Proceedings Paper:

Alharbi, A. and Stevenson, R. orcid.org/0000-0002-9483-6006 (2019) A dataset of systematic review updates. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 21-25 Jul 2019, Paris, France. ACM , pp. 1257-1260. ISBN 978-1-4503-6172-9

https://doi.org/10.1145/3331184.3331358

A Dataset of Systematic Review Updates

Amal Alharbi*
King Abdulaziz University
Jeddah, Saudi Arabia
ahalharbi1@sheffield.ac.uk

Mark Stevenson
University of Sheffield
Sheffield, United Kingdom
mark.stevenson@sheffield.ac.uk

ABSTRACT

Systematic reviews identify, summarise and synthesise evidence relevant to specific research questions. They are widely used in the field of medicine where they inform health care choices of both professionals and patients. It is important for systematic reviews to stay up to date as evidence changes but this is challenging in a field such as medicine where a large number of publications appear on a daily basis. Developing methods to support the updating of reviews is important to reduce the workload required and thereby ensure that reviews remain up to date. This paper describes a dataset of systematic review updates in the field of medicine created using 25 Cochrane reviews. Each review includes the Boolean query and relevance judgements for both the original and updated versions. The dataset can be used to evaluate approaches to study identification for review updates.

KEYWORDS

Systematic review; systematic review update; test collection; evaluation

ACM Reference Format:
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 (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3331184.3331358

1 INTRODUCTION

Systematic reviews are widely used in the field of medicine where they are used to inform treatment decisions and health care choices. They are based on assessment of evidence about a research question which is available at the time the review is created. Reviews need to be updated as evidence changes to continue to be useful. However, the volume of publications that appear in the field of medicine on a daily basis makes this difficult [2]. In fact, it has been estimated that 7% of systematic reviews are already out of date by the time of publication and almost a quarter (23%) two years after they have appeared [19].

A review can be updated at any point after it has been created and would ideally be carried out whenever new evidence becomes available but the effort required makes this impractical. The Cochrane Collaboration recommends that reviews should be updated every two years. Cochrane's Living Evidence Network have recently started developing living systematic reviews for which evidence is reviewed frequently (normally monthly) [7] but it is unclear whether this effort is sustainable. The Agency for Healthcare Research and Quality suggests that reviews are updated depending on need, priority and the availability of new evidence [15].

The process that is applied to update a systematic review is similar to the one used to create a new review [6]. A search query is run and the resulting citations screened in a two stage process. In the first stage (*abstract screening*) only the title and abstract of the papers retrieved by the Boolean search are examined. It is common for the majority of papers to be removed from consideration during the abstract screening stage. The remaining papers are considered in a second stage (*content screening*) during which the full papers is examined. If any new relevant studies are found then data is extracted and integrated into the review. The review's findings are also updated if the evidence is found to have changed from the previous version. The screening stages are one of the most time consuming parts of this process since an experienced reviewer takes at least 30 seconds to review an abstract and substantially longer for complex topics [22]. The problem is made more acute by the fact that the search queries used for systematic reviews are designed to maximise recall, with precision a secondary concern, while the volume of medical publications increases rapidly.

Developing methods to support the updating of reviews are therefore required to reduce the workload required and thereby ensure that reviews remain up to date. However, previous work on the application of Information Retrieval (IR) to the systematic review process has only paid limited attention to the problem of updating reviews (see Section 2).

This paper describes a dataset created for evaluating automated methods applied to the problem of identifying relevant evidence for the updating of systematic reviews. It is, to our knowledge, the first resource made available for this purpose. In addition, this paper also reports performance of some baseline approaches applied to the dataset. The dataset described in this paper is available from https://github.com/Amal-Alharbi/Systematic_Reviews_Update.

2 RELATED WORK

A significant number of previous studies have demonstrated the usefulness of IR techniques to reduce the workload involved in the systematic review screening process for new reviews, for example [3, 5, 12–14, 16, 17, 22]. A range of datasets have been made available to support the development of automated methods for study identification. Widely used datasets include one containing 15 systematic reviews about drug class efficiency [3] and another

*Currently studying at the University of Sheffield

SIGIR '19, July 21–25, 2019, Paris, France

containing two reviews (on Chronic Obstructive Pulmonary Disease and Proton Beam therapy) [22]. Recently the CLEF eHealth track on Technology Assisted Reviews in Empirical Medicine [9, 20] developed datasets containing 72 topics created from diagnostic test accuracy systematic reviews produced by the Cochrane Collaboration. Another test collection has also been derived from 94 Cochrane reviews [18]. However, none of these datasets focus on the review updates.

Only a few previous studies have explored the use of IR techniques to support the problem of updating reviews [3, 11, 21]. In the majority of cases this work has been evaluated against simulations of the update process, for example by "time slicing" the included studies and treating those that appeared in the three years before review publication as being added in an update [11]. An exception is work that used update information for nine drug therapy systematic reviews [4], but this dataset is not publicly available.

To the best of our knowledge there is no accessible dataset that focuses on the problem of identifying studies for inclusion in a review update. The problem is subtly different from the identification of studies for inclusion in a new review since relevance judgements are available (from the original review) which have the potential to improve performance. A suitable dataset for this problem would include the list of studies considered for inclusion in both the original and updated reviews together with a list of the studies that were actually included in each review. This paper describes such a resource.

3 DATASET

The dataset is constructed using systematic reviews from the Cochrane Database of Systematic Reviews[1], a standard source of evidence to inform healthcare decision-making. Intervention reviews, that is reviews which assess the effectiveness of a particular healthcare intervention for a disease, are the most common type of reviews carried out by Cochrane. A set of 25 published intervention systematic reviews were selected for inclusion in the dataset. Reviews included in the dataset must have been available in an original and updated version (i.e. an updated version of the review has been published) and at least one new relevant study identified during the abstract screening stage for the update.

The following information was automatically extracted from each review: (1) review title, (2) Boolean query, (3) set of included and excluded studies (for both the original and updated versions) and (4) update history (including publication date and URL of original and updated versions).

3.1 Boolean Query

Candidate studies for inclusion in systematic reviews are identified using Boolean queries constructed by domain experts. These queries are designed to optimise recall since reviews aim to identify and assess all relevant evidence. Queries are often complex and include operators such as AND, OR and NOT, in addition to advanced operators such as wildcard, explosion and truncation [10].

Boolean queries in the reviews included in the dataset are created for either the OVID or PubMed interfaces to the MEDLINE database of medical literature. For ease of processing, each OVID query was automatically converted to a single-line PubMed query using a Python script created specifically for this purpose (see Figure 1).

(a) Multi-line query in OVID format

`1. endometriosis/`
`2. (adenomyosis or endometrio$).tw.`
`3. or/1-2`

(b) One-line PubMed translation
`endometriosis[Mesh:NoExp] OR adenomyosis[Text Word] OR endometrio*[Text Word]`

Figure 1: Example portion of Boolean query [8] in (a) OVID format and (b) its translation into single-line PubMed format. This portion of the query contains three clauses and the last clause represents the combining results of clause 1 and 2 in a disjunction (OR).

3.2 Included and Excluded Studies

For each version of the reviews (original and updated) the dataset includes a list of all the studies that were included after each stage of the screening process (abstract and content). The set of studies included after the content level screening is a subset of those included after abstract screening and represents the studies included in the updated review.

Included and excluded studies are listed in the dataset as PMIDs (unique identifiers for PubMed citations that make it straightforward to access details about the publication). If the PMID for a study was listed in the systematic review (which accounted for a majority of cases) then it was used. If it was not then the title of the study and year of publication were used to form a query that is used to search PubMed (see Figure 2). If the entire text of the title, publication year and volume of the retrieved record match the details listed in the systematic review then the PMID of that citation is used.

Study title: Clinical experience treating endometriosis with nafarelin.
Publication Year: 1989
Search Query: `clinical[Title] AND experience[Title]` AND treating[Title] `AND endometriosis[Title] AND nafarelin [Title] AND 1989[Date - Publication]`

Figure 2: Example of search query generated from title and publication year for study from Topic CD000155 [8].

3.3 Update History

Details of the date of publication of each version (original and update) are also extracted and included.

3.4 Dataset Characteristics

Descriptive statistics for the 25 systematic reviews that form the dataset are shown in Table 1. It is worth drawing attention to the small number of studies included after the initial abstract screening stage.

[1]https://www.cochranelibrary.com/cdsr/about-cdsr

Table 1: List of the 25 systematic reviews with the total number of studies returned by the query (Total) and the number included following the abstract (Abs) and content (Cont) screening stages. The average (unweighted mean) number of studies is shown in the bottom row. Note that for the updated review, the number of included studies in the table lists only the new studies that were added during the update.

	Original Review			Updated Review		
Review	**Total**	**Abs**	**Cont**	**Total**	**Abs**	**Cont**
CD000155	397	42	14	101	6	4
CD000160	433	7	6	1980	1	1
CD000523	34	6	3	18	1	1
CD001298	1384	22	15	1020	17	13
CD001552	2082	2	2	844	2	2
CD002064	38	2	2	9	1	0
CD002733	13778	30	10	6109	6	6
CD004069	951	5	2	771	9	7
CD004214	57	5	2	21	4	1
CD004241	838	25	9	193	5	3
CD004479	112	6	1	153	4	3
CD005025	1524	43	8	1309	46	4
CD005055	648	8	4	353	3	0
CD005083	462	46	16	107	9	2
CD005128	25873	5	4	5820	9	3
CD005426	6289	13	8	1413	3	0
CD005607	851	11	7	103	2	1
CD006839	239	8	6	93	3	3
CD006902	290	18	6	106	10	5
CD007020	348	47	4	47	4	3
CD007428	157	7	3	190	9	3
CD008127	5460	7	0	6720	2	1
CD008392	5548	15	5	1095	2	0
CD010089	41675	22	10	4514	4	0
CD010847	571	15	1	111	6	0
Average	4402	17	6	1335	7	3

4 EXPERIMENTS AND RESULTS

Experiments were conducted to establish baseline performance figures for the dataset. The aim is to reduce workload in the screening stage of the review update by ranking the list of studies retrieved by the Boolean query.

Performance at both abstract and content screening levels was explored. The collection was created by using the Boolean query to search MEDLINE using the `Entrez` package from `biopython.org`. The list of studies included after abstract screening was used as the relevance judgements for abstract level evaluation and the list of studies included after the content screening was used for content level evaluation.

4.1 Approaches

4.1.1 Baseline Query. A "baseline query" was formed using the review title and terms extracted from the Boolean query. This query is passed to BM25 [1] to rank the set of studies returned from the Boolean query for the review update.

4.1.2 Relevance Feedback. A feature of the problem of identifying studies for inclusion in updates of systematic reviews is that a significant amount of knowledge about which studies are suitable is available from the original review and this information was exploited using relevance feedback. Rocchio's algorithm [1] was used to reformulate the baseline query by making use of relevance judgements derived from the original review. Content screening judgements (included and excluded studies) were used for the majority of reviews. Abstract screening judgements were used if these were not available, i.e. no studies remained after content screening.

4.2 Evaluation Metrics

Mean average precision (MAP) and work saved over sampling (WSS) are the metrics most commonly used to evaluate approaches to study identification for systematic reviews, e.g. [5, 9, 20]. MAP represents the mean of the average precision scores over all reviews. WSS measures the work saved to retrieve a defined percentage of the included studies. For example WSS@95 measures the work saved to retrieve 95% of the included studies. WSS at recall 95 and 100 (WSS@95 and WSS@100) was used for the experiments reported in this paper.

4.3 Results

Results of the experiment are shown in Table 2. As expected, performance improves when relevance feedback is used. The screening effort required to identify all relevant studies (100% recall) is reduced by 63.5% at abstract level and 74.9% at content level. This demonstrates that making use of information from the original review can improve study selection for review updating.

Table 2: Performance ranking abstracts for updated reviews at (a) abstract and (b) content levels. Results are computed across all reviews at abstract level (25 reviews) and only across reviews in which a new study was added in the updated version for content level (19 reviews).

Approach	MAP	WSS@95	WSS@100
(a) abstract level (25 reviews)			
Baseline Query	0.213	51.70%	56.60 %
Relevance Feedback	0.413	58.80%	63.50%
(b) content level (19 reviews)			
Baseline Query	0.260	65.50%	70.50%
Relevance Feedback	0.382	69.90%	74.90%

Figure 3 shows the results of AP scores for all 25 reviews. Relevance feedback improved AP for 23 (92%) of the reviews.

There are also four reviews where the use of relevance feedback produced an AP score of 1, indicating that the approach reduces work required by up to 99.9%.

5 CONCLUSION

Updating systematic reviews is an important problem but one which has largely been overlooked. This paper described a dataset containing 25 intervention reviews from the Cochrane collaboration that

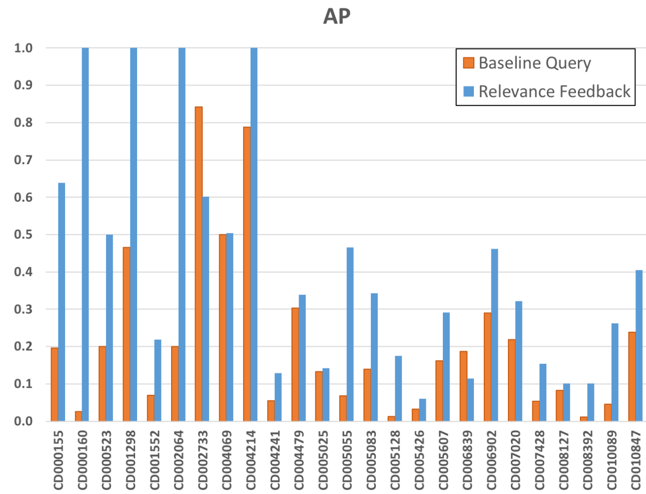

Figure 3: Abstract screening AP scores for each review using Baseline Query and Relevance Feedback.

can be used to support the development of approaches to automate the updating process. The title, Boolean query, relevance judgements for both the original and the updated versions are included for each systematic review.

Standard approaches were applied to the dataset in order to establish baseline performance figures. Experiments demonstrated that information from the original review can be used to improve study selection for systematic review updates. It is hoped that this resource will encourage further research into the development of approaches that support the updating of systematic reviews, thereby helping to keep them up to date and valuable.

REFERENCES

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 2011. *Modern Information Retrieval* (2nd ed.). Addison-Wesley Publishing Company, Boston, MA, USA.

[2] Hilda Bastian, Paul Glasziou, and Iain Chalmers. 2010. Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up? *PLOS Medicine* 7, 9 (Sep 2010), 1–6. https://doi.org/10.1371/journal.pmed.1000326

[3] Aaron Cohen. 2008. Optimizing feature representation for automated systematic review work prioritization. *AMIA ... Annual Symposium proceedings* (2008), 121–125.

[4] Aaron Cohen, Kyle Ambert, and Marian McDonagh. 2012. Studying the potential impact of automated document classification on scheduling a systematic review update. *BMC Medical Informatics and Decision Making* 12, 1 (2012), 33. https://doi.org/10.1186/1472-6947-12-33

[5] Aaron Cohen, William Hersh, Kim Peterson, and Po-Yin Yen. 2006. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association : JAMIA* 13, 2 (2006), 206–19. https://doi.org/10.1197/jamia.M1929

[6] Mark R Elkins. 2018. Updating systematic reviews. *Journal of Physiotherapy* 64, 1 (2018), 1–3. https://doi.org/10.1016/j.jphys.2017.11.009

[7] Julian H. Elliott, Anneliese Synnot, Tari Turner, Mark Simmonds, Elie A. Akl, et al. 2017. Living systematic review: 1. Introduction - the why, what, when, and how. *Journal of Clinical Epidemiology* 91 (November 2017), 23–30. https://doi.org/10.1016/J.JCLINEPI.2017.08.010

[8] Edward Hughes, Julie Brown, John Collins, Cindy Farquhar, Donna Fedorkow, et al. 2007. Ovulation suppression for endometriosis for women with subfertility. *Cochrane Database of Systematic Reviews* 3 (2007). https://doi.org/10.1002/14651858.CD000155.pub2

[9] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. 2017. CLEF 2017 technologically assisted reviews in empirical medicine overview. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017, CEUR Workshop Proceedings*, Vol. 1866. 1–29.

[10] 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), 1–20. https://doi.org/10.1186/1472-6947-10-58

[11] Madian Khabsa, Ahmed Elmagarmid, Ihab Ilyas, Hossam Hammady, Mourad Ouzzani, et al. 2016. Learning to identify relevant studies for systematic reviews using random forest and external information. *Machine Learning* 102, 3 (Mar 2016), 465–482. https://doi.org/10.1007/s10994-015-5535-7

[12] Halil Kilicoglu, Dina Demner-Fushman, Thomas C Rindflesch, Nancy Wilczynski, and Brian Haynes. 2009. Towards automatic recognition of scientifically rigorous clinical research evidence. *AMIA* 16 (2009), 25–31. https://doi.org/10.1197/jamia.M2996

[13] Seunghee Kim and Jinwook Choi. 2014. An SVM-based high-quality article classifier for systematic reviews. *Journal of Biomedical Informatics* 47 (2014), 153–159.

[14] Athanasios Lagopoulos, Antonios Anagnostou, Adamantios Minas, and Grigorios Tsoumakas. 2018. Learning-to-Rank and Relevance Feedback for Literature Appraisal in Empirical Medicine. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings.* 52–63. https://doi.org/10.1007/978-3-319-98932-7_5

[15] Ersilia Lucenteforte, Alessandra Bettiol, Salvatore De Masi, and Gianni Virgili. 2018. *Updating Diagnostic Test Accuracy Systematic Reviews: Which, When, and How Should They Be Updated?* Springer International Publishing, Cham, 205–227. https://doi.org/10.1007/978-3-319-78966-8_15

[16] David Martinez, Sarvnaz Karimi, Lawrence Cavedon, and Timothy Baldwin. 2008. Facilitating biomedical systematic reviews using ranked text retrieval and classification. In *13th Australasian Document Computing Symposium (ADCS)*. Hobart Tasmania, 53–60.

[17] Makoto Miwa, James Thomas, Alison O'Mara-Eves, and Sophia Ananiadou. 2014. Reducing systematic review workload through certainty-based screening. *Journal of Biomedical Informatics* 51 (2014), 242–253. https://doi.org/10.1016/j.jbi.2014.06.005

[18] Harrisen Scells, Guido Zuccon, Bevan Koopman, Anthony Deacon, Leif Azzopardi, et al. 2017. A test collection for evaluating retrieval of studies for inclusion in systematic reviews. In *40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Tokyo, Japan, 1237–1240. https://doi.org/10.1145/3077136.3080707

[19] Kaveh G Shojania, Margaret Sampson, Mohammed T Ansari, Jun Ji, Steve Doucette, et al. 2007. How quickly do systematic reviews go out of date? A survival analysis. *Annals of Internal Medicine* 147 (2007), 224–233. https://doi.org/10.7326/0003-4819-147-4-200708210-00179

[20] Hanna Suominen, Liadh Kelly, Lorraine Goeuriot, Evangelos Kanoulas, Leif Azzopardi, et al. 2018. Overview of the CLEF eHealth Evaluation Lab 2018. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Springer International Publishing, Cham, 286–301.

[21] Byron C Wallace, Kevin Small, Carla E Brodley, Joseph Lau, Christopher H Schmid, et al. 2012. Toward modernizing the systematic review pipeline in genetics: efficient updating via data mining. *Genetics in Medicine* 14 (2012), 663. https://doi.org/10.1038/gim.2012.7

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