Protocol-driven Searches for Medical & Health-Sciences Systematic Reviews

Matt-Mouley Bouamrane[1], Craig Macdonald[2], Iadh Ounis[2], Frances Mair[1]

[1]Institute of Health and Wellbeing
College of Medical, Veterinary and Life Sciences
[2]School of Computing Science
University of Glasgow, Scotland, UK
{*FirstName.LastName*}*@glasgow.ac.uk*

Abstract. Systematic reviews are critically important search tasks in medicine and health services research. Along with large and well conducted randomised control trials, they are considered to provide the highest levels of clinical evidence. We here provide a brief overview of the methodologies used to conduct systematic reviews, and report on our recent experience of conducting a *meta-review* – i.e. a systematic review of reviews – of pre-operative assessment. We discuss issues associated with the large *human manual* effort currently necessary to conduct systematic reviews when using available search engines. We therefore suggest ways in which more dedicated and sophisticated information retrieval tools may enhance the efficiency of systematic searches and increase the recall of results. Finally, we discuss the potential for the development of systematic reviews tests collections – as well as standard evaluation methodologies – as future benchmarks for comparative studies and evaluation of automated information retrieval tools.

1 Introduction

Systematic reviews (SR) and meta-analyses (MA) of the medical literature are considered to provide – along with large and well-conducted Randomised Control Trials (RCT) – the highest existing level of clinical evidence (level I) [1]. SRs are now routinely used as the starting point for developing clinical guidelines [2]. Guidelines affect the promotion of health care interventions by policy-makers and clinical managers, as well as the provision of care to patients at the point of care. Where systematic reviews fail to produce sufficient evidence to issue guidelines, clinical recommendations are typically based on expert opinions, considered to be the lowest level of clinical evidence, (level IV) [1]. SRs often do not provide definitive and authoritative answers to a research question, for a lack of sufficient available or reliable evidence in the scientific literature. In these cases, the SR highlights gaps in the existing evidence, which in turn may subsequently shape the agenda for medical interventions, services organisation, policy development, as well as research funding priorities [3].

From an information retrieval (IR) perspective, a SR is a search task, with clearly defined information requirements *(the research question)*, which are explicitly specified as systematically-developed and constrained notions of relevance, in the form of a *search protocol*. Indeed, the process underpinning a SR

is guided by published peer-standards, including a *protocol* for deriving search queries and the relevance screening of search results. Hence, we describe a SR to be a *protocol-driven* search task. Moreover, a SR can be seen as a *recall-focused* task, as *all relevant* literature must be found.

This paper contributes as an overview on the background, motivations and methodologies for conducting SRs, which we believe are both unfamiliar and useful to the IR community. Moreover, we make comparisons with other recall-focused IR tasks, and discuss how IR research can potentially contribute to aiding SRs.

The structure of this paper consists of: a brief introduction to the motivations of SRs; an overview of the methodologies used to conduct SRs in medicine and health services research; we discuss our recent experience of conducting a meta-review of preoperative assessment and the challenges encountered using currently available search tools; finally, we discuss the potential development of systematic reviews tests collections – as well as standard evaluation methodologies – as future benchmarks for comparative studies and evaluation of automated information retrieval tools.

2 Systematic Reviews

2.1 Issues with the Reporting of Clinical Outcomes

Several studies have highlighted substantial issues within the reporting of clinical outcomes in the scientific literature [4]. *Publication bias* is the tendency for scientific publications to be biased towards the reporting of statistically significant results or studies with a proven demonstration of practical efficiency [5]. This effect can be substantial with evidence of highly exaggerated effectiveness of treatments [6]. *Outcomes reporting bias* occurs when only a selected subset of measures are reported, which produces an *incomplete* or *inaccurate* evaluation of study outcomes [7].

To minimise the potential for occurrences of these biases to misrepresent or exaggerate the effectiveness of treatments, the assessment of clinical evidence in the medical literature is increasingly relying on systematic reviews and meta-analyses of the literature. A SR identifies and aggregates *all available evidence* pertaining to a specific research question, using a rigorous and transparent methodological *protocol* guided by peer standards. The protocols specifies clear eligibility and exclusion criteria – in order to provide reliable, accurate, and critically appraised evidence-based clinical reports with a minimum of bias [8, 9]. A *meta-analysis* (MA) uses statistical methods to aggregate the quantitative results of independent RCT studies [10]. SRs are now common in medicine and many other fields. Indeed, using data from 2004, Moher et al. [11] estimated that in excess of 2500 SRs were published annually – a figure which is likely to have risen since the results of this study was published.

2.2 IR Searches in Systematic Reviews

From an IR perspective, a SR represents an instance of a search task with well-defined information needs and highly constrained definitions of relevance. Moreover, as an SR must assert that *all potentially relevant documents* are retrieved

– i.e. full recall *must* be achieved – typically all papers matching the query are examined, leading to very low overall precision of the results. The entire retrieved set is screened for relevance, with many potentially topically relevant papers being excluded if they do not meet strict pre-defined inclusion or exclusion criteria. For instance, exclusion criteria may be based on methodological or studies' type criteria (e.g. RCT, case-control, cohort studies, etc.) This assures the ultimate integrity of the clinical evidence, by discarding lower quality studies or inadequate methodological approaches which could undermine the validity of the clinical evidence. The latter can be particularly difficult for existing search engines to detect. In many cases, documents may have been indexed with some meta-data, such as study type or study categories (e.g. Medical Subject Headings (MESH) terms[1]) but this meta-data often remains insufficiently reliable for practical high precision searches, often due to the coarse granularity of the indexing categories given the high specificity of the SR search task [12].

Overall, to attain quality and reproducible SRs, the entire search process is driven by the search protocol. Moreover, these searches are "manually" labour-intensive, with low precision, and are mainly conducted – or at least designed and overseen – by domain-specific experts. Hence they are *expensive* in time, labour and expertise. In this paper, we describe SRs as representing an archetypal example of a *recall-focused* and *protocol-driven* search task. The IR community can make a significant contribution to supporting search tasks underpinning SRs, if it were capable of developing tools to *optimise searches*, increase the *precision* of searches, while guaranteeing the full recall of *all relevant* documents. In the following, we review the existing standard protocols for systematic reviews, before reporting on the authors' recent experience of conducting a SR, and discussing how IR can contribute to the process of performing systematic reviews.

3 Protocols for Systematic Reviews

As highlighted above, SRs must abide by a search protocol, which detail the survey methodology. In particular, steps such as formulating the query from the information need, screening of results and reporting of conclusions are discussed. In this section, we provide details on the current PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) protocol. This will facilitate the explanation of the SR case study that follows in Section 4.

The PRISMA statement [13] is a peer-recommended methodology for conducting SRs. PRISMA updated the earlier QUOROM (Quality of Reporting of Meta-analyses) statement [10]. The statements were devised to address the issues of perceived inconsistencies and biases in the reporting of meta-analyses of RCTs. Common issues include the failure of explicitly reporting the status of intervention concealment[2].

[1] `http://www.nlm.nih.gov/mesh/`

[2] Intervention concealment ensure that in an RCTs, the patients receiving - or not - the treatment, *as well* as the health professionals directly involved in the provision of the treatment are both blinded to whether the patient belongs to the intervention or control groups, in order to minimise the bias in the estimates of treatment effects.

To address identified common shortcomings in the methodology of reporting clinical evidence, the PRISMA statement recommends a protocol-based methodological process of reporting critical items identified in the literature reviewed. The omission of these items could undermine the validity of the results reported. PRISMA recommends that SRs report (i) a study selection trial flow in order to determine the criteria that led to studies being included or rejected from the review as well as (ii) a methodological item check list. The items check list provides a quality assessment tool of the searches strategies used to identify clinical evidence, the selection criteria, data and characteristics of studies, as well as processes for validity assessment and quantitative analysis. A structured methodology for reporting meta-analyses is recommended, in order to ensure the consistency and reliability, as well as peer-accepted standard of reporting of SRs and MAs of RCTs.

In addition, PRISMA specifies how the authors of a SR should filter the identified RCTs at each successive step of the review, in the form of a flow diagram, shown in Figure 1. In doing so, the quality of the studies included in the report must be pro-actively assessed in order to exclude lower quality studies, the inclusion of which would risk undermining the validity of the synthesised clinical evidence. Hence, the reliability and validity of the results reported in the review could potentially be critically assessed and guaranteed through a process of third-party replication. Moreover, to enable the full reproducibility of the search, authors need to thoroughly describe their own methodological protocol while conducting the SR, as well as reporting the quality assessment of included studies, against the set of reporting criteria defined in PRISMA. The PRISMA statement recognises that a SR is inherently an iterative search process. The modification of the review protocol in the course of the study is therefore possible as long as it is both justified and explicitly reported. Moreover, the statement stresses that the review protocol ought to be publicly accessible for peer-review.

Finally, the different forms of biases should be addressed to assure the validity of the results reported in the review. In particular, specific attention needs to be paid to minimise the risk of biases by performing both *"study-level"* and *"outcomes-level"* assessment, while selection bias should be explicitly addressed by reporting the publication status of included studies.

As will be seen in the following case study, a SR is a complex IR task, whereby a well-motivated and developed information need is formulated into a cumulative series of queries through an interactive development process. Matching papers are obtained using the queries on a database of publications – such as *Medline*®[3] – which are then exhaustively screened for relevance. The inclusion and exclusion criteria used when screening for relevance are specified *a-priori* in the search protocol. By providing checklists and methodology steps for query formulation, relevance screening and summarising, a search protocols such as PRISMA ensures the reproducibility of a systematic review study, but does not reduce the time or expense in conducting it.

[3] http://www.nlm.nih.gov/bsd/pmresources.html

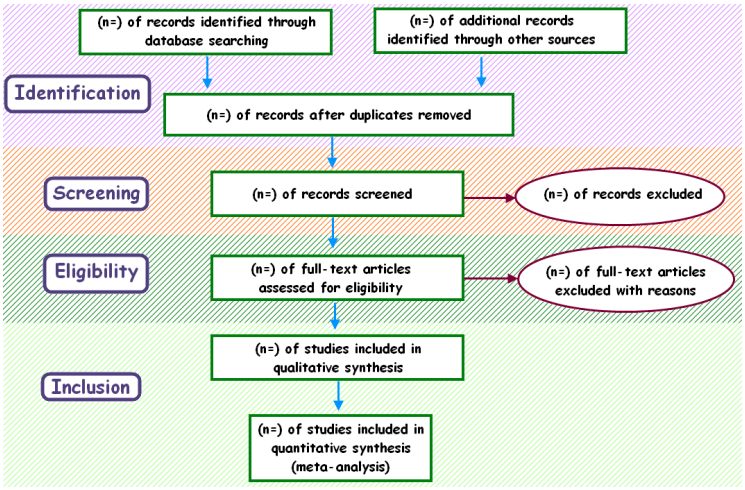

Fig. 1. The PRISMA phases flow of studies selection for systematic review [9].

Currently, IR tools are only involved in the first "Identification" stage of the SR flow process (as per Figure 1). Instead, we believe that they could, and should, provide support for the later Screening, Eligibility and Inclusion stages by providing more advanced retrieval models and search interfaces. Yet, several aspects of a SR iterative process represent substantial challenges for existing IR techniques. In the following section, we summarise our recent experiences in conducting a systematic review. Later, we relate these experiences to other investigated tasks in IR, such as legal and patent retrieval. Moreover, we provide challenges for IR and suggest how models and tools should be enhanced to further support future SR search tasks.

4 Systematic Review Case Study

In this section, we provide the motivations behind our case study SR, and the methodology used. Moreover, we provide an overview of the efforts expended and issues identified while screening a large sample of retrieved documents. We use these to formulate the motivation behind improving IR systems to reduce the efforts of conducting SRs.

4.1 Motivation & Methodology

Our systematic review is concerned with the effectiveness of existing practises of assessing patients before a surgical operation, which is usually referred to as *pre-operative assessment*. The World Health Organisation has estimated that in excess of 230 million surgical procedures are carried out every year [14]. However, patient-factors (e.g. hypertension on the day of surgery) can lead to the cancellation of the surgery. Up to two-thirds of day case cancellations and 50% of inpatients cancellations are caused by patient-related factors and it has been

(anesth* or anaesth* or surgery or surgical or ambulatory or orthopedic procedure* or neurosurg* or preoperative* or elective or minimally invasive of minor surg* or peri-operativ* or pre-procedur* or preoperativ* or preprocedure* or pre-anaesthe* or preanesthe* or preanaesthe* or pre-anaesth* or posteroperative complication* or intraoperative complication* or intra-operative complication*) and (risk* or assess* or test* or scor* or screen* or evaluat* or stratif*)

Fig. 2. Example of a complex Boolean sub-query used within our SR.

suggested that more efficient pre-operative processes could prevent a significant number of these cancellations [15]. Moreover, Pearse et al. found that, while 12.5% of surgical operations were performed on high-risk surgical populations (defined as mortality rates of 5% or over for a specific procedure), this population accounted for more than 80% of post-operative deaths [16].

Between March 2010 and February 2011, we performed a systematic review of the medical literature in search of the available levels of evidence underpinning the effectiveness of existing practises of preoperative assessment. As such, our systematic review is a meta-review, in that we sought to identify all previous SRs and MAs on the specific topic of pre-operative assessment processes, against well defined eligibility criteria. We developed a search protocol according to the NHS Centre For Reviews and Dissemination guidance for undertaking reviews in health care [8]. This guidance provides step-by-step instructions in developing a search protocol, which we used to complement the PRISMA checklist.

We performed an Medline database search in July 2010 using the Ovid search portal tool[4]. Medline is the U.S. National Library of Medicine's bibliographic database. It contains over 18 million references to articles in biomedicine and life sciences. The search was performed and refined over a series of five meetings between the SR protocol team and an information scientist of the University of Glasgow who specialises in bibliographic searches for the life sciences. We initially used a broad search strategy for identifying reviews of preoperative assessment using generic query terms such as "preoperative risk assessment, evaluation, screening, testing," etc., combining keywords, MESH terms, and study types such as "reviews" and "meta-analyses". MESH terms returned entirely unmanageable numbers of results (i.e. in excess of 30 000). The search strategy was progressively refined, until a manageable 8522 abstracts were retrieved from Medline (after restricting to English language). The final used query comprised a series and combination of over 20 complex Boolean sub-queries. An example of one subquery is given in Figure 2, demonstrating manual stemming and Boolean aspects.

The retrieved titles and abstracts were screened independently by two expert reviewers, using strict inclusion and exclusion criteria, such that only SRs and MAs of pre-operative assessments were included in the study.

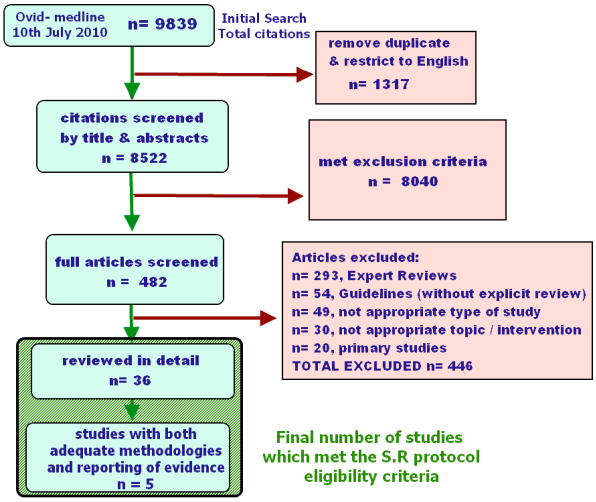

Fig. 3. Flow chart of our SR on pre-operative assessments.

4.2 Results Overview

The detailed results are reported in details in [17]. However, Figure 3 shows the trial flow (as per the PRISMA protocol), and the number (n) of studies included or excluded at each stage of the study. In particular, a majority (n=8040) of titles and abstracts screened, did not meet our strict inclusion criteria, because of either 2 reasons (i) *low relevance*: the studies focused on a clinical intervention that was only marginally relevant to preoperative assessment (e.g. preoperative therapy, intervention or medication, intra- or post-operative interventions and treatments, etc.) or was not the appropriate type of study (e.g. a primary study of a clinical intervention rather than a systematic review of the effectiveness of processes) or (ii) *too high specificity*: the studies described highly specific clinical processes, interventions or populations (e.g. surgery on a specific organ for a specific type of disease) which were not generic and thus not deemed useful for the purpose of our review (i.e. evidence under-pinning standard processes of generic preoperative assessment for elective surgery). The full text of a further n=482 studies were screened, with only 36 studies meeting the final study selection criteria both for relevance and quality of methodology. These 36 studies were reviewed in detail, from which n=5 were relevant with respect to all inclusion and exclusion criteria, including the reporting of adequate clinical levels of evidence. In conducting this systematic review, we identified several issues:

High Screening Workload: The screening of 8522 abstracts, retrieving and assessing 482 full articles and selecting and extracting data from the final set of 36 included studies - with respect to the search protocol - is an extremely laborious process. Indeed, it took 2 doctoral researchers almost 6 months of full-time work to complete the screening. Each abstract or paper was reviewed by both researchers, in order to minimise the risk of bias, as recommended by

[4] `http://www.ovid.com/site/products/ovidguide/medline.htm`

Table 1. Ranks of papers examined within the ranking ranges of the initial Ovid Medline search.

Criteria	Ranking ranges
n=482 full paper screened	rank 35 to 8457
n=36 relevant & met inclusion criteria	rank 419 to 8457
n=5 met inclusion criteria & specified level of clinical evidence + n= 3 further documents identified	rank 521 to 4096 (521, 1989, 2281, 3502, 4096)

PRISMA. Initially, a large amount of time was spent screening and discarding clearly irrelevant studies. Of the full paper screening, many of these studies were subsequently rejected based on methodological criteria. Here, although the topic of the study was clearly relevant to the search, the quality of the methodology of the studies was deemed insufficient according to the search protocol criteria, to guarantee the identification of reliable clinical evidence.

Relevance Ranking: This comment above explains why the studies that fully met our inclusion criteria for relevance (n=419) and clinical evidence (n=521) were very lowly ranked in the Ovid search results (see Table 1). In addition, one relevant document was identified at a very low rank (8427) which would suggest that the number of results examined was justified for achieving high recall.

Search Snapshots: A practical issued faced during the systematic review was the reproducibility of searches using the Ovid tool. Indeed, while it is possible to save queries in Ovid (i.e. as the combination of complex Boolean queries), it is not possible to save the actual *search results* themselves. About 700 000 new records are added to Medline every year, which translates to in excess of tens of thousands of new studies every week. Instead, to facilitate easy management of the systematic review, and reproducibility of the results, the ability to obtain a snapshot of the search results at a given point in time would have been very useful.

Manual indexing: Many articles indexed as "reviews" were not necessarily reviews - never mind systematic reviews - while search queries typically retrieved a large volume of articles that had only a very remote relevance to that of the core topic of pre-operative assessment. Of those articles which were deemed relevant to our research topic (full-paper screening), a vast majority were expert opinion reviews or expert opinion-based guidelines (over 70% of full papers screened) and thus did not meet the study type selection criteria (SR or MA).

Overall, of the 8522 papers screened in the initial Ovid Medline search, only 5 studies both adequately reported the study search and studies selection criteria, as well as providing explicit and adequate grading of clinical recommendations. The mean rank at which these 5 studies were retrieved in the initial Medline search was 2478, demonstrating the lack of precision in the search results. A further 3 relevant studies were identified through a complementary search in other local repositories - a problem also noted by [18]. Effectively, this provided us with less then 1 in 236 (8522/36) studies meeting our search protocol inclusion criteria and less than 1 in a 1000 documents explicitly providing levels of clinical evidence in our systematic search. Moreover, only the most basic functionalities

of the IR process were used. We believe that new retrieval models and search engines that can support complex constraints and the recall focused nature of the task can significantly benefit researchers conducting SRs. For instance, a faceted search interface which contains facets pertaining to common inclusion or exclusion criteria (e.g. built using classification techniques) would allow the exploration and iterative reduction of the set to be screened. In the next section, we discuss other recall-focused tasks that have recently been investigated in IR, describe a roadmap for improving the IR technology used for SRs, and discuss possible evaluation methodologies for such techniques.

5 Towards Protocol-Driven IR

In this section, we compare and contrast the systematic review of medical literature with other domain-specific recall-focused IR tasks, then we propose a roadmap of how IR research can address the systematic review search task. Finally, we discuss how evaluation methodologies may be enhanced to facilitate IR research in systematic reviews.

5.1 Recall-Focussed Search Tasks

e-Discovery is the process of a negotiated discovery of electronically-stored documentary evidence during a legal case. In particular, lawyers negotiate a complex Boolean query, which aims to identifies relevant (known as 'responsive') documents to the plaintiff party, and excludes private documents belonging to the defendant that should not be revealed to the plaintiff. Similar to systematic reviews, recall is important, as a legal argument may hinge on the discovery of a supporting responsive document, however the search protocols within systematic reviews place more constraints on relevance. Since 2006, the TREC Legal track has operationalised the e-Discovery task within an IR research setting. Results thus far indicate that the negotiated Boolean queries can miss up to 50% of the responsive documents (a similar problem has been reported for SRs [19]), but that single retrieval systems could only demonstrate small improvements over the retrieval using the negotiated Boolean queries [20].

Patent Prior-Art Search is the process whereby other patents relating to a given patent are identified. Such patents may be cited within the patent, or maybe used to invalidate the patent. Once again, recall is an essential aspect of this task. However, in contrast to a systematic review or e-Discovery, the given patent can be used as the query, instead of a complex Boolean query developed within a search protocol or by legal negotiations. Patent prior-art search has been investigated within in several evaluation forums (e.g. TREC & CLEF). In particular, the TREC Chemical track has ran since 2009 [21], focusing on chemical patents alone. Participating groups made use of citations, as well as advanced entity tagging of chemical entities [21].

5.2 Roadmap of IR Research for Systematic Reviews

The systematic review search task is characterised by several dimensions. In particular, while recall is very important and the notion of relevance is very

constrained, for every SR there is some practical maximum number of abstracts that can be screened. In the following, we enumerate ways in which recall can be enhanced, screening effort can be minimised and the IR system can be more fully utilised in the SR process:

Maximising Screening Effort: In SRs, Boolean queries have been classically used to limit the size of the retrieval set (but at the risk of reduced recall [22]). However, more intelligent methods of deciding on a cutoff for the retrieved set are possible. For instance, [23] suggests using already-identified studies to find a cutoff point which ensures recall but minimises the size of the retrieved set to be screened. Instead, as an alternative, in a similar manner to the Legal track, relaxing Boolean queries to use proper relevance rankings should permit more relevant papers to be identified [22]. Moreover, we believe that it is possible to make a probabilistic guarantees on the number of identified relevant documents attained by a given rank cutoff, inspired by [24].

Enhancing Relevance Ranking: [23] found that simply adding the term 'versus' to the query improved the results quality for 61 SR searches in Medline. While heuristical, this suggests that enhanced query reformulation and ranking models could enhance SR searches. For instance, classical recall enhancement techniques such as query expansion and collection enrichment [25] may introduce further relevant documents not found using the strict Boolean queries. Moreover, the work of the TREC Genomics track (2003-2007) [27] is of note, for its handling of genome-related retrieval from Medline, however it did not tackle recall-focused tasks such as SRs. Lastly, with the prevalence of feature-based models in modern IR, we see the potential for further improving retrieval by the deployment of learning to rank techniques [26] adapted to high recall environments.

Constrained Relevance: With many dimensions of relevance prescribed by the search protocol and the various inclusions and exclusion criteria, the IR system should aim to facilitate common criteria by providing various document classification models [28]. For instance, in our systematic review, studies that have been conducted or written according to agreed standards (e.g. graded clinical recommendations) are relevant, and could be identified using citation analysis. Moreover, we found that many expert opinions were retrieved. Techniques from NLP [29] and IR [30] for identifying subjective documents may be appropriate at identifying expert opinions (which should not be relevant to a systematic review). This is an example of a negative relevance problem, which has been found to be challenging in areas such as relevance feedback.

Exploratory Search Interface: With many options for constraining the retrieved documents, an improved search engine over Medline could provide a faceted search interface [31], allowing the researchers to iteratively explore the retrieved documents and develop inclusion and exclusion criteria in a manner directly supported by the engine. While faceted retrieval systems have been popular in supporting transactional search tasks such as shopping, recent developments have encompassed their applications to identifying key blogs on the blogosphere [30], as well as automatically suggesting appropriate facets to support on a per-query basis [31].

As can be seen from the list above, research addressing problems in systematic reviews encompass different problem areas in IR, including machine learning, models and interfaces. In the next section, we provide suggestions on how the success of research within the systematic review search task may be evaluated through appropriate evaluation methodologies.

5.3 Evaluation Methodologies

Given the volume [11] and the expense [23] of the systematic reviews that are conducted every year, as well as the potential for IR technology to improve the systematic review process, we believe that there is a case for the development of standard IR test collections covering this search task. A test collection for systematic reviews could leverage the experience garnered by the TREC Legal and Chemical tracks in evaluating recall-focused tasks. Of note, a characteristic common to both the Legal and Chemical tracks is the use of stratified sampling in the relevance assessment of the documents identified by the participating systems, to reduce the assessing workload to a manageable level. However, by using such sampling methods means that only estimates of recall of the participating systems are obtained in practice. In contrast, a systematic review test collection could be created in co-operation with an active systematic review, thereby potentially enhancing the recall of the study, as well as obtaining the relevance assessments as a side-product of the review. Another alternative methodology is described by Boudin et al. [32], where relevance assessments for systematic reviews are bootstrapped from those already published, but with the disadvantage of missing relevant documents not identified by the originial systematic reviews.

In the same way that judges are recognising that technology derived from the TREC Legal track may become acceptable for e-Discovery [33], once lessons learned from IR result in improved search systems available to medical researchers conducting systematic reviews, revisions to the PRISMA search protocols may relax the burden in the searching for relevant literature in systematic reviews when enhanced IR tools are utilised.

6 Conclusions

We have described the motivations behind conducting systematic reviews of the medical literature, namely: the identification of *all relevant* clinical results pertaining to a highly specific research question, typically the clinical effectiveness of treatments or interventions. We have provided an overview of peer-accepted standards search protocols that are observed when conducting SRs. Moreover, as a case study, we reported our experience of conducting a recent SR. We compared SRs to similar recall-focused IR tasks, and provided a roadmap for future IR research, based on enhancing SRs search tasks. Finally, we discussed the possibility and benefits of conducting a TREC-style evaluation efforts for systematic reviews search tasks.

Acknowledgements

This research is funded by the Scotland Chief Scientist Office (CSO) through a postdoctoral training fellowship (2010/2013 - M-M. Bouamrane).

References

1. Evans, D.: Hierarchy of evidence: a framework for ranking evidence evaluating healthcare interventions. Journal of Clinical Nursing **12**(1) (2003) 77–84
2. NICE: The guideline development process - an overview for stakeholders, the public and the NHS (3rd edition). National Institute for Health and Clinical Excellence (2007)
3. Doyle, J., Waters, E., Yach, D., McQueen, D., De Francisco, A., Stewart, T., Reddy, P., Gulmezoglu, A.M., Galea, G., Portela, A.: Global priority setting for Cochrane systematic reviews of health promotion and public health research. Journal of Epidemiology and Community Health **59**(3) (2005) 193–197
4. Chan, A.W., Hrbjartsson, A., Haahr, M.T., Gtzsche, P.C., Altman, D.G.: Empirical evidence for selective reporting of outcomes in randomized trials. Journal of the American Medical Association **291**(20) (2004) 2457–2465
5. Sterne, J.A.C., Egger, M., Smith, G.D.: Investigating and dealing with publication and other biases in meta-analysis. BMJ **323**(7304) (2001) 101–105
6. Begg, C.B., Berlin, J.A.: Publication bias and dissemination of clinical research. Journal of the National Cancer Institute **81**(2) (1989) 107–115
7. Kirkham, J.J., Dwan, K.M., Altman, D.G., Gamble, C., Dodd, S., Smyth, R., Williamson, P.R.: The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. BMJ **340** (2010)
8. NHS-CRD: Centre for Reviews and Dissemination's guidance for undertaking systematic reviews in health care. University of York (2009), http://www.york.ac.uk/inst/crd/
9. Liberati, A., Altman, D.G., Tetzlaff, J., Mulrow, C., Gotzsche, P.C., Ioannidis, J.P., Clarke, M., Devereaux, P., Kleijnen, J., Moher, D.: The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Journal of Clinical Epidemiology **62**(10) (2009) e1–e34
10. Moher, D., Cook, D.J., Eastwood, S., Olkin, I., Rennie, D., Stroup, D.F., the QUOROM Group: Improving the quality of report of meta-analyses or randomised controlled trials: the QUOROM statement. The Lancet **354** (1999) 1896–1900
11. Moher, D., Tetzlaff, J., Tricco, A.C., Sampson, M., Altman, D.G.: Epidemiology & reporting characteristics of systematic reviews. PLoS Medicine **4**(3) (2007) e78
12. Liu, Y.H.: On the potential search effectiveness of MeSH (medical subject headings) terms. In: Proceedings of IIiX 2010, New York, NY, USA, ACM (2010) 225–234
13. Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., The PRISMA Group: Preferred reporting items for systematic reviews & meta-analyses: The PRISMA statement. PLoS Medicine **6**(7) (2009)
14. World Health Organization: Safe surgery saves lives. WHO world alliance for patient safety. WHO report (2008)
15. NHS Modernisation Agency: National good practice guidance on pre-operative assessment for in patient surgery. (2003)
16. Pearse, R., Harrison, D., James, P., Watson, D., Hinds, C., Rhodes, A., Grounds, R.M., Bennett, E.D.: Identification and characterisation of the high-risk surgical population in the United Kingdom. Critical Care **10**(3) (2006) R81
17. Bouamrane, M.-M., Gallacher, K., Marlborough, H., Jani, B. , Kinsella, J., Richards, R., van Klei, W., Mair, F.S. Processes of preoperative assessment in elective surgery: a systematic review of reviews. (2011) Under review.
18. Beahler, C.C., Sundheim, J.J., Trapp, N.I.: Information retrieval in systematic reviews: Challenges in public health arena. American Journal on Preventitive Medicine **18** (2000) 6–10

19. Golder, S., McIntosh, H., Loke, Y.: Identifying systematic reviews of the adverse effects of health care interventions. BMC Medical Research Methodology **6**(1) (2006) 22
20. Oard, D.W., Baron, J.R., Lewis, D.D.: Some lessons learned to date from the TREC Legal track (2006-2009). University of Maryland (2010)
21. Lupu, M., Huang, J., Zhu, J., Tait, J.: TREC-CHEM: large scale chemical information retrieval evaluation at TREC. SIGIR Forum **43** (2009) 63–70
22. Pohl, S., Zobel, J., Moffat, A.: Extended Boolean retrieval for systematic biomedical reviews. In: Proceedings of ACCS 2010. (2010) 117–126
23. Zhang, L., Ajiferuke, I., Sampson, M.: Optimizing search strategies to identify randomized controlled trials in MEDLINE. BMC Medical Research Methodology **6**(23) (2006)
24. Arampatzis, A., Kamps, J., Robertson, S.: Where to stop reading a ranked list?: threshold optimization using truncated score distributions. In: Proceedings of SIGIR 2009. (2009) 524–531
25. Kwok, K.L., Grunfeld, K., Chan, M., Dinstl, N., Cool, C.: TREC-7 adhoc, high precision & filtering experiments using PIRCS. In: Proceedings of TREC-7. (1998)
26. Liu, T.Y.: Learning to rank for information retrieval. Foundations and Trends in Information Retrieval **3**(3) (2009) 225–331
27. Roberts, P.M., Cohen, A.M., Hersh, W.R.: Tasks, topics and relevance judging for the trec genomics track: five years of experience evaluating biomedical text information retrieval systems. Inf. Retr. **12**(1) (2009) 81–97
28. Cohen, A., Hersh, W., Peterson, K., Yen, P.Y.: Reducing workload in systematic review preparation using automated citation classification. Journal of the American Medical Informatics Association **13**(2) (2006) 206 – 219
29. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval **2**(1-2) (2008) 1–135
30. Macdonald, C., Santos, R.L.T., Ounis, I., Soboroff, I.: Blog track research at TREC. SIGIR Forum **44** (2010)
31. Hearst, M., Elliott, A., English, J., Sinha, R., Swearingen, K., Yee, K.P.: Finding the flow in web site search. Commun. ACM **45** (2002) 42–49
32. Boudin, F., Nie, J.Y., Dawes, M.: Deriving a test collection for clinical information retrieval from systematic reviews. In: Proceedings of DTMBIO 2010. (2010) 57–60
33. Oard, D.W., Hedin, B., Tomlinson, S., Baron, J.R.: Overivew of the TREC 2008 Legal track. In: Proceedings of TREC 2008. (2008)