




A comparison of automatic Boolean query formulation for systematic reviews

Harrisen Scells^{1,2}  · Guido Zuccon¹ · Bevan Koopman²

Received: 1 July 2020 / Accepted: 6 October 2020 / Published online: 27 October 2020
© Springer Nature B.V. 2020

Abstract

Systematic reviews are comprehensive literature reviews that target a highly focused research question. In the medical domain, complex Boolean queries are used to identify studies. To ensure comprehensiveness, all studies retrieved are screened for inclusion or exclusion in the review. Developing Boolean queries for this task requires the expertise of trained information specialists. However, even for these expert searchers, query formulation can be difficult and lengthy: especially when dealing with areas of medicine that they may not be knowledgeable about. To this end, two computational adaptations of methods information specialists use to formulate Boolean queries have been proposed in prior work. These adaptations can be used to assist information specialists by providing a good starting point for query development. However, a number of limitations with these computational methods have been raised, and a comparison between them has not been made. In this study, we address the limitations of previous work and evaluate the two. We found that, between the two computational adaptations, the objective method is more effective than the conceptual method for query formulation alone, however, the conceptual method provides a better starting point for manual query refinement. This work helps to inform those building search tools that assist with systematic review construction.

Keywords Information retrieval · Systematic reviews · Boolean queries · Query formulation

✉ Harrisen Scells
h.scells@uq.net.au

Guido Zuccon
g.zuccon@uq.edu.au

Bevan Koopman
bevan.koopman@csiro.au

¹ The University of Queensland, Brisbane, QLD, Australia

² CSIRO, Brisbane, QLD, Australia

1 Introduction

Systematic reviews are highly important within the medical domain. They are used to inform clinical decision making, and are seen as the highest form of medical evidence (Lavis et al. 2005). The process for developing a systematic review has many steps, and requires the support of clinical researchers, librarians, and review committees (McGowan and Sampson 2005). Systematic reviews are guided by a highly specific research question, and executed through a methodological study protocol (Chandler et al. 2019). Figure 1 illustrates the particular phase in systematic review creation that this study targets. Arguably, one of the most important processes in the creation of a systematic review is the identification of medical literature which will be synthesised later in the process. This identification process involves searching and screening studies (e.g., randomised controlled trials) from large medical databases (e.g., PubMed, which contains approximately 30 million studies at the time of writing). Screening literature is an important task that constitute a significant amount of time and effort in the systematic review creation process. To complete this task, a Boolean query is used, as it allow for complete control over the search system and enables the explicit encapsulation of the information need of the research question into the query syntax. The Boolean query has a major impact on the screening process: a query may retrieve all of the relevant studies but may also retrieve an excessive amount of non-relevant studies. It is typical for the screening process to involve upwards of 10,000–1,000,000 studies that require screening. Moreover, the screening process is usually performed twice or thrice by independent screeners to reduce bias. Although there has been much progress to reduce the workload associated with this screening process, including screening prioritisation (Scells et al. 2020a; Kanoulas et al. 2017, 2018; Lee and Sun 2018), active learning for study inclusion classification (Cohen et al. 2006; Miwa et al. 2014), text mining (O'Mara-Eves et al. 2015; Olorisade et al. 2016; Shemilt et al. 2014), and automated second screeners (Wallace et al. 2010), the search query can have a much more significant effect on screening workload reduction, simply by reducing the number of studies retrieved (Scells and Zucco 2018a; Scells et al. 2019). However, effective query development is extremely difficult and time consuming (Golder et al. 2008; Bullers et al. 2018). It requires the

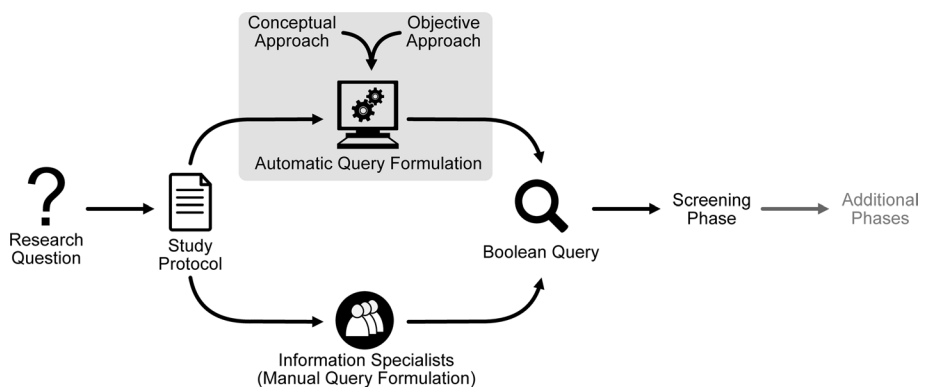


Fig. 1 The initial phases of systematic review creation that this study focuses on. The highlighted area indicates the aspect of the creation phase that we focus on: query formulation. In particular, we propose to automate a currently manual task, indicated below the highlighted area

expertise of trained librarians (i.e., information specialists), and the use of highly complex Boolean queries (Lefebvre et al. 2008)

1.1 Automatic query formulation

Recently, there have been efforts to automate the processes information specialists use to develop Boolean queries (Scells et al. 2020b, c). These methods are fully automatic, computational adaptations of two manual approaches to query development that are employed by information specialists. The first manual approach is a *conceptual* approach (Hausner et al. 2012, 2015), where a query is developed by identifying high-level concepts, and finding synonyms for these concepts. The second manual approach is an *objective* approach (Clark 2013), where a query is developed by identifying and classifying terms using a statistical approach. Both of these manual query formulation methods are used extensively for Boolean query development for systematic review literature search. The automatic query formulation methods that have been derived from these approaches have been shown to achieve similar performance. However, a comparison between both automatic query formulation methods has not been performed. Furthermore, a number of limitations have been raised that warrant implementation and further comparison to manually formulated queries. To this end, this article seeks to not only compare the differences between queries derived from the two automatic query formulation methods with each other, but also between queries derived from each of the automatic query formulation methods and the original, manually formulated queries.

1.2 Limitations of prior research

Both computational adaptations were suggested to be appropriate as a starting point for refinement and use in literature searches. However, the two approaches have not been compared to each other to determine which one is more appropriate for the task of systematic review literature search, and the suggestion that they can be used as a good starting point for further formulation or refinement was not appropriately evaluated in past research. For example, prior research did not study the effectiveness of the automatic query formulation methods after the resulting queries had been manually refined by a human. Furthermore, a number of extensions to both computational adaptations have been suggested that may further improve the effectiveness of the automatically formulated queries; namely, the use of *phrases* in addition to terms for keywords in Boolean clauses, and the use of *seed studies* as relevance feedback to tune queries. Seed studies are commonly used in the development of queries for systematic review literature search. Specifically, they are highly relevant studies that are identified before starting the review. However, there currently does not exist a test collection with the original seed studies used for manual query formulation in this domain. For this reason, we instead study how sensitive the two automatic query formulation approaches are to different initial seed studies.

1.3 Research questions

Investigation into the comparison of automatic query formulation methods between themselves and manually formulated queries, as well as addressing the limitations of prior research underpin the research questions of this study:

- RQ1* How does automatic query formulation compare to manual query formulation in terms of search effectiveness?
- RQ2* What factors of automatically formulated queries contribute the most to effectiveness?
- RQ3* Which automatic query formulation method provides the most effective starting point for manual refinement?
- RQ4* How sensitive to variation in the initial seed studies are the automatic query formulation approaches?

The first two research questions guide the investigation into the comparison of automatic query formulation methods between themselves and manually formulated queries. Specifically, *RQ1* aims to identify which automatic query formulation is most effective when compared to a manually formulated query. This is achieved through a batch-style evaluation of automatically formulated and manual queries. Meanwhile, *RQ2* aims to identify the factors of automatically formulated queries that contribute to their effectiveness, e.g., choice of terms vs. phrases or the number of seed studies. Note that here the focus is on the comparison between the two automatic query formulation methods, and not the manually formulated queries. The next two research questions guide the investigation into limitations identified in prior research. Specifically, *RQ3* aims to identify which automatic query formulation method provides the most effective query once manually refined, and how these manually refined queries compare to the same originally manually formulated queries. This is achieved through a small-scale case study. Finally, *RQ4* aims to identify how sensitive each automatic query formulation method is to the initial set of seed studies. For this, different portions of relevance judgements are sampled for seed studies as input, and statistical variances are studied. In answering these research questions, this study makes the following contributions:

- Extensions to automatic query formulation methods that have been identified as limitations in previous work.
- A comprehensive comparison of two automatic query formulation methods (as extended in this study) to manually formulated queries, and between each other.
- A case study investigating the suggestion that the automatic query formulation methods are good starting points for further manual refinement.
- A comprehensive investigation into the effect seed studies have on the automatic query formulation methods, in particular how sensitive the effectiveness of queries are to a given set of seed studies.

2 Related work

Systematic review are critical in medicine and numerous studies have investigated methods to ensure their quality. However, there has been surprisingly little research towards the development and comparison of query formulation methods for systematic review literature search. Two primary methods have arisen to guide information specialists to develop effective queries for search, however, while these methods strive to be methodical, they still are subject to the experience and bias of the information specialist developing the query.

2.1 Conceptual query formulation

The first is the conceptual method (Clark 2013). Here, an initial set of high level concepts are identified that encapsulate the research question of the systematic review. Each high level concept becomes a clause in a conjunctive Boolean query. The high level concepts are then expanded into synonyms by the information specialist manually. Seed studies are used to repeatedly gauge the effectiveness of the query, manually refining as needed, and formulation ends when the information specialist believes that the number of studies retrieved will (i) contain all (or most) of the studies that will be included in the systematic review; and (ii) be screenable within a certain budget and amount of time.

2.2 Objective query formulation

The second is the objective method (Hausner et al. 2012). Here, seed studies are divided into two sets: development and validation. The development set is used to identify terms using statistical methods and the validation set is used to gauge the effectiveness of the query. The information specialist still must decide which terms to add to the query, and, while more objective, still makes the method bias. The information specialist developing a query using this method also attempts to ensure the query retrieves all (or most) of the studies that will be included in the review and that the retrieved studies are screenable within a certain budget and amount of time.

These two methods are the primary methodologies used to develop queries for systematic reviews. While there has been a small study to compare the objective and conceptual methods (Hausner et al. 2015) for 13 topics, there has not been an extensive evaluation or comparison of either method. Recently, fully automatic computational adaptations of the conceptual and objective methods have arisen (Scells et al. 2020b, c). These adaptations allow us to perform an inexpensive large scale evaluation of these methods by simulating the processes information specialists use to develop queries.

2.3 Automatic query formulation in other domains

While the automatic development of queries for the systematic review literature search domain has not been widely investigated, a number of studies have investigated automatic query formulation methods in other domains. For example, Kim et al. (2011) have developed a decision tree based method for automatic query formulation in the legal eDiscovery domain. This method, like the conceptual and objective methods, relies on seed studies to select which keywords to add to a query and the location of those keywords in the query. The difficulty in applying this method to systematic review literature search is that it requires many more seed studies to be effective: in reality this is not feasible. Other works have also investigated the generation of structured queries from natural language statements. These works focus on the generation of SQL queries (Androutopoulos et al. 1995; Pazos et al. 2013; Popescu et al. 2004; Zhong et al. 2017), and this cannot express the full range of requirements needed for systematic review literature search such as field restrictions, complex Boolean clauses, and phrase and free text searching. Finally, closer to the systematic review domain, Scells and Zuccon (2018a); Scells et al. (2019) have investigated the automatic refinement of Boolean queries for systematic review literature

search. This work differs to the work in this article as it was concerned with the refinement to *existing* Boolean queries, where this article focuses on the automatic formulation of Boolean queries from *scratch*.

This study is novel as it is the first of its kind to perform a large scale comparison of two fully automatic Boolean query formulation methods for the systematic review domain. The identification of effective automation methods for systematic review creation can have a significant impact on the costs and time to create systematic reviews (Clark et al. 2020; Tsafnat et al. 2014).

3 Methods

In this section, we describe the conceptual and objective methods, the computational adaptations we make to automatically formulate queries, and the limitations of previous work and how we plan to address them. This section first provides an description of the *manual methodologies* for the conceptual and objective query formulation methods, the *computational adaptations* that have been made to them in prior research, and what *extensions* to each of the methods are made in this research. This provides the basis for the investigation into *RQ1* and *RQ2*, which are concerned with the comparison between automatically formulated queries and manually formulated queries. Following on from these sections, the next two sections provide a method for how the *manual query refinement* process will be undertaken for the case study addressing *RQ3*, and how *seed studies will be sampled* in order to measure how sensitive the automatic query formulation methods are to them for *RQ4*.

3.1 Conceptual query formulation

The conceptual method is the most commonly used approach to develop Boolean queries for systematic review literature search. Under this approach, a number of high-level concepts are identified, either from seed studies or through initial searches, that represent the research question of the review. Often, these concepts are categorised using the PICO question scheme. PICO stands for Population, Interventions, Controls, and Outcomes. It is a way of framing medical questions, in terms of the information needed to answer them. It is common for the title and research question of a systematic review to be framed using PICO. Once the information specialists has identified the high-level concepts they will use to develop the search, they use both their expertise and a number of tools to assist them to identify synonyms and related keywords to their high-level concepts. They add to and refine the query in an iterative process until they feel they are satisfied. This refinement process is achieved using a number of seed studies to gauge the effectiveness of the search. Commonly, only a handful of seed studies are used in the conceptual query formulation process (Clark 2013).

3.1.1 Computational adaptations to the conceptual method

The computational adaptation of the conceptual method is seeded using a single sentence describing the high level research statement the systematic review aims to address, as well as a number of seed studies. Seed studies are split into Development ($\frac{2}{3}$) and Unseen ($\frac{1}{3}$). The Development set is used to perform a term reduction step in query logic composition which removes non-contributing terms from the query. The computational approach of the

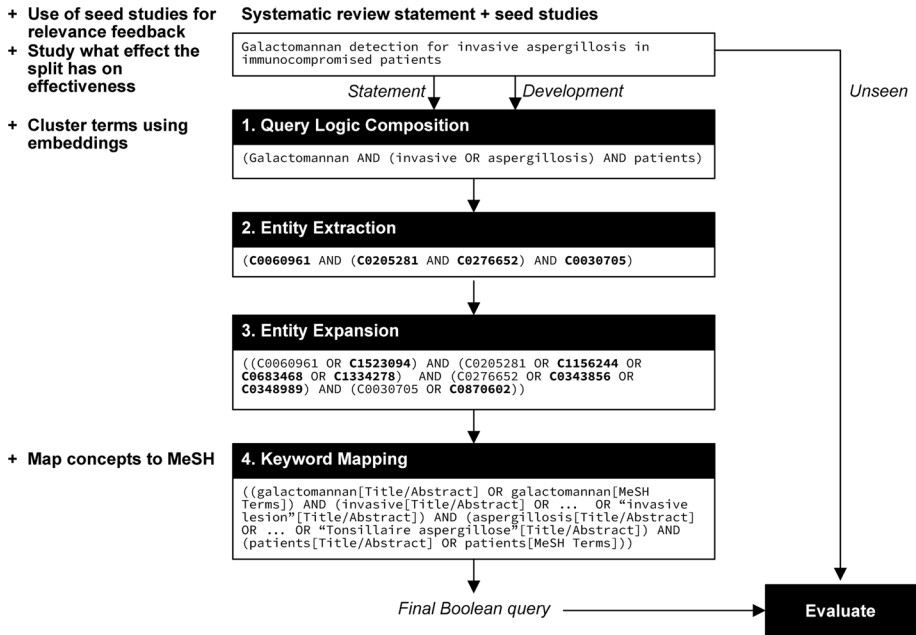


Fig. 2 Overview of the computational adaptation of the conceptual approach. Additions made as part of this paper are indicated with + symbols

conceptual method proposed by Scells et al. (2020b) uses the titles of systematic reviews as the input statement as they are typically written in PICO format. In this paper, we also use the systematic review title. This computational approach differs from the manual approach, primarily it does not perfectly reflect the process an information specialist would use to add keywords to a query as there is no universally agreed upon approach for this. The input seed studies are identified from the relevance assessments for topics. The original conceptual adaptation proposed a pipeline of processing steps to transform a statement into a Boolean query. This pipeline is illustrated in Fig. 2. The steps in the pipeline are as follows:

Query logic composition	Terms from the seed sentence are composed into a logical structure for a Boolean query. This step approximates the information specialist in identifying the initial high-level concepts and categorising them into the different Boolean clauses that will eventually contain the synonyms for these terms. Previous work, for example, uses an unlexicalised, English probabilistic context-free grammar parser to segment word boundaries. These word boundaries are then used to specify where terms should be added into Boolean clauses of a query.
Entity extraction	Once the structure of the Boolean query is defined, the terms in these clauses are mapped to UMLS concepts. ¹ There are a number of methods which perform the mapping of free-text to UMLS concepts such as QuickUMLS (Soldaini and Goharian 2016) and MetaMap (Aronson 2001). In previous work, MetaMap is used for this purpose.
Entity expansion	An optional step, entity expansion uses UMLS concept embeddings to expand the query. This step models the information specialist in identifying synonyms to the chosen high-level concepts. Previous work uses embeddings of UMLS concepts crafted by van der Vegt et al. (2019), obtained by applying word2vec on the entire PubMed database.
Keyword mapping	After the mapping and optional expansion of UMLS concepts, the concepts must then be mapped into appropriate keywords (a single UMLS concept may have a number of aliases, i.e., textual representations—alternate spellings, word orderings, etc.—due to the origin of the concept in different ontologies). The Keyword Mapping step performs this action. Previous work uses a number of techniques including using the preferred term in UMLS, using all of the aliases for a concept, or using only the most frequently used term [a method proposed by Jimmy et al. (2018)].

3.1.2 Extensions to the computational conceptual method

The following extensions have been made to the computational adaptation of the conceptual method, as indicated by the + labels in Fig. 2.

MeSH terms	The method proposed in previous work did not consider MeSH terms. These terms can significantly improve the effectiveness of queries (Scells et al. 2020c). We integrate MeSH terms into query formulation by mapping entities directly to MeSH concepts during the keyword mapping step.
------------	---

¹ UMLS stands for the Unified Medical Language System. It is an umbrella ontology, containing representations of medical terminology using several other ontologies, e.g., MeSH.

- Query logic composition** Previous work used an NLP approach to automatically extract keywords, as well as a manual approach. It was found that the NLP approach was significantly less successful compared to the manual approach at creating the logical structure of a query. Instead in this work, we use an embedding approach to cluster similar terms into the same Boolean clauses. We propose two methods to extract terms and phrases from the input statement: the first splits the statement into uni-grams, and the second splits the statement into phrases using the RAKE algorithm (Rose et al. 2010). Keywords are first mapped to UMLS concepts using an Elasticsearch index. The choice for this method and how it is used is explained in Sect. 4. An embedding for each UMLS concept is then obtained using the model proposed by van der Vegt et al. (2019) Keywords are clustered by measuring the cosine similarity of their embedding between other keyword embeddings. A minimum similarity threshold is used to determine if a keyword belongs in an existing cluster or if a new cluster should be created. In our empirical testing a value of 0.3 was found to provide the best separation of concepts. A keyword is added into the cluster which contains the most similar keyword. If the keyword does not meet the minimum similarity threshold, it is added to a new cluster containing itself.
- Seed studies** The use of seed studies were not previously considered in earlier work, although they are indeed used in reality by information specialists (Chandler et al. 2019). Seed studies can be used to tune the effectiveness of queries at different stages in the computational conceptual pipeline. We extend the computational adaptation of the conceptual method by integrating seed studies into the logical composition step. We use a portion of the relevance assessments to tune the query by reducing keywords that do not contribute to the search while maximising coverage. This is done by first constructing a set of binary keyword vectors K for each seed study corresponding to each extracted keyword; $\mathbf{s}_k \in K$. Once the keywords have been clustered as in the query logic composition step, the result is a set containing each set of clustered keyword vectors $C = \{K_1, K_2, \dots, K_n\}$. The maximum coverage for a new Boolean clause $K_i \in C$ is the logical disjunction of all term vectors corresponding to that clause: $coverage(K_i) = \mathbf{s}_{k_1} \vee \mathbf{s}_{k_2} \vee \dots \vee \mathbf{s}_{k_n}$. Each keyword in the clause is then tested to determine if it reduces the coverage, or in other words, the removal of the keyword causes a change the in coverage vector. If no change is detected, the keyword is discarded. This process is formalised as follows: For each $K_i \in C$ let

$$K'_i = \{\mathbf{s}_{k_i} \in K_i \mid coverage(K_i - \mathbf{s}_{k_i}) \neq coverage(K_i)\}$$

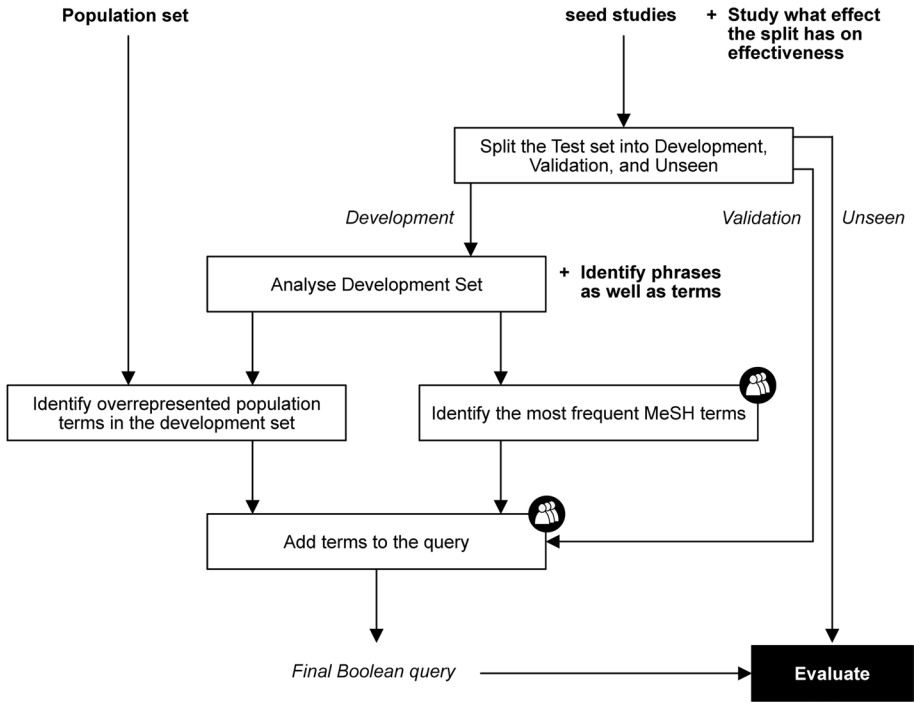



Fig. 3 Overview of the computational adaptation of the conceptual approach. Manual steps associated with the objective method are indicated by ; the computational adaptation seeks to automate these steps. Additions made as part of this paper are indicated with + symbols

Now let $C' = K'_1 \cap K'_2 \cap \dots \cap K'_n$. Each set of keywords $K'_i \in C'$ then becomes a clause in the Boolean query, where each keyword in a K'_i set is joined by an OR operator, and each set of keywords in C' is joined by an AND operator.

3.2 Objective query formulation

The objective method is a relatively recent approach to develop Boolean queries for systematic review literature search. Figure 3 provides a visualisation of the objective query formulation process. The phases in this query development method are more well-defined than the conceptual approach, and therefore it is easier to more closely simulate this method computationally. It involves the use of statistical methods to identify which terms should be added to the query. The objective method uses seed studies to both identify terms and evaluate the effectiveness of terms. Seed studies are split into development ($\frac{2}{3}$) and validation ($\frac{1}{3}$) sets. Over-represented terms are then identified in the titles and abstracts of studies from the development set and then filtered using a population (background) set. The process for identifying over-represented terms is as follows: terms are first ranked using document frequency in the development set. The top 20% of these terms are then re-ranked using document frequency in the population set. The bottom 2% of these terms are those which will be added to the query. At the same time, the 20 most frequent MeSH terms in the development set are also identified to be added to the query. Terms and MeSH

terms are then classified into three categories: (i) health conditions; (ii) treatments; and (iii) study design. Each of these categories becomes a Boolean OR clause. These clauses then become subclauses of an AND clause. The query is then refined using the validation set of seed studies until the information specialists are satisfied with the query.

3.2.1 Computational adaptations to the objective method

The computational adaptation of the objective method is seeded using seed studies and a background collection. A pipeline of statistical term extraction and classification steps is used to transform the contents of studies in the seed studies and background collection into a Boolean query. This pipeline is illustrated in Fig. 3. The steps in the pipeline are as follows:

Parameter tuning	The parameters that control how terms are filtered or how many MeSH terms to add (i.e., those steps indicated by the \ominus symbol are not tuned in the original proposed objective method by Hausner et al. (2012). This is one of the major computational adaptations made. Tuning these parameters using different evaluation measures allows for queries to be automatically designed for specific purposes (i.e., broad searches for far-reaching systematic reviews, or highly specific searches for rapid reviews). Furthermore, to more fairly evaluate queries, the seed studies are split into three sets: development ($\frac{2}{4}$), validation ($\frac{1}{4}$), and <i>unseen</i> ($\frac{1}{4}$). The unseen set is used to evaluate how effective the query is on studies not used to construct or tune the query.
Term categorisation	The categorisation of terms into one of the three <i>health categories</i> is a manual process. To automate this process, the semantic type of a term is used. In previous work, the semantic type for terms is obtained by mapping the term into a UMLS concept using Meta-Map. Terms are added to different clauses depending on the classification of the semantic type. Each of these clauses combines the terms using a Boolean OR operator.

3.2.2 Extensions to the computational objective method

The following extensions have been made to the computational adaptation of the objective method, as indicated by the + labels in Fig. 3.

Phrase search	Queries were only formulated using single terms (i.e., uni-grams). Using phrases (i.e., n-grams) may improve the precision of queries by making them more specific. Here, we address this limitation by using the rapid automatic keyword extraction (RAKE) algorithm (Rose et al. 2010). RAKE extracts n-grams using term co-occurrence and term frequency statistics. Due to the reliance on statistics, rather than linguistic features, RAKE is domain-independent (thus suitable for the medical domain).
---------------	--

3.3 Manual query refinement

Previous work, which proposed the computational adaptations of the conceptual and objective query formulation approaches (Scells et al. 2020b, c), also made the claim that the queries that are automatically formulated provide a good starting point for further manual refinement. To this end, this section describes the query refinement method that will form the basis for a case study to investigate this claim. Queries chosen for refinement have a manual query reduction applied to them. Specifically, keywords in the query that retrieve relatively high numbers of studies and no seed studies are removed. Keywords that also retrieve relatively few numbers of studies and no seed studies are removed. The outcome of the reduction process can improve both precision and recall, depending on where the keyword was removed (i.e. if the keyword was in a clause grouped by an AND or OR operator). To assist in the query refinement process, the searchrefiner tool is used (Scells and Zuccon 2018b). This tool is used by information specialists to refine their own manually formulated searches, and the effectiveness of the tool at this task has been validated by others (Clark et al. 2020). The query refinement process is completed by one of the authors of this study, who also developed the tool. The refinement process is also lengthy (10–30 min per query), therefore a random subsection of automatically formulated queries are chosen for the refinement process.

3.4 Seed study sensitivity analysis

In addition to improving and comparing the automatic conceptual and objective query formulation methods between each other and manually formulated queries, we also perform a sensitivity analysis to determine how seed studies influence the effectiveness of the two automatic query formulation methods. Both computational adaptations of the conceptual approach and the objective approach use seed studies for relevance feedback. Each computational adaptation method closely models the way information specialists use seed studies in the respective manual methods. However, we do not have access to the original seed studies for each topic. The way we address this is by randomly sampling seed studies from the studies included in the systematic review from each topic. We then analyse the effect of this sampling by performing a 1-way ANOVA test. The two groups are the retrieval effectiveness of (i) the set of queries automatically formulated for a topic using different initial seed studies and; (ii) the original query formulated for the topic. Both sets of queries are evaluated on the ‘unseen’ portion of seed studies. Note that the objective method uses a development and validation portion of seed studies, and the conceptual method uses only a development portion. When formulated using the same seed studies, the development portion for the conceptual method is the combined development and validation portion for the objective method. Also note that the manually formulated query does not change depending on the input seed studies in the ANOVA test: in reality this may not be the case; however, it will show how much the automatically formulated queries differ from a method that is ‘perfect’ at formulating effective queries. Manually formulating different queries for each topic for however many samples of seed studies are desired is infeasible as it would be highly costly and time consuming, requiring the expertise of trained information specialists (although may provide for a fairer comparison of how seed studies influence the sensitivity of query effectiveness).

Table 1 Results of each automatic query formulation method averaged across each topic, averaged across each of the 30 iterations of query formulation

		Precision	F _{0.5}	F ₁	F ₃	Recall	WSS
Cptl.	Original	0.0217	0.0267	0.0407	0.1439	0.9338	0.9181
	Phrase	0.0023*	0.0027*	0.0038*	0.0107*	0.5129*	0.4878*
	Term	0.0021*	0.0026*	0.0037*	0.0114*	0.6286*	0.5990*
Objective	Phrase/F ₃	0.0031*	0.0037*	0.0055*	0.0213*	0.3572*†	0.3571*†
	Phrase/F ₃ /MeSH	0.0029*	0.0036*	0.0055*	0.0235*	0.5657*	0.5653*
	Phrase/Recall/	0.0006*	0.0007*	0.0012*	0.0053*	0.5532*	0.5513*
	Phrase/Recall/MeSH	0.0005*	0.0006*	0.0010*	0.0048*	0.7935*†	0.7899*†
	Term/F ₃	0.0053*‡	0.0065*‡	0.0099*‡	0.0365*‡	0.4432*‡	0.4430*‡
	Term/F ₃ /MeSH	0.0050*‡	0.0061*‡	0.0092*‡	0.0356*‡	0.5482*	0.5478*‡
	Term/Recall	0.0004*‡	0.0004*‡	0.0007*‡	0.0032*‡	0.8126*‡	0.8058*‡
	Term/Recall/MeSH	0.0002*‡	0.0003*‡	0.0005*‡	0.0022*‡	0.8780‡	0.8692‡

Two-tailed statistical significance ($p < 0.05$) between the original queries is indicated by *, between conceptual (Cptl.) phrase and objective phrase formulation is indicated by †, between conceptual (Cptl.) term and objective term formulation is indicated by ‡.

4 Experimental setup

Experiments are conducted on the CLEF TAR 2018 set of diagnostic test accuracy systematic reviews (Kanoulas et al. 2018). The CLEF TAR 2018 collection was designed as a shared Information Retrieval task for the purpose of developing methods to support the screening phase of systematic review creation. We adapt this collection for the use of automatic query formulation. The CLEF TAR task has run for three years, however the 2017 collection is a subset of the 2018 collection, and the 2019 collection contains systematic reviews of various types (including the 2017 collection). In this work only the 2018 collection is used so as to control for the type of systematic review [i.e., diagnostic systematic reviews are much more difficult to search literature for than intervention reviews (Leeflang et al. 2013)]. This test collection contains titles, relevance assessments, and queries for 75 systematic reviews. The PubMed entrez API (Sayers 2010) is used for retrieval and statistics (e.g., document frequency). As there are multiple ways for the conceptual and objective methods to be run (i.e., terms versus phrases), we make this distinction clear as an *instantiation* of one of the methods. We define a set of queries formulated with different samples of seed studies for the same topic as an *iteration*. This creates a new problem, however, which is that the query formulation methods may be sensitive to the seed studies used. In total, we perform 30 iterations per query formulation instantiation, per topic. This was found to provide us a sufficiently powered statistical test. Note that the random split for a given iteration is the same across all query formulation instantiations and approaches. Next, to be able to compare the originally formulated queries for each topic to the queries we automatically formulate, we evaluate the original queries on the unseen portion of each iteration (giving 30 runs also for the original queries). In our results, we compare the average performance of a given evaluation measure across all iterations and across all topics, for each instantiation of a query formulation approach.

In previous work, for both computational adaptations, UMLS entities (CUIs) were extracted using MetaMap. There is a major limitation of MetaMap: it is not computationally

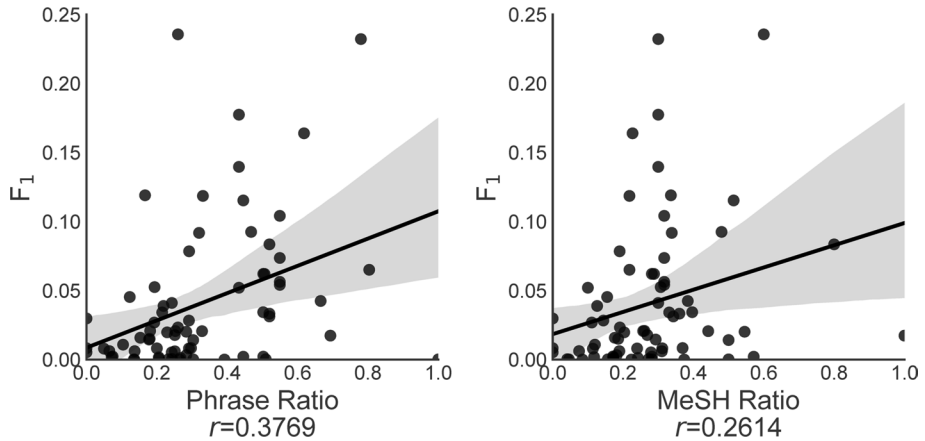


Fig. 4 Correlations between ratio of phrases ($\frac{|\text{phrases}|}{|\text{keywords}|}$) and ratio of MeSH terms ($\frac{|\text{MeSH terms}|}{|\text{keywords}|}$) in the original, manually formulated queries, and the effectiveness of those queries. Pearson's correlation coefficient r is indicated beneath each plot

efficient. In this work, for mapping terms to and from UMLS entities, we use a custom Elasticsearch index containing the UMLS terminology. The matching of terms to entities is handled by the default ranking function of Elasticsearch 7.5.2 (BM25). For mapping term to a single entity, we always choose the top-most ranked entity. This method of entity mapping has been shown to be empirically comparable to MetaMap (Mirhosseini et al. 2014).

5 Results

5.1 RQ1: comparison to original queries

This section aims to address the RQ1: *How does automatic query formulation compare to manual query formulation in terms of search effectiveness?* We address this question by comparing the automatic query formulation methods to the original, manually formulated queries. This comparison made in Table 1. The results in this table are the average performance across all topics and all iterations of seed study splits. Bold values indicate the highest performing method for the conceptual and objective approaches.

We find that none of the automatic query formulation methods can outperform the original, manually formulated queries. Indeed, often the queries formulated automatically are significantly worse than the original queries. The highest performing automatic method in terms of recall and WSS (evaluation measures that are critical to the construction of an effective systematic review) is the Term/Recall/MeSH instantiation of the objective method. The highest performing automatic method in terms of precision is the Term/ F_3 instantiation of the objective method.

In almost all measures, the objective instantiations outperform the conceptual instantiations. Often, there are significant differences between the effectiveness of the objective and conceptual methods.

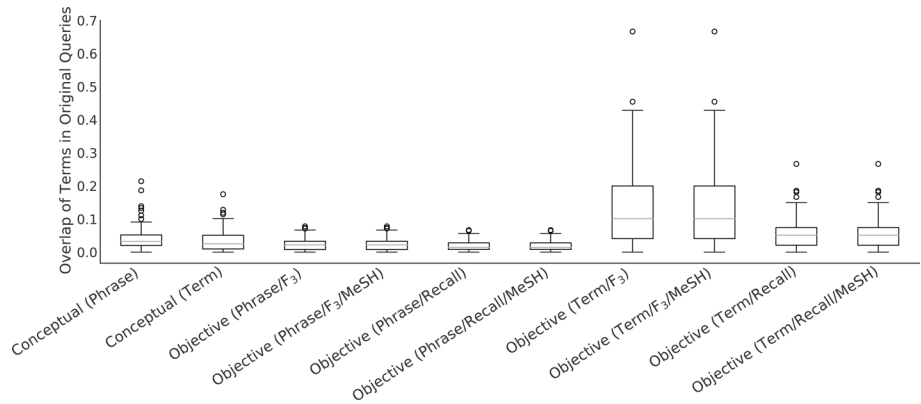


Fig. 5 Overlap between of keywords (e.g., terms, phrases, MeSH terms) between the originally, manually formulated queries and each instantiation of the automatic query formulation methods. The overlap is normalised by the total number of terms in an automatically formulated query. Terms from all iterations of each instantiation of an automatic query formulation method are combined to compute the overlap

Between both the conceptual and the objective methods, the term-based instantiations are almost always more effective than the phrase instantiation counterparts. This suggests that more effective queries for systematic review literature search are those that contain single terms instead of a mixture of terms and phrases.

The impact of MeSH terms on queries (at least within the objective instantiations) is clear also: the gains achieved in recall are typically higher in objective instantiations that add MeSH terms to those queries that do not. This suggests also that the use of MeSH terms is critical for maximising recall.

5.2 RQ2: factors contributing to effectiveness

This section aims to address RQ2: *what factors of automatically formulated queries contribute the most to effectiveness?*. To answer this, we study the correlations between factors that may contribute to the effectiveness of queries and the actual effectiveness of queries.

5.2.1 Factor 1: choice of keywords

The first factor that is investigated is the choice of keywords in queries. Specifically, how phrases and MeSH terms contribute to effectiveness, and whether the effectiveness of automatically formulated queries is due to the choice of keywords. Figure 4 illustrates how phrases and MeSH terms affect the originally manually formulated queries. Interestingly, while phrases caused a decrease in retrieval effectiveness for most automatically formulated queries (as illustrated in Table 1, term vs. phrase instantiations), the addition of phrases is correlated with retrieval effectiveness for the original queries. This is also true when MeSH terms are added to the automatically formulated queries (also presented in Table 1, MeSH vs. no MeSH instantiations), indeed in the original queries, there is a moderately strong correlation between the number of MeSH terms and effectiveness. These

findings suggest that more effective queries contain both phrases and MeSH terms, and that both phrases and MeSH terms are conducive of effectiveness. However, as the results in Table 1 illustrate, the identification and combination of these keywords are the most important factors in terms of effectiveness. Furthermore, while more phrases can have a positive impact on the effectiveness of a query, the choice of phrases is more important.

Figure 5 furthers the point that not only is the identification of correct terms important, but also the way in which those terms are combined in a Boolean expression. This figure illustrates the normalised overlap of terms between the original, manually formulated queries and each instantiation of an automatic query formulation method. Interestingly, most automatically formulated queries have less than 10% of terms in common with the original queries. The intuition that a high keyword overlap between automatically formulated queries and original queries results in high retrieval effectiveness does not hold when considering the objective Term/ F_3 and Term/Recall instantiations. The objective Term/Recall/MeSH instantiation achieved the highest recall and WSS of all automatic formulation instantiations (including conceptual methods), while the objective Term/ F_3 achieved the highest precision and F-measures. However, the objective Term/Recall instantiations have a lower term overlap than objective Term/ F_3 . This suggests that to obtain high recall, the choice of keywords is less important than the way keywords are combined in a Boolean expression, as although the objective Term/ F_3 instantiations have a higher overlap in terms than the objective Term/Recall instantiations (suggesting that they are more similar to the original queries), they obtain almost double the recall.

5.2.2 Factor 2: number of seed studies

The second factor that we investigate is the number of seed studies used in the automatic query formulation methods. This is because each topic has different numbers of seed studies for input. Note that this is reflective of reality as there is no set number of seed studies used in query formulation: it is possible that an information specialist is provided many, one or even no seed studies to develop a query. We perform this analysis with the intuition that the more seed studies that can be used, the more successful a query formulation method should be at producing an effective query. The correlations between the number of seed studies used for query formulation and query effectiveness is presented in Fig. 6.

First, looking at the conceptual methods: The number of seed studies is weakly negatively correlated with precision but more strongly correlated with recall, with respect to both the phrase and term instantiations. The conceptual term instantiation is indeed strongly correlated with recall.

Next, investigating the objective instantiations: The inverse is true for the phrase-based instantiations: more seed studies are more strongly correlated with precision and less correlated with recall. Indeed for many of the objective approaches, more seed studies does not necessarily correlate with any increase in recall. However, the term-based instantiations of the objective method which optimise for recall do see a moderate correlation in both precision *and* recall; indicating that for at least these instantiations, more seed studies do indeed correlate with more effective queries.

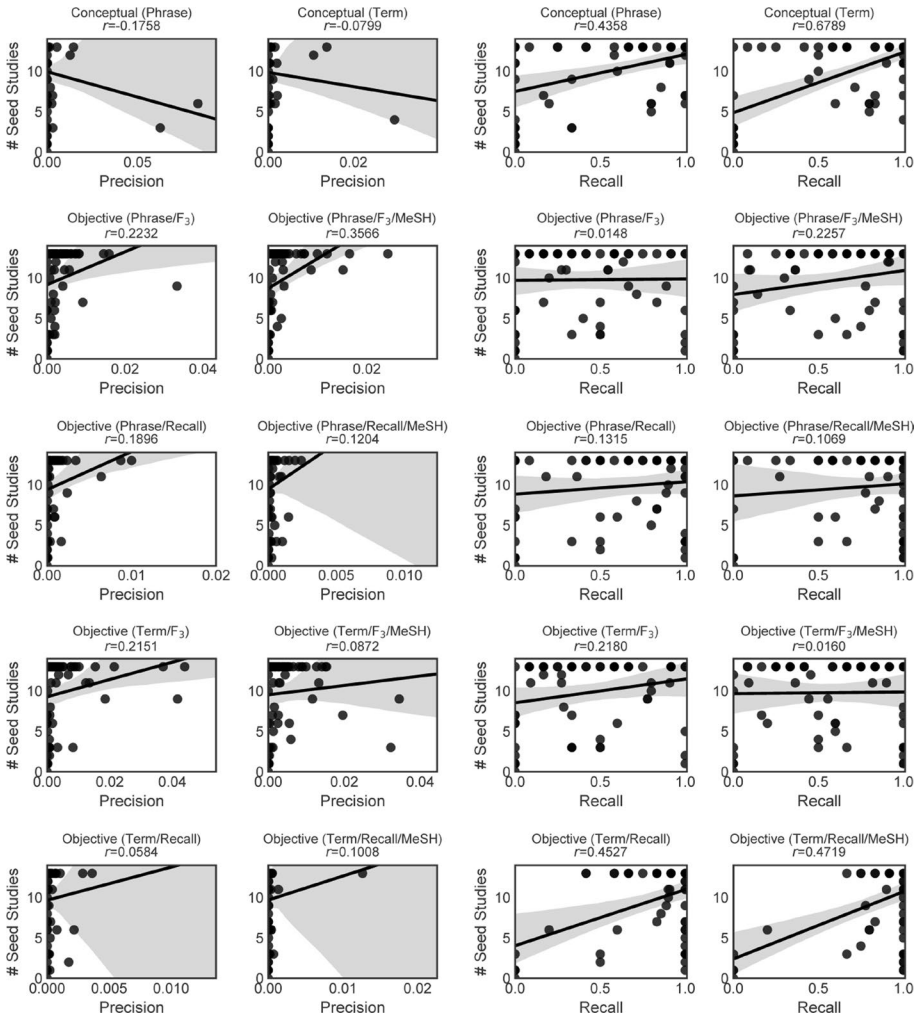


Fig. 6 Correlation between number of seed studies and effectiveness of queries. Each point in a plot refers to an iteration of the automatic query formulation approach given in the title (thus the x-axis is averaged across topics). Plots on the left correspond to precision, plots on the right refer to recall. Pearson's *r* is indicated between the two variables beneath the title of each plot

Table 2 Results of case study using manual query refinement after automatic query formulation

	Precision	F _{0.5}	F ₁	F ₃	Recall	WSS
Original	0.0263	0.0324	0.0494	0.1686	0.8869	0.8232
Conceptual (formulated)	0.0025	0.0031	0.0049	0.0220	0.6458	0.6177
Conceptual (refined)	0.0020	0.0025	0.0040	0.0188	0.9166	0.9159
Objective (formulated)	0.0001	0.0002	0.0003	0.0017	0.9687	0.9607
Objective (refined)	0.0009	0.0011	0.0018	0.0090	0.9375	0.9368

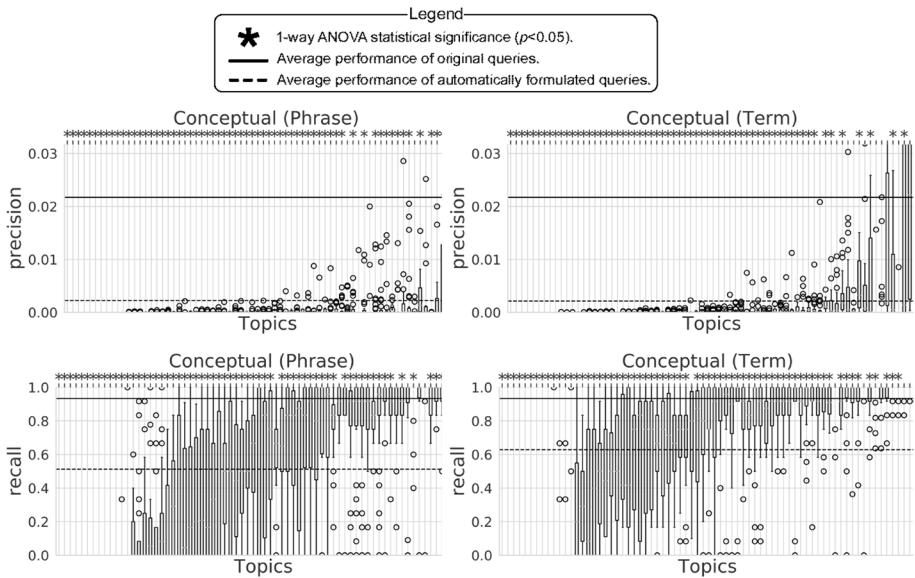


Fig. 7 Sensitivity of the two instantiations of the automatic conceptual query formulation method, measured using precision and recall. Left: phrase-based instantiation, right: term-based instantiation. Approximately one-third of topics obtain high recall with relatively low sensitivity (far right topics in bottom plots), while another third is highly sensitive to seed studies (middle topics in bottom plots). The last third of topics did not retrieve any relevant studies (far left topics in bottom plots). Meanwhile, for both instantiations, many topics obtained very low precision, except for a handful from the conceptual Term method; although these topics mostly vary in effectiveness

5.3 RQ3: effectiveness after manual refinement

This section aims to address RQ3: *Which automatic query formulation method provides the most effective starting point for manual refinement?* We address this question through a case study where we manually refine a small subset of the automatically formulated queries by removing terms from the query. We take a small subset of queries (approximately 10% of topics, 8 in total) from the highest performing iterations in terms of recall and manually apply query reduction using the validation set to validate the effectiveness of the queries. The results of this case study are presented in Table 2. We first compare the results of the automatically formulated queries and the same queries, but manually refined. The queries automatically formulated using the conceptual approach perform the worst (mirroring the results in Table 1). However, once refined, the recall of these queries outperforms the same original queries, with a marginal drop in precision. On the other hand, manually refining the objective queries resulted in a small drop in recall and a small increase in precision. This suggests that queries automatically formulated using the conceptual approach have the ability to be much more effective once refined, and that the objective approach formulates queries with a very high recall which is difficult to maintain while increasing precision when refining. This leads to an overarching result about query formulation in this domain: it may be easier to refine a query to increase recall when precision is high than it is to refine a query to maintain recall and increase precision when recall is high.

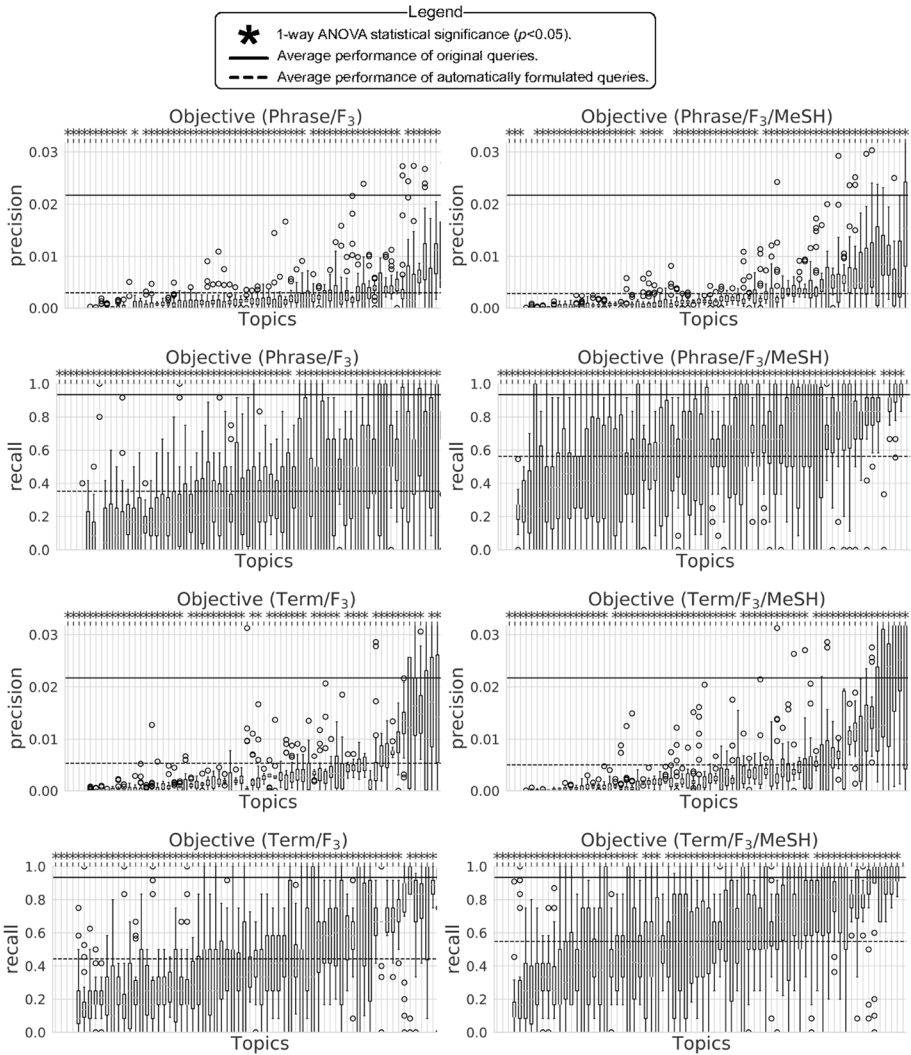


Fig. 8 Objective F_3 -optimised query formulation. Left: without MeSH terms, right: with MeSH terms. Overall, these instantiations have the highest variability in terms of both recall and precision. There is little difference between the phrase-based instantiations and the term-based instantiations

5.4 RQ4: sensitivity to seed studies

This section aims to address RQ4: *How sensitive to variation in the initial seed studies are the automatic query formulation approaches?* We address this question by analysing the per-topic performance of each instantiation of both of the fully automatic Boolean query formulation methods. This is shown in Figs. 7, 8, and 9. Each of these figures illustrates the per-topic breakdown of the effectiveness of each instantiation of an automatic Boolean query formulation method (given in the title) in terms of precision or recall (the y-axis). Boxes in each plot are ordered by the average performance of each topic to better show the

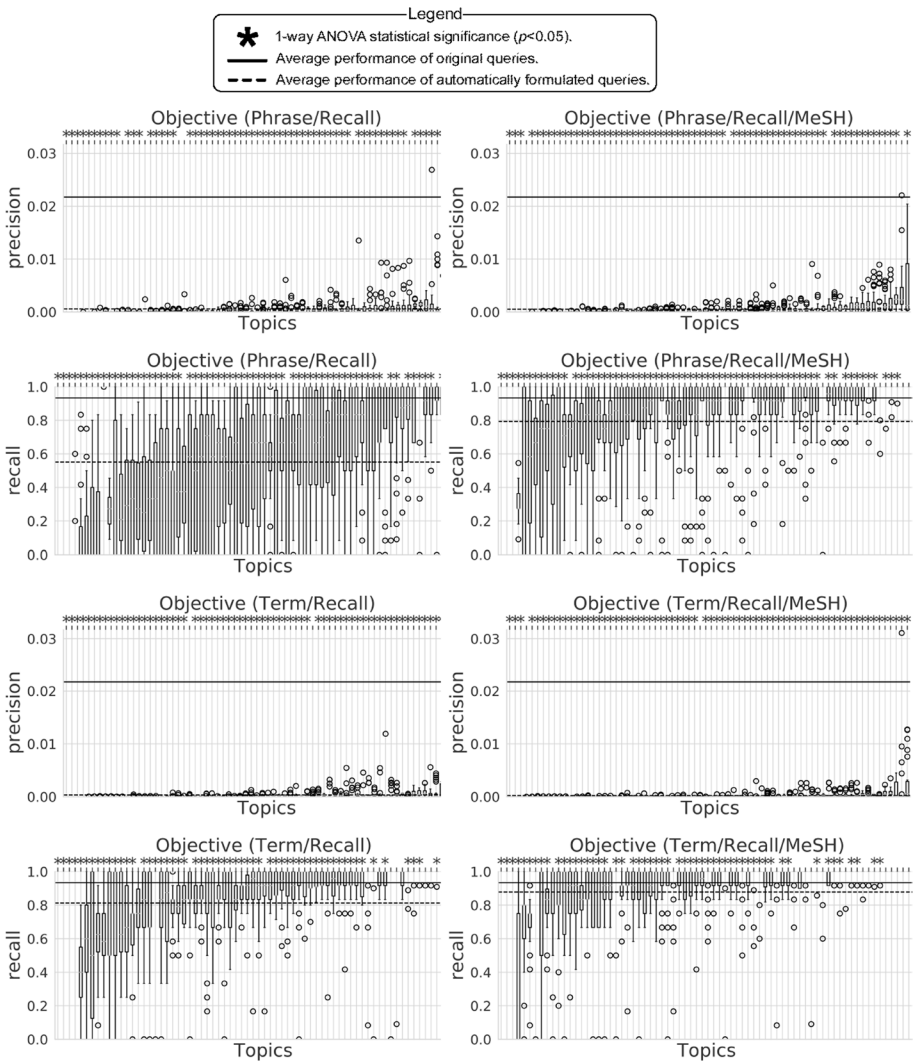


Fig. 9 Objective recall-optimised query formulation. Left: without MeSH terms, right: with MeSH terms. The term-based instantiations have the lowest variability in recall overall, while also achieving the highest recall overall. Meanwhile, the phrase-based instantiations are similar in variability to the F_3 -optimised instantiations

differences between instantiations of the query formulation methods. Note that this means that plots cannot be compared to each other using the x-axis.

First, examining the topic breakdown plots for the conceptual instantiations in Fig. 7, we note that while some topics are able to achieve reasonably high performance, a number of topics result in an overall poor average performance. The recall plots specifically tell an interesting story: a number of high-performing topics show low variability *and* closely match the performance of the original queries. Meanwhile, the majority of topics display

very high variability in effectiveness. The precision of topics in the conceptual instantiations is overall poor, and for most topics there is little variations in the low performance. However, approximately one quarter of topics from both instantiations are sensitive to seed studies, causing variation in retrieval effectiveness—suggesting that for these topics, the choice of seed studies does indeed have an. Approximately one quarter of topics in both instantiations retrieve no relevant studies across all iterations. For the automatic conceptual method, the choice of seed studies can have a considerable impact on the overall effectiveness of certain queries, especially recall.

Next, we report the variability for the objective instantiations which optimise for F_3 in Fig. 8. Overall, there is a high amount of variability in the effectiveness of queries for both precision and recall, for all instantiations. Between the phrase and the term instantiations, there is little difference in the amount of variability due to the seed studies. Indeed for the majority of topics, there is a high statistical difference between the variability of the original queries and the automatically formulated queries. Comparing these instantiations to the previous conceptual instantiations, on a per-topic basis, these instantiations are more likely to produce queries that retrieve more relevant studies: fewer topics overall retrieve zero studies. However, when averaged across each iteration, the conceptual instantiations achieve a higher recall as there is less variability. This may be due to the fact that these instantiations attempt to put some weight towards precision during the tuning process: overall these instantiations obtain the highest precision out of all other instantiations studied in this article.

Finally, we report the variability for the objective instantiations which optimise for recall in Fig. 9. The most immediate result is the relatively low variability in recall for the objective Term/Recall/MeSH instantiation. This instantiation is the most tolerant to sensitivity in seed studies among all methods. Comparing across all other instantiations, the queries automatically formulated for this topic have the lowest overall variability. The average recall across all topics and iterations for this particular instantiation is very close to the original queries. However, this must be balanced with the significant difference in precision between these queries and the original ones. Comparing these queries to the objective queries tuned for F_3 , there is a large difference in the variability between the phrase-based instantiations and the term-based instantiations. This is unlike the instantiations tuned for F_3 , where there is little difference in the variability of queries between phrase-based and term-based instantiations. For all of the instantiations in Fig. 9, the precision of almost all topics is low, as is expected when queries are tuned for recall. Unlike the conceptual instantiations or the objective instantiations tuned for F_3 , the variability in precision among these topics is low for almost all topics.

6 Discussion

6.1 RQ1: comparison to original queries

The first research question, *how does automatic query formulation compare to manual query formulation in terms of search effectiveness?* guided the investigation into comparing Boolean queries derived from two automatic query formulation methods to original, manually formulated Boolean queries. We found that the *automatic query formulation methods* investigated in this work, are *only somewhat effective compared to the original, manually formulated queries*. This demonstrates the utility of information specialists in

applying their expertise to query formulation. That being said, we suspect that the intellectual burden involved in query formulation can be massively reduced through the use of automatic query formulation. It is also worth noting that the original, manually formulated queries have undergone the scrutiny of colleagues and peer review: these steps are likely to greatly improve the quality and effectiveness of queries making the comparison performed in this study more stringent (as we do not have access to the queries prior to these quality control steps).

6.2 RQ2: factors contributing to effectiveness

The second research question, *what factors of automatically formulated queries contribute the most to effectiveness?* guided the investigation into two specific factors that were seen to likely contribute the most to effectiveness of queries. We performed an extensive analysis to determine if these factors of the automatically formulated queries lead to effective queries. Firstly, we found that in order to have a high recall, it is not necessary to have the same terms as the original queries. In fact, we found that a high term overlap with the original query instead lead to high precision. Next, we identified that queries with MeSH terms are more conducive of higher recall, and that the more MeSH terms in a query, the more effective that query is. For all instantiations of the conceptual and objective methods, we also found that the use of phrases reduces recall while increasing precision. Although higher numbers of keywords in the original queries strongly correlated with higher effectiveness, suggesting that the use of many phrases, MeSH terms, and terms could be beneficial, the key finding for us was that *the choice of keywords is more important than the number of keywords*. We also found that generally, *the more seed studies that were used for automatic query formulation, the more effective queries were*, in both precision and recall. However, there is still work to be done to reduce the variability of query formulation. As the conceptual and objective methods are deterministic, the only variable introduced to each method is the set of seed studies used to start the query formulation process for each method.

6.3 RQ3: effectiveness after manual refinement

The third research question, *which automatic query formulation method provides the most effective starting point for manual refinement?* guided a case study that involved the manual refinement of automatically formulated queries. We found that when some manual effort is expended to refine the automatically formulated queries, i.e., through query reduction, *the queries can become as effective as the original queries*. Specifically, we found that the automatic conceptual approach should be chosen when recall is the preferred measure to optimise a search for, and the automatic objective approach should be chosen when precision is the preferred measure to optimise a search for. The automatically formulated, manually refined queries obtain a higher recall than the manual, original queries. However, the precision of the manually refined queries is still lower than original queries. Furthermore, it was found that the conceptual queries obtained a much higher recall once manually refined, while approximately maintaining their precision. Note that the refinement was performed by an author of the paper and not an experienced information specialist. It is likely that if an experienced information specialist were to refine the automatically formulated

queries, then both precision and recall could be increased, in line with the effectiveness of the original queries.

6.4 RQ4: sensitivity to seed studies

The fourth research question, *how sensitive to variation in the initial seed studies are the automatic query formulation approaches?* guided the investigation into the sensitivity of the automatic query formulation methods in terms of retrieval effectiveness. Almost all of the automatic Boolean query formulation methods investigated in this work were *highly sensitive to the initial seed studies*. While we cannot know for certain if manually formulated queries using the conceptual or objective approaches are as sensitive to seed studies, as this would require humans to develop searches, the effectiveness of manually formulated queries is almost always higher than automatic approaches. While the conceptual query formulation methods offer a more consistent base for manual query refinement than the objective F_3 instantiations, the most consistent and least sensitive to seed studies was the objective (Term/Recall/MeSH) instantiation. However, as the results of the manual refinement show, these queries are more difficult to refine (to increase precision while maintaining recall) than the conceptual (Term) instantiation which was easier to refine (to increase recall while maintaining precision). To truly determine the most effective base for manual query refinement, a large scale user study must be undertaken. We leave this for future research.

7 Conclusions

This article presented extensions to two existing automatic Boolean query formulation methods. Instantiations of these methods were compared with each other between formulation methods and within instantiations of a method. Automatic instantiations of the objective and conceptual methods were also compared to queries formulated manually for the same topics. An analysis to determine which factors produced more effective queries was undertaken as well as an analysis on how sensitive the automatic query formulation instantiations are to seed studies. We also performed a small case study to determine which instantiation of the highest performing formulation method provides the best starting point for manual query refinement and how the sensitivity to seed studies may affect this. Our main findings are that while the automatic Boolean query formulation instantiations of the objective and conceptual methods on their own cannot beat the performance of the original queries, with some manual refinements (in this case query reduction), they can be more effective. The conceptual computational adaptations should be used for this purpose as they achieved the highest precision once refined, and a comparable recall to the original queries. If no manual query refinements are desired the objective adaptations are a more suitable choice; however, the trade-off between precision and recall depends on the evaluation measure optimised. We also found that both automatic methods are sensitive to seed studies, and that the instantiations of these methods that are term-based are generally less sensitive to variation and more effective. Instantiations that use MeSH terms generally have a higher recall with a trade-off in precision, and instantiations that use phrases generally have a higher precision with a trade-off in recall.

The results of this article impact both new techniques for automatic Boolean query formulation for systematic review literature search, as well as manual approaches. Our empirical

findings confirm the intuitions that queries should prefer terms to increase recall and carefully chosen phrases to increase precision. The choice of seed studies can have a significant impact on the resulting query, and these should be chosen carefully to ensure maximum coverage of relevant studies.

In our future work, we plan to undertake a user study to measure how sensitive manually formulated queries are to seed studies (rather than approximating as in this work) and to investigate query reduction methods to automatically refine queries to further improve the performance of the automatic objective and conceptual methods. Another aspect of query formulation which was not investigated in this work is the quality of seed studies. One possible direction of research is to develop evaluation criteria to predict the effectiveness of a resulting query given a set of seed studies.

The end goal of this line of research is to integrate it into tools for information specialists to use to reduce the cognitive burden of query formulation, to provide a less subjective basis for query formulation, and to ultimately improve the systematic review creation process by reducing the total number of studies to screen for inclusion in the systematic review.

Acknowledgements Harrisen Scells is the recipient of a CSIRO Ph.D. Top Up Scholarship. Dr. Guido Zucco is the recipient of an Australian Research Council DECRA Research Fellowship (DE180101579).

References

- Androustopoulos, I., Ritchie, G. D., & Thanisch, P. (1995). Natural language interfaces to databases—An introduction. *Natural Language Engineering*, 1(1), 29–81.
- Aronson, A. R. (2001). Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In: *Proceedings of the AMIA symposium* (p. 17). American Medical Informatics Association.
- Bullers, K., Howard, A. M., Hanson, A., Kearns, W. D., Orriola, J. J., Polo, R. L., et al. (2018). It takes longer than you think: Librarian time spent on systematic review tasks. *Journal of the Medical Library Association: JMLA*, 106(2), 198.
- Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (2019). *Cochrane handbook for systematic reviews of interventions*. Chichester: Wiley.
- Clark, J. (2013). Systematic reviewing. In G. M. W. Suhail & A. R. Doi (Eds.), *Methods of clinical epidemiology*. Berlin: Springer.
- Clark, J., Glasziou, P., Del Mar, C., Bannach-Brown, A., Stehlik, P., & Scott, A. M. (2020). A full systematic review was completed in 2 weeks using automation tools: A case study. *Journal of Clinical Epidemiology*, 121, 81–90.
- Cohen, A., Hersh, W., Peterson, K., & Yen, P. (2006). Reducing workload in systematic review preparation using automated citation classification. *JAMA*, 13(2), 206–219.
- Golder, S., Loke, Y., & McIntosh, H. M. (2008). Poor reporting and inadequate searches were apparent in systematic reviews of adverse effects. *Journal of Clinical Epidemiology*, 61(5), 440–448.
- Hausner, E., Guddat, C., Hermanns, T., Lampert, U., & Waffenschmidt, S. (2015). Development of search strategies for systematic reviews: Validation showed the noninferiority of the objective approach. *Journal of Clinical Epidemiology*, 68(2), 191–199.
- Hausner, E., Waffenschmidt, S., Kaiser, T., & Simon, M. (2012). Routine development of objectively derived search strategies. *Systematic Reviews*, 1(1), 19.
- Jimmy, Zucco, G., & Koopman, B. (2018). Choices in knowledge-base retrieval for consumer health search. In G. Pasi, B. Piwowarski, L. Azzopardi, & A. Hanbury (Eds.) *Advances in information retrieval* (pp. 72–85). Cham: Springer.
- Kanoulas, E., Li, D., Azzopardi, L., & Spijker, R. (2017). CLEF 2017 technologically assisted reviews in empirical medicine overview. In: *CLEF'17*.
- Kanoulas, E., Spijker, R., Li, D., & Azzopardi, L. (2018). Clef 2018 technology assisted reviews in empirical medicine overview. In: *CLEF 2018 evaluation labs and workshop: Online working notes, CEUR-WS*.
- Kim, Y., Seo, J., & Croft, W. B. (2011). Automatic Boolean query suggestion for professional search. In: *SIGIR'11*.

- Lavis, J., Davies, H., Oxman, A., Denis, J. L., Golden-Biddle, K., & Ferlie, E. (2005). Towards systematic reviews that inform health care management and policy-making. *JHSRP*, 10(1–suppl), 35–48.
- Lee, G. E., & Sun, A. (2018). Seed-driven document ranking for systematic reviews in evidence-based medicine. In: *Proceedings of the 41st annual international ACM SIGIR conference on research & development in information retrieval, SIGIR '18* (pp. 455–464).
- Leeflang, M. M., Deeks, J. J., Takwoingi, Y., & Macaskill, P. (2013). Cochrane diagnostic test accuracy reviews. *Systematic Reviews*, 2(1), 82.
- Lefebvre, C., Manheimer, E., & Glanville, J. (2008). Searching for studies. *Cochrane handbook for systematic reviews of interventions: Cochrane book series* (pp. 95–150).
- McGowan, J., & Sampson, M. (2005). Systematic reviews need systematic searchers (IRP). *Journal of the Medical Library Association*, 93(1), 74.
- Mirhosseini, S., Zuccon, G., Koopman, B., Nguyen, A., & Lawley, M. (2014). Medical free-text to concept mapping as an information retrieval problem. In: *Proceedings of the 2014 Australasian document computing symposium* (pp. 93–96).
- Miwa, M., Thomas, J., O'Mara-Eves, A., & Ananiadou, S. (2014). Reducing systematic review workload through certainty-based screening. *JBI*, 51, 242–253.
- Olorisade, B. K., de Quincey, E., Brereton, P., & Andras, P. (2016). A critical analysis of studies that address the use of text mining for citation screening in systematic reviews. In: *Proceedings of the 20th international conference on evaluation and assessment in software engineering* (p. 14).
- O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: A systematic review of current approaches. *Systematic Reviews*, 4(1), 5.
- Pazos R, R. A., González B, J. J., Aguirre L, M. A., Martínez F, J. A., & Fraire H, H. J. (2013). Natural language interfaces to databases: An analysis of the state of the art. In: *Recent advances on hybrid intelligent systems* (pp. 463–480).
- Popescu, A. M., Armanasu, A., Etzioni, O., Ko, D., & Yates, A. (2004). Modern natural language interfaces to databases: Composing statistical parsing with semantic tractability. In: *Proceedings of the 20th international conference on computational linguistics*.
- Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text Mining: Applications and Theory*, 1, 1–20. <https://doi.org/10.1002/9780470689646.ch1>.
- Sayers, E. (2010). *A general introduction to the e-utilities. Entrez programming utilities help [internet]*. Bethesda: National Center for Biotechnology Information.
- Scells, H., & Zuccon, G. (2018a). Generating better queries for systematic reviews. In: *Proceedings of the 41st annual international ACM SIGIR conference on research & development in information retrieval, SIGIR '18*.
- Scells, H., & Zuccon, G. (2018b). Searchrefiner: A query visualisation and understanding tool for systematic reviews. In: *Proceedings of the 27th ACM international conference on information and knowledge management* (pp. 1939–1942). ACM.
- Scells, H., Zuccon, G., & Koopman, B. (2019). Automatic Boolean query refinement for systematic review literature search. In: *The world wide web conference* (pp. 1646–1656).
- Scells, H., Zuccon, G., & Koopman, B. (2020a). You can teach an old dog new tricks: Rank fusion applied to coordination level matching for ranking in systematic reviews. In: *European conference on information retrieval* (pp. 399–414). Springer.
- Scells, H., Zuccon, G., Koopman, B., & Clark, J. (2020b). Automatic Boolean query formulation for systematic review literature search. *Proceedings of the Web Conference, 2020*, 1071–1081.
- Scells, H., Zuccon, G., Koopman, B., & Clark, J. (2020c). A computational approach for objectively derived systematic review search strategies. In: *European conference on information retrieval* (pp. 385–398). Springer.
- Shemilt, I., Simon, A., Hollands, G., Marteau, T., Ogilvie, D., O'Mara-Eves, A., et al. (2014). Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. *RSM*, 5(1), 31–49.
- Soldaini, L., & Goharian, N. (2016). Quickmuls: A fast, unsupervised approach for medical concept extraction. In: *MedIR workshop. SIGIR*.
- Tsafnat, G., Glasziou, P., Choong, M. K., Dunn, A., Galgani, F., & Coiera, E. (2014). Systematic review automation technologies. *Systematic Reviews*, 3(1), 74.
- van der Vegt, A. H., Zuccon, G., & Koopman, B. (2019). Learning inter-sentence, disorder-centric, biomedical relationships from medical literature. In: *AMIA fall symposium*.
- Wallace, B., Trikalinos, T., Lau, J., Brodley, C., & Schmid, C. (2010). Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics*, 11(1), 55.

Zhong, V., Xiong, C., & Socher, R. (2017). Seq2sql: Generating structured queries from natural language using reinforcement learning. ArXiv preprint, [arXiv:1709.00103](https://arxiv.org/abs/1709.00103).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.