows (final) recall and precision for all queries. The larger markers show the mean and standard deviation for both groups. The right plot shows the Mean Average Precision and mean F_β-score for each group.

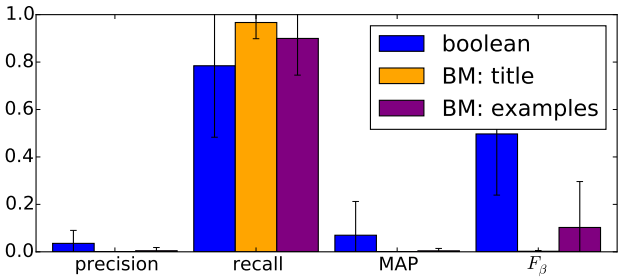

Fig. 3. **Search results for boolean vs best mach (BM) search on abstract.** Best match queries were based on the review articles title (title) or on the content of two provided documents which should have been included (examples). Replication on the full text yielded similar results.

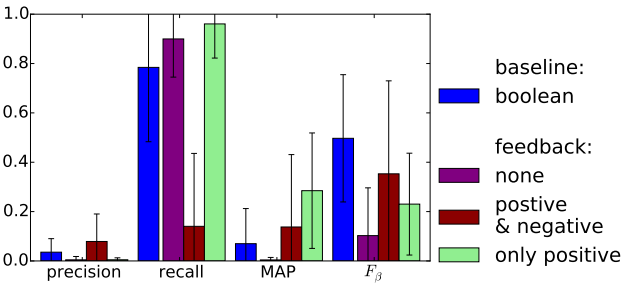

Fig. 4. **Search results for relevance feedback during example search.** Baseline depicts boolean full text search. The initial query consisted of two example documents (no feedback condition). The feedback loop consisted of labeling of the first ranked document, followed by reranking of the search results.

2) Boolean vs. best match

The results of the second experiment are depicted in figure 3. BM search on title results in almost total recall. However, this is a misleading result, because almost all the documents in the database were retrieved by this search (745,043 on average). This is reflected by the extremely low precision. The low F_β and MAP values confirm that BM search on the title is not

effective.

BM search using examples also improves recall. Again, this improvement comes at the cost of reduced precision. Example search did return a reasonable amount of documents, as reflected by the higher $F_\beta score$. Due to low precision, the MAP and F_β score are well below the *boolean* baseline, suggesting that these BM searches do not improve the overall search accuracy.

Repetition of this experiment on the full text yielded similar results.

3) Relevance Feedback

Figure 4 shows the results of the relevance feedback experiment. Both *boolean* search and *example BM* search from the previous experiment, are used as control conditions. Any form of feedback increases both the F_β and the MAP score, compared with the normal example search. Positive only feedback increases recall, whereas complete feedback (positive and negative) reduces recall but improves precision.

On average, the complete feedback searches returned 13.8 documents. The reason for this is that the negative counterexamples would dominate the query, resulting in a quick exclusion of all articles. This result explains the big gap in recall for this condition.

Positive only feedback seems to greatly improve performance. Not only compared to BM search without feedback, but also compared with the boolean baseline. Recall and MAP both exceed the boolean baseline in the positive feedback condition. The F_β value is higher compared with no feedback, but does not reach the value of the baseline.

4) Relevance Feedback Tuning

The previous experiment revealed that the inclusion of negative examples can greatly affect the performance of relevance feedback. As described in appendix A, the inclusion of negative examples is not a boolean parameter, but rather a continuous scale. Figure 5 shows the effect of varying this parameter.

A negative feedback que size of 0, corresponds to positive feedback only. In line with the previous experiment, recall decreases as negative feedback increases. Precision on the other hand, increases. The F_β value shows that the optimum could lie somewhere between 0 and 20.

Expanding the maximum number of query terms also decreases recall, but increases precision. MAP and F_β both peak around 20 query terms, suggesting a local optimum.

Both experiments show that parameters can greatly affect the performance of relevance feedback.

V. Conclusion

Three improvements to document retrieval in systematic reviewing have been suggested. Effectiveness of each suggestion varies, but in general there seem to be many opportunities for improvement.

Full-text, as opposed to abstract search increases recall, but does so at the cost of decreasing precision.

Investigating best-match search is difficult, because an alternative query needs to be formulated. Out of the investigated options, the boolean query is the most effective.

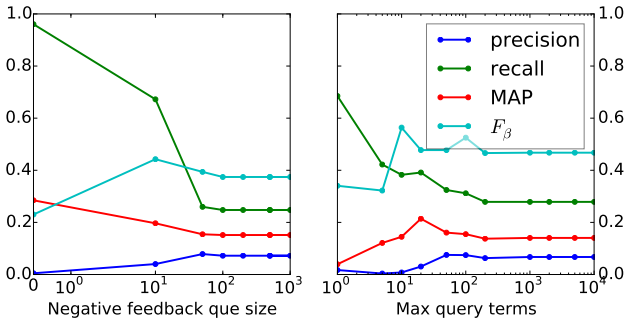

Fig. 5. **Finetuning of relevance feedback parameters** Both plots show the performance of relevance feedback varying one parameter (keeping tho other fixed). See appendix A for an explanation of the parameters.

Relevance feedback seems to be the most promising feature. Positive relevance feedback outperforms the current boolean search on most measures. Furthermore, our research suggest that tweaking the parameters may lead to even more improvement. Further research should investigate the optimal combination of settings.

VI. Discussion

A. The gold standard

Both the number of queries and the number of reference per query used in our evaluation was rather small. By using a bigger dataset, the signal to noise ratio can be increased, leading to more consistent findings.

Next to this, all reference types were grouped together. The difference between included and excluded references is essential in systematic document retrieval. Because of the sparsity of the data we grouped them together. It is unclear how different the reference types are. An articles could be excluded because it is irrelevant, but also because it does not meet the requirements stated in the protocol (e.g. the number of participants is not high enough). Further research could examine how the categories relate to each other.

Another concern is the possible bias of our golden standard. The references in the Cochrane review articles are obtained using a boolean search. Therefore our dataset might be biased towards these queries. However, Cochrane queries multiple libraries to obtain their results. Next to this, manual selection poses an additional filters. Again, a larger dataset will benefit more from this manual selection.

Lastly, we failed to fully replicate the boolean queries from the review articles. Ideally a recall of 1.0 would be expected. Even though the queries were somewhat modified, we tried to ensure that the original recall was remained. A possible explanation for the reduced recall is the fact that we omitted the MeSH terms. These terms are widely used in the Cochrane queries, and are a powerful tool. It would be very useful to repeat this study including MeSH terms in the queries to investigate what the added value of our improvements is.

B. Evaluation

Trends in MAP seemed to be consistent with trends observed at the query level. Precision and recall (at the final rank) also described important characteristics, although the exact rank at which they were measured (or the total number of documents retrieved) was needed in order to interpret them right. F_β measures did not seem to describe the data well.

Taken altogether MAP seems to be the best aggregated descriptor. However, a downside for the systematic review domain, is that we cannot increase the relative importance of recall compared with precision.

C. Relevance Feedback

Relevance feedback seems very promising, but more research is needed. Apart from finetuning the parameters, the exact implementation will have to be investigated. Sending feedback after each document could slow down the review process. We found that it did require a lot more time, and it may also be impractical for the researcher. Giving feedback in batches of a certain size may be more desirable.

Next to this, relevance feedback is used to improve an initial BM query. The impact of different initial queries on relevance tuning efficiency remains to be discovered.

D. Example search

Example search is essential for relevance feedback. Plain example search did achieve high not achieve high MAP values due to low precision. However, parameter tuning may lead to improvements.

Nonetheless, example search will always find documents that are similar to the ones provided. A downside of this may be that documents that are of a different order (yet equally important) could be less likely to be found. Research using a bigger dataset should investigate this possible issue.

A main advantage of example search is that it is very easy to incorporate expert knowledge. User studies could investigate it's effectiveness in practise.

Taken altogether the suggested improvements seem to indicate some interesting improvements. Yet this paper only explored some options, and thorough follow-up is needed.

APPENDIX A
PARAMETERS IN RELEVANCE FEEDBACK

stopwords Wether stopwords should be ignored, default: True

fields Fields to search trough, default: All

numdocs Number of documents to build the initial (example) query with, default: 2

terms per doc The number of terms that are extracted from the examples is linearly dependent on document, this parameter defines the slope, default: 40.

max query terms A limiter of the function described above, default: 100.

min doc freq The number of documents in which a term has to occur before it will be considered as a query term, default: 2.

min term freq The minimum number of times that a term has to occur before it will be considered as a query term, default: 4.

irrelvant queue size Not all negative examples are used, they are kept in a deque, with a fixed size, as an item is added to the top, the item at the bottom is flushed, default value: 20.

minimum should match The percentage of query terms that a document should match in order to be retrieved. This parameter ensures the search is limited. Default: 10%

max scroll Because the relevance feedback simulation is relatively slow, it can be useful to terminate the search at a certain number of documents. Default: 2500

REFERENCES

[1] Cochrane diagnostic test accuracy reviews. http://methods.cochrane.org/sdt/cochrane-diagnostic-test-accuracy-reviews. Accessed: 2016-06-20.

[2] Elastic revealing insights from data. https://www.elastic.co/. Accessed: 2016-06-20.

[3] Ftp service - ncbi - national institutes of health. http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/. Accessed: 2016-06-20.

[4] Hayda Almeida, Marie-Jean Meurs, Leila Kosseim, Greg Butler, and Adrian Tsang. Machine learning for biomedical literature triage. *PloS one*, 9(12):e115892, 2014.

[5] Benjamin G Druss and Steven C Marcus. Growth and decentralization of the medical literature: implications for evidence-based medicine. *Journal of the Medical Library Association*, 93(4):499, 2005.

[6] Mariska MG Leeflang, Jonathan J Deeks, Yemisi Takwoingi, and Petra Macaskill. Cochrane diagnostic test accuracy reviews. *Systematic reviews*, 2(1):1, 2013.

[7] Makoto Miwa, James Thomas, Alison OMara-Eves, and Sophia Ananiadou. Reducing systematic review workload through certainty-based screening. *Journal of biomedical informatics*, 51:242–253, 2014.

[8] David L Sackett. Evidence-based medicine. In *Seminars in perinatology*, volume 21, pages 3–5. Elsevier, 1997.

[9] James Thomas, John McNaught, and Sophia Ananiadou. Applications of text mining within systematic reviews. *Research Synthesis Methods*, 2(1):1–14, 2011.

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

[11] Byron C Wallace, Kevin Small, Carla E Brodley, and Thomas A Trikalinos. Active learning for biomedical citation screening. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 173–182. ACM, 2010.

[12] Erfan Younesi, Luca Toldo, Bernd Müller, Christoph M Friedrich, Natalia Novac, Alexander Scheer, Martin Hofmann-Apitius, and Juliane Fluck. Mining biomarker information in biomedical literature. *BMC medical informatics and decision making*, 12(1):1, 2012.