

# A Computational Approach for Objectively Derived Systematic Review Search Strategies

Harrisen Scells<sup>1</sup>[0000-0001-9578-7157], Guido Zuccon<sup>1</sup>[0000-0003-0271-5563],  
Bevan Koopman<sup>2</sup>[0000-0001-5577-3391], and Justin Clark<sup>3</sup>[0000-0003-0133-1613]

<sup>1</sup> The University of Queensland, St Lucia, Australia {h.scells@uq.net.au}

<sup>2</sup> CSIRO, Brisbane, Australia

<sup>3</sup> Institute for Evidence-Based Healthcare, Bond University, Gold Coast, Australia

**Abstract.** Searching literature for a systematic review begins with a manually constructed search strategy by an expert information specialist. The typical process of constructing search strategies is often undocumented, ad-hoc, and subject to individual expertise, which may introduce bias in the systematic review. A new method for *objectively* deriving search strategies has arisen from information specialists attempting to address these shortcomings. However, this proposed method still presents a number of manual, ad-hoc interventions, and trial-and-error processes, potentially still introducing bias into systematic reviews. Moreover, this method has not been rigorously evaluated on a large set of systematic review cases, thus its generalisability is unknown. In this work, we present a computational adaptation of this proposed objective method. Our adaptation removes the human-in-the-loop processes involved in the initial steps of creating a search strategy for a systematic review; reducing bias due to human factors and increasing the objectivity of the originally proposed method. Our proposed computational adaptation further enables a formal and rigorous evaluation over a large set of systematic reviews. We find that our computational adaptation of the original objective method provides an effective starting point for information specialists to continue refining. We also identify a number of avenues for extending and improving our adaptation to further promote supporting information specialists.

**Keywords:** Systematic Reviews · Boolean Queries · Query Formulation

## 1 Introduction

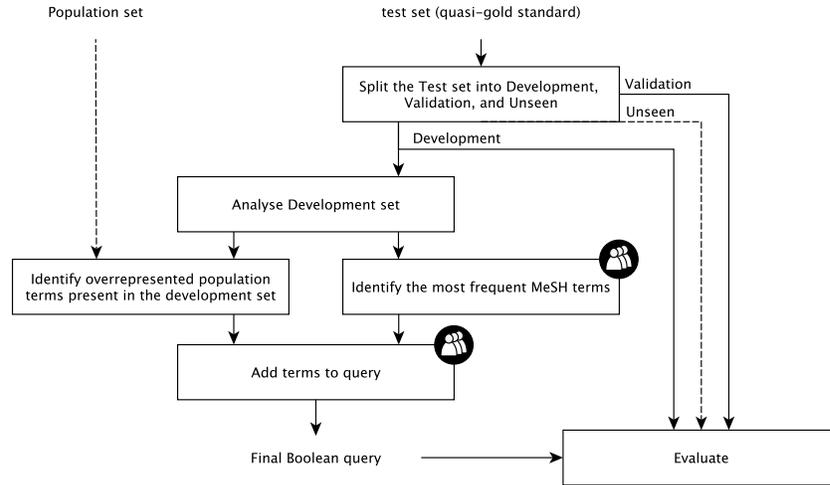
The goal of a systematic review is to synthesise *all* relevant literature for a highly focused research question. Systematic reviews are used extensively in evidence based medicine (this is the domain we consider in the rest of the paper), both for healthcare decision making and institutional policy mandates concerning health topics. While systematic reviews strive to be methodical and comprehensive, there are still a number of processes associated with them which introduce bias and subjectivity. Arguably, the process which contributes the most bias is the construction of *search strategies*. The main element of a search strategy is a complex Boolean query. This is issued to one or more publication databases

(e.g., PubMed, EMBASE). Retrieved studies are first screened (i.e., the title and abstracts are assessed) for potential inclusion in the review. Then, the full-text of screened studies deemed potentially relevant to the review are assessed to determine if they should be synthesised in the final review [5].

The most common method for developing a search strategy is the *conceptual* method [2,12] (although other methods have been investigated that do not produce a Boolean query [7,6]). Here, the query is formulated by dividing the research question of a systematic review into multiple high-level concepts, and then choosing suitable synonyms for each concept. Query formulation is typically performed by trained information specialists (e.g., librarians), who use domain expertise and intuition to decide, for example, what keywords to add to a query and where they should be added, what kind of field restrictions should be applied, and when to stop formulating. Often, information specialists also have access to a handful of studies (seeds) that the researchers are certain will be included in the synthesis of the review. In the conceptual approach, information specialists use the few seed studies to repeatedly gauge the effectiveness of the queries they formulate in an ad-hoc manner.

An *objective* [20,4] method for deriving systematic review search strategies has recently been proposed which aims to avoid the unrigorous, subjective aspects of the conceptual approach. In this method, a small set of ‘gold standard’ studies are first identified — these serve to (semi-)automatically identify keywords to add to the query, and to validate its effectiveness. The gold standard set is akin to the seed studies considered in the conceptual approach, but generally much larger (conceptual: a handful; objective: in the order of 10s-100s). Despite the name, this method is still ad-hoc and involves manual trial-and-error with respect to choosing a subset of the identified keywords to add to the query, and where to place keywords in the query. In addition, this method has only been evaluated on a handful of use-case systematic reviews, thus its effectiveness and generalisability beyond these cases is as yet unknown.

In this paper, we propose a computational adaptation of the objective methodology proposed by Hausner et al.[4] for objectively deriving medical systematic review search strategies. Our approach does not require manual human involvement, nor trial-and-error procedures, and, in fact, is capable of generating a query automatically, given a set of relevant studies as input. Furthermore, we evaluate this method on a large set of 40 systematic reviews from a collection used for the evaluation of automation methods in this context [6] and further replicate a small study by Hausner et al. [4]. We also consider the cost factors of systematic review development in our evaluation. The primary goal of this research is to develop a more transparent and less subjective method to search strategy development by computationally adapting and extending the current objective approach. Achieving total recall of the relevant literature for a study is important. However, the effectiveness of systematic reviews is often hampered by the fact that they are resource-intensive and often become out-of-date at the time of publication [21]: it takes on average 2 years and AUD\$350K to create a systematic review [13,9] and currently only 36% of Cochrane SRs are deemed



**Fig. 1:** The process used for deriving queries using the objective method. The dashed line signifies an extension of the objective method not in the originally proposed method. The  symbol refers to the processes in the objective method which currently require manual intervention. Automating these manual processes is the main focus of this work.

up-to-date. The largest time and monetary cost involved in systematic review creation is the *screening* phase, which is directly influenced by the number of studies retrieved by the search strategy. Furthermore, the exponentially growing body of published research casts doubts on how effective the reviews are in identifying all relevant research; potentially introducing bias.

With the method presented in this research, our overarching goal is to automatically and objectively derive suitable queries which can be used as a starting point for query formulation, to derive more effective search strategies (higher precision while maintaining recall) than manually derived queries. The method of this study is expected not to replace information specialists, but to provide support by reducing bias and subjectivity in the search development process.

## 2 Computational, Objective Method

Our method for automatically and objectively deriving search strategies is an adaptation of the method originally proposed by Hausner et al. [4]. A high-level process overview of the objective method is shown in Figure 1. This figure highlights manual aspects of the original method which we computationalise, and extensions to the original method our method makes, which seek to further reduce bias. The following two sections first describe the original method, and then the specific computational adaptations and extensions we make.

### 2.1 The Objective Method

The objective method [4] starts by identifying studies either by scanning the references of similar, already published systematic reviews, or by issuing broad

queries to medical databases (e.g., PubMed, MEDLINE) and screening a subset for identifying gold-standard references. These references form the ‘test’ set, which is divided into a development set ( $\frac{2}{3}$ ) and a validation set ( $\frac{1}{3}$ ). The title and abstract of each of the references in the test set is then analysed by ranking each term by the frequency it appears in each of the references (i.e., document frequency – DF). Next, these terms are filtered to include the top 20% of terms according to DF. At the same time, a population set (i.e., a background collection) of 10,000 studies is identified by issuing an empty search to PubMed and restricting results to the last 12 months (the default ranking of PubMed is by descending publication date). The previously filtered terms are filtered yet again to include the bottom 2% of terms according to DF. Finally, the 20 most frequent MeSH terms are identified from studies in the development set. A Boolean query is then developed by dividing terms into three categories (which form three clauses, linked with the Boolean AND operator): (1) terms relating to health conditions, (2) terms relating to a treatment, and (3) terms relating to the types of study design to be included. Through a time consuming process of trial-and-error, filtered terms and MeSH terms are then added to one of the three clauses of the query depending on the category of the term. Terms inside a category are combined using an OR operator, and the three categories are then combined with an AND operator. The effectiveness of the query is then compared to the validation set.

## 2.2 Automating the Objective Method

We propose a number of computational modifications to this method which seek to further remove human subjectivity from the process. We also improve the process by which evaluation of the resulting queries is undertaken in a number of ways. In our modified methodology, we begin with the same test set, however we split into ( $\frac{2}{4}$ ) development, ( $\frac{1}{4}$ ) validation, and ( $\frac{1}{4}$ ) unseen. The unseen set is used to approximate how the query will perform on studies which are not assessed (i.e., a study which is relevant, but which may never be retrieved by the query, and therefore never screened for potential relevance). It also allows us to develop the query on the development set, tune parameters values on the validation set, and study their effectiveness on the unseen set. We then follow the same method of filtering terms using the development set and the population set, as well as identifying MeSH terms to use. To automatically assign a category for a term, the semantic type of a term is used. The semantic type is obtained automatically by mapping terms to UMLS concepts via MetaMap [1] (version 2018 with options set to their default values and using UMLS2018AA). If a term does not map to a concept in MetaMap, it is discarded. Once a semantic type is obtained for a term, it is automatically categorised in a two fold process: (i) if the semantic type is present in Table 1a, then the term is mapped accordingly [22], (ii) if the semantic type is not present, then the semantic group of the semantic type is used to broadly categorise the term according to Table 1b. Note that in step (ii), some terms may be discarded due to the semantic group they belong to, denoted by ‘None’ in the table. Following this process, the identified MeSH terms are then added to one or more of the three categories according to the top-

Uzuner et al. [22] Relationship    Hausner et al. [4] Category

Test → Treatment  
 Treatment → Treatment  
 Diagnosis → Condition

(a) How a relationship as identified by Uzuner et al. [22] maps to a category.

Semantic Group	Hausner et al. [4] Category	MeSH top-level heading	Hausner et al. [4] Category
ACTI	→ Treatment	Anatomy	→ Condition
ANAT	→ Condition	Organisms	→ Condition
CHEM	→ Treatment	Diseases	→ Condition
CONC	→ None	Chemicals and Drugs	→ Treatment
DEVI	→ Treatment	Analytical, Diagnostic and Therapeutic Techniques, and Equipment	→ Treatment
DISO	→ Condition	Psychiatry and Psychology	→ Condition
GENE	→ Condition	Phenomena and Processes	→ Condition
GEOG	→ Study Type	Disciplines and Occupations	→ Condition
LIVB	→ Condition	Anthropology, Education, Sociology, and Social Phenomena	→ None
OBJC	→ Treatment	Technology, Industry, and Agriculture	→ None
OCCU	→ Condition	Humanities	→ None
ORGA	→ Study Type	Information Science	→ Study Type
PHEN	→ Condition	Named Groups	→ Condition
PHYS	→ Condition	Health Care	→ None
PROC	→ Treatment	Publication Characteristics	→ Study Type
		Geographicals	→ Study Type

(b) How a semantic group maps to a category.

(c) How a top-level MeSH heading maps to a category.

**Table 1:** Processes for mapping terms to a Hausner et al. [4] category in a query.

level MeSH parent in Table 1c. Once all of the identified terms are categorised, the computational assembly of the Boolean query takes place.

We also take a computational approach to the assembly of the Boolean query. A naïve approach could involve trying all combinations of terms in a category with all combinations of all other terms in all other categories (similar to the manual trial-and-error employed in the original method). The complexity of this approach, however, presents itself as infeasible:  $O(n!^3)$  (where  $n$  is the number of terms for a given category, assuming all categories have the same number of terms, in the worst case). Instead, we compute the maximum number of studies in the development set retrievable using the filtered terms and MeSH terms by first representing the set of relevant studies retrieved for each category as a binary array (e.g., health conditions  $\mathbf{c}=[1,1,1,1,1,1,1,1]$ , treatments  $\mathbf{t}=[0,1,1,1,1,1,1,1]$  and study type  $\mathbf{s}=[1,1,0,1,1,1,1,1]$ ), where 1 indicates that the relevant study referred by that position in the array is retrieved. Then we perform conjunction (bitwise AND) on the three binary arrays to obtain a new binary array (i.e.,  $\mathbf{c} \wedge \mathbf{t} \wedge \mathbf{s} = \mathbf{q}=[0,1,0,1,1,1,1,1]$ ) which represents the set of relevant studies in the development set that can be retrieved by the query that includes all terms (i.e. the maximal query). Note that when there are no terms present for a category<sup>4</sup> then the category is removed from the conjunction which forms  $\mathbf{q}$ . The logical conjunction of the three vectors has the same effect as executing the query, thus

<sup>4</sup> e.g., there are no terms that can be categorised into Study Type ( $\mathbf{s}$ ), but there are terms categorised into Conditions ( $\mathbf{c}$ ) and Treatments ( $\mathbf{t}$ ).

greatly increasing the number of comparisons that can be made (i.e, it reduces computation time). Further note that it is not guaranteed that the set of categories which contain terms from the development set, when combined using a Boolean AND operator, will retrieve all the relevant studies in the validation set — this is true regardless of using our technique for speeding up query assembly, or trying all combinations. Next, in an iterative manner, each term from each category is temporarily removed and a new binary vector ( $\mathbf{v}_i$ ) is computed, containing the set of relevant studies in the development set retrieved without that term. If  $\mathbf{q} \wedge \mathbf{v}_i = \mathbf{q}$ , that is, if the removal of the term has no effect on the number of relevant studies in the development set retrieved by the rest of the terms, then the removal of that term from the category is made permanent. In other words, that term contributes nothing overall to the query (or its contribution is redundant as its contributing studies are also retrieved by other query terms) and is removed from its respective clause. Note that this technique could also be used in an interactive system to highlight to a user those terms that do not contribute to the set of retrieved documents, or alternatively for evaluating existing search strategies. The iteration proceeds by considering one candidate term for removal at a time; terms are ordered descending by the sum of the components of their document vectors, i.e., their total document frequency, thus the order of terms removed is deterministic. The complexity of this approach is  $O(3n)$ : each term in the query is required to be only tested once for inclusion in the final query, rather than for all possible combinations. The resulting query is guaranteed to retrieve the maximum number of relevant studies possible in the development set (based on the terms which have been identified in the previous process).

We further propose to tune the term cut-off thresholds parameters for the filtering steps. Rather than fixing the thresholds at 20% for development and 2% for population (as done by Hausner et al.[4]), we perform a grid search (independently for each query) over combinations of thresholds to find the parameters best suited (optimising for  $F_1$ ,  $F_3$ , recall) for a particular query. We also apply the same strategy to identify the number of MeSH terms to add to a query. The development set is used to identify terms; then we evaluate queries on the validation set to select the best combination of parameters. The query can then be evaluated fairly on the unseen set.

### 3 Empirical Evaluation

We evaluate the computational method for objectively deriving systematic review search strategies on the CLEF 2018 Technology Assisted Reviews (TAR) collection [6]. This collection of diagnostic test accuracy systematic review protocols contains 75 topics (i.e., systematic reviews use-cases).<sup>5</sup> Diagnostic test accuracy reviews are highly specific and are considered one of the most difficult

<sup>5</sup> The CLEF 2018 TAR collection is a superset of queries from the 2017 TAR collection. The CLEF 2017 TAR collection was not used as the overlap of queries in the 2017 and 2018 collections for our purposes was the same, once we removed topics that had less than 50 relevant studies.

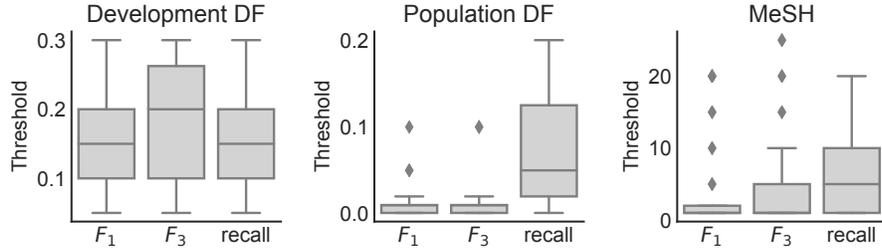
	$F_{0.5}$	$F_1$	$F_3$	NNR	Precision	Recall	
Original	0.0078 <sup>†</sup>	0.0123 <sup>†</sup>	0.0558 <sup>†</sup>	1040.03	0.0062 <sup>†</sup>	0.9384	
Original (Relaxed)	0.0015*	0.0024*	0.0115*	230824.58	0.0012*	0.9078	
Automated objective tuned for:	$F_1$ +MeSH	0.0056 <sup>†</sup>	0.0086 <sup>†</sup>	0.0340 <sup>†</sup>	<b>895.38</b>	0.0046 <sup>†</sup>	0.5329 <sup>†</sup>
	$F_1$	0.0148 <sup>†</sup>	0.0194 <sup>†</sup>	0.0442 <sup>†</sup>	2186.58	0.0129 <sup>†</sup>	0.2418 <sup>†</sup>
	$F_3$ +MeSH	0.0060 <sup>†</sup>	0.0094 <sup>†</sup>	0.0384 <sup>†</sup>	921.29	0.0049 <sup>†</sup>	0.5095 <sup>†</sup>
	$F_3$	<b>0.0166<sup>†</sup></b>	<b>0.0219<sup>†</sup></b>	<b>0.0510<sup>†</sup></b>	1217.09	<b>0.0146<sup>†</sup></b>	0.2672 <sup>†</sup>
	Recall+MeSH	0.0005 <sup>*†</sup>	0.0007 <sup>*†</sup>	0.0035 <sup>*†</sup>	84809.31*	0.0004 <sup>*†</sup>	<b>0.9523</b>
	Recall	0.0002 <sup>*†</sup>	0.0004 <sup>*†</sup>	0.0017 <sup>*†</sup>	102020.38*	0.0002 <sup>*†</sup>	0.8561*

**Table 2:** Evaluation results on unseen documents, with and without MeSH terms applied to queries. Relaxed indicates original queries where MeSH exploration is removed and phrases converted to Boolean OR clauses. Significant differences (paired two-tailed t-test  $p < 0.05$ ) between original queries indicated by \*. Significant differences between relaxed queries indicated by †. Highest values are **bolded**. Original queries do not achieve 100% recall because (i) errors in reporting of queries [3,14], and (ii) all queries are issued to PubMed even if the original query was reported as a MEDLINE query (i.e., translated automatically [17]).

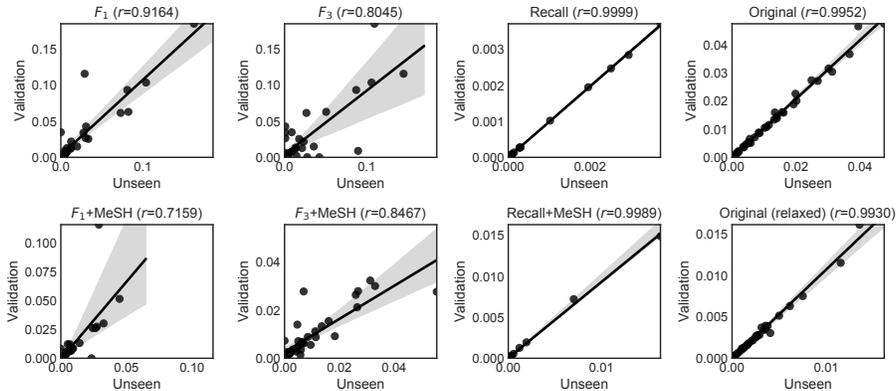
types of systematic reviews to search literature for [11]. Each topic comprises the title of the review, the Boolean query used to retrieve studies, and relevance assessments for the studies retrieved by the query. To determine the effectiveness of queries, we execute them in PubMed through the entrez API [16]. In our experiments, the test set for each topic is derived from the studies labelled relevant at an abstract level: these are studies that were retrieved by the Boolean query of the original systematic review and were screened for inclusion. We set the minimum size of the development set to 25, therefore excluding topics from the collection where the number of studies labelled relevant was less than 50 (development = 2/4 of total size). This number was chosen as the size of the development set in the study by Hausner et al. [4] was 25 (for a single topic). For comparison, the development set in a study by Simon et al. [20] was 78 (single topic). After removing topics in this way, 40 topics remained (still considerably larger than the previous studies), and the average number of relevant studies per topic was  $180.65 \pm 157.8$  (min: 52, max:604). When optimising the threshold parameters, we performed a grid search over the values [0.05, 0.30] with step 0.05 for the development and 0.001, 0.01, 0.02, 0.05, 0.10, 0.20 for the population sets. The number of MeSH terms to add to a query were parametrised to 1, 5, 10, 15, 20, 25. Evaluation on the final query is performed on the validation (as it was for Hausner et al.[4]) and unseen sets. As in the work by Simon et al.[20], queries are evaluated using precision, recall (sensitivity), and number needed to read (NNR). Additionally, we compute  $F_\beta$  for  $\beta = 0.5, 1, 3$  (standard values used to evaluate automatic systematic review methods [15]).

## 4 Results

We first compare the results obtained using the computational method for objectively deriving search strategies against those the queries originally used to



**Fig. 2:** Variance across topics for the best selected values for different parameters, via optimisation of the considered evaluation measures on validation set.



**Fig. 3:** Correlation of  $F_1$  between validation and unseen for the results presented in Table 2. Pearson's  $r$  correlation is signified in the title of each subplot.

retrieve studies. We then study the effect selection of terms has on queries. Next, we contrast the differences between adding versus not adding MeSH terms and report the differences between the most effective and the least effective query when MeSH terms are not added and when they are added. And finally, we compare our adaptation to the original method.

Empirical results obtained by applying our method to queries is reported in Table 2. Our approach produces queries that are tuned for different evaluation measures. We show that for the  $F_\beta$  variations, NNR, precision, and recall, there are queries which outperform the original queries for each of these measures. While tuning parameters produces gains over the original queries, it introduces a trade-off. Generally, after tuning, queries with gains for precision measures obtain significant losses for recall measures compared to the original queries (e.g., queries tuned for  $F_1$  without adding MeSH terms obtain the highest precision, but suffer a significant loss in recall). Likewise, where there are gains in recall, there are significant losses in precision compared to the original queries (e.g., queries tuned for recall with the addition of MeSH terms obtained the highest recall, however suffer a significant loss in precision). Figure 2 highlights

(a) Prevalence ( $p$ ) of terms for CD009135. (b) Prevalence ( $p$ ) of terms for CD010276.

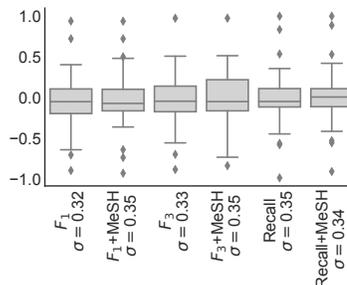
	development n=39				Population n= 30m				
	$p$		$n$		$p$		$n$		
	$p$	$n$	$p$	$n$	$p$	$n$	$p$	$n$	
visceral	0.9231	36	0.0026	76907	oral	0.9630	26	0.0356	1072216
in	0.8974	35	0.0235	707649	in	0.9630	26	0.0235	707649
test	0.8974	35	0.0847	2553063	to	0.8519	23	0.0237	715449
leishmaniasis	0.8462	33	0.0009	27225	patients	0.6667	18	0.1826	5503941
to	0.7436	29	0.0237	715449	specificity	0.6667	18	0.0401	1209795
patients	0.6154	24	0.1826	5503941	lesions	0.6296	17	0.0202	610428
positive	0.5641	22	0.0513	1546457	sensitivity	0.5926	16	0.0423	1275222
specificity	0.5385	21	0.0401	1209795	detection	0.5556	15	0.0289	872069
is	0.5385	21	0.0223	671619	is	0.5556	15	0.0223	671619
sensitivity	0.4872	19	0.0423	1275222	malignant	0.5185	14	0.0122	369252

**Table 3:** Prevalence of the top 10 terms in development and Population sets for the most effective (Sub-table 3a), and least effective (Sub-table 3a) topics in  $F_1$ .

the differences in parameter choices tuned for each evaluation measure. Higher DF thresholds for the development and lower DF thresholds for the population lead to queries with higher precision and lower recall. Lower DF thresholds for the development and higher thresholds for the population lead to queries with higher recall and lower precision. Furthermore, adding more MeSH terms increased recall but at the expense of precision, as expected. Finally, while using the validation set appears to be a good indication of how the query will perform on unseen data, over-fitting leads to the trade-off in precision and recall. The correlation between the performance on the validation data and the unseen data is presented in Figure 3. The figure suggests that performance obtained when tuning parameters on the validation set are strongly correlated with those obtained on unseen data (for the same parameters values). However, we find that as more weight is placed on precision, the correlation between performance on the unseen data and validation data becomes weaker.

We further analyse the queries by studying the terms that were considered for inclusion by thresholding: we focus on queries that did not have MeSH terms added. Table 3 provides a comparison between the highest performing topic in terms of  $F_1$  on the unseen set (Table 3a, topic CD009135: *Rapid tests for the diagnosis of visceral leishmaniasis in patients with suspected disease*) and the lowest performing topic (Table 3b, topic CD010276: *Diagnostic tests for oral cancer and potentially malignant disorders in patients presenting with clinically evident lesions*). Both topics contain high prevalence terms from the title of the systematic review – indicating they are likely relevant to the topic. However, this is the case for terms identified in the best performing query, as well as for those in the worst. This suggests that the identification of terms likely plays only a partial role in the effectiveness of the query – and that selection and location within the Boolean syntax of the query may also be conducive of effectiveness.

CD009579	CD009647
Parasite Egg Count	Aged, 80 and over
Schistosomiasis haematobia	Water-Electrolyte Balance
Antigens, Helminth	Body Water
Sensitivity and Specificity	Osmolar Concentration
Schistosoma haematobium	Electric Impedance
Schistosomiasis mansoni	Dehydration
Hematuria	Sodium
Schistosoma mansoni	Reproducibility of Results
Feces	Prospective Studies
Prevalence	Urine



**Table 4:** Top 10 MeSH terms identified in development set for CD009579 (left, highest  $F_1$ ) and CD009647 (right, lowest  $F_1$ ). **Fig. 4:** Normalised differences in no. of terms in queries between our method and relaxed original.

We also study the interplay between the number of studies provided in the development set and query effectiveness. One may hypothesise that a higher number of studies in the development set is associated to higher effectiveness (given that topic CD009135 contains 39 references in the development set, while topic CD010276 contains 27). Indeed there is a moderate positive correlation (Pearson’s  $r = +0.51$ ) between the size of the development set (used to derive terms) and  $F_1$  on unseen data. Similarly, we study the interplay between the number of terms in the final query and effectiveness. We found a weak negative correlation (Pearson’s  $r = -0.2$ ) between the number of terms in queries and the actual effectiveness on the unseen set ( $F_1$ ). This suggests that including few representative terms is more conducive of effectiveness than many broad terms.

Differences between queries with MeSH terms and those without are also analysed. The results in Table 2 suggest a trade-off in precision and recall when MeSH terms are added. When MeSH terms are added, we observe a higher recall, as expected, but lower precision than when MeSH terms were not added. Queries with MeSH terms did not retrieve a significantly higher number of studies than those without. We now study the effect of adding MeSH terms in more detail, specifically on queries where parameters were tuned for  $F_3$ , where the highest gains were observed overall. The most effective and least effective queries in terms of  $F_1$  were for topics CD009579, *Circulating antigen tests and urine reagent strips for diagnosis of active schistosomiasis in endemic areas* (precision: 0.0330, recall: 0.1765,  $F_1$ : 0.0556), and CD009647, *Clinical symptoms, signs and tests for identification of impending and current water-loss dehydration in older people* (precision: 0.0002, recall: 0.4286,  $F_1$ : 0.0003). The MeSH terms identified for addition to these queries are listed in Table 4. Figure 5 presents the two queries for comparison. Sub-figures 5a and 5b contain the aforementioned best and worst queries derived by optimising  $F_1$ . The identified MeSH terms lead to small improvements in recall when they are added to the query, at the expenses of a substantial drop in precision. When observing the performance on the validation data, topic CD009579 still performs better than topic CD009647. This suggests that this particular topic was more difficult to search for. This is reflected in the queries originally formulated for these topics. The original

```
((schistosomiasis OR cca OR used OR found OR hematuria OR either) AND (strips OR urinalysis
OR dipstick OR dipsticks OR mg OR Parasite Egg Count) AND (village OR villages OR kg))
```

(a) Highest  $F_1$ ; query for topic CD009579.

```
((found OR urine OR balance OR Aged, 80 and over) AND fluid)
```

(b) Lowest  $F_1$ ; query for topic CD009647.

**Fig. 5:** Computationally derived queries. Queries refer to the most effective (Sub-figure 5a) and the least effective (Sub-figure 5b) in terms of  $F_1$  (by optimising for  $F_1$ ) among those in the collection. The queries do not contain field restrictions for space reasons. MeSH terms indicated by *italics* ([Mesh Terms:noexp]). In all other cases the [Title/Abstract] field restriction was applied.

```
1. prostate.ti,ab.
2. psa.ti,ab.
3. used.ti,ab.
4. either.ti,ab.
5. seed.ti,ab.
6. symptom.ti,ab.
7. ml.ti,ab.
8. toxicities.ti,ab.
9. prostatic.ti,ab.
10. Prostatic Neoplasms/
11. or/1-10
12. beam.ti,ab.
13. brachytherapy.ti,ab.
14. radical.ti,ab.
15. prostatectomy.ti,ab.
16. ebrt.ti,ab.
17. cox.ti,ab.
18. androgen.ti,ab.
19. implantation.ti,ab.
20. consensus.ti,ab.
21. pretreatment.ti,ab.
22. sexual.ti,ab.
23. neoadjuvant.ti,ab.
24. mailed.ti,ab.
25. implant.ti,ab.
26. curative.ti,ab.
27. or/12-26
29. 11 and 27
```

```
1. cancer.ti,ab,sh.
2. adenocarcinoma.ti,ab,sh.
3. 1 or 2
4. prostat*.ti,ab,sh.
5. 3 and 4
6. Prostatic Neoplasms/
7. 5 or 6
8. seed*.rs.
9. permanent*.ti,ab,sh.
10. 8 or 9
11. implant*.ti,ab,sh.
12. 10 and 11
13. Brachytherapy/
14. Brachytherapy.ti,ab,sh.
15. or/12-14
16. 7 and 15
```

**Fig. 7:** Manually derived objective query (commonalities in *italics*).

	# Ret	NNR	Precision	Recall
Manual	78913	6070.23	0.0002	1.0000
Computational	48945	3496.14	0.0003	1.0000

**Table 5:** Difference in effectiveness between the computationally derived query and the manually derived query.

**Fig. 6:** Computationally derived objective query.

query for topic CD009579 (precision: 0.0019, recall: 1,  $F_1$ : 0.0038) performs better than the original query for topic CD009647 (precision:  $2.6^{10e-5}$ , recall: 1,  $F_1$ :  $5.2^{10e-5}$ ). Note that although the performance of the queries are similar, the manually formulated queries have many more terms (CD009579 – derived: 15, original (relaxed): 40; CD009647 – derived: 6, original (relaxed): 199). However, looking at the distribution of query lengths for each method in Figure 4, not only is there little variation in the total number of terms in queries, but there is little variance in amount of terms added in the automatic method and in the relaxed versions of the original queries.

Finally we study a query derived manually (by Hausner et al. [4], Figure 7) and the same query derived computationally (Figure 6). Table 5 presents the

differences in effectiveness given the number of documents retrieved, NNR, precision, and recall on the same set of validation documents (note that only the development and validation sets are used to make a fair comparison to the manual method; thus no tuning was used). The query derived objectively using our computational method retrieves less documents, but maintains recall: this results in a saving of approximately USD\$90,000 (considering double screening and the costs/times per study reported by McGowan et al. [13]).

## 5 Conclusion

We presented a computational approach to objectively deriving search strategies for systematic reviews. This approach adapts and extends the proposal of Hausner et al.[4], to further reduce human subjectivity in an otherwise objective methodology. The computational method can be used as a starting point for query formulation, as demonstrated by our results. The manual objective method included human intervention; our computational adaptations and extensions approximated the steps a human would take. To better approximate these steps, we will set up an interactive query formulation study with information specialists. The feedback and results from this can be used to improve computational methods and would provide us with the means to fairly compare our computational approach with the ad-hoc method.

We have identified a number of avenues for further extending the fully automatic approach and its empirical evaluation. Firstly, randomness is introduced in this method when the test set is split into development, validation, and unseen. The use of 3-fold cross validation would reduce experimental bias in the subsequent phases. Next, we have observed that in the current approach only unigrams are used as candidate terms for possible inclusion in queries (this is also the case in the original method and has already been identified as an issue [4]). This is a limitation because the semantic context that may have been encoded as a phrase (e.g., in an n-gram such as “myocardial infarction”) is lost. We suggest that by automatically identifying medical phrases using automatic tools such as MetaMap, this shortcoming can be overcome.

A limitation of this work is that a prospective study was not undertaken. New queries formulated using our method may have retrieved unjudged but relevant studies. A future extension of this work could be to use our proposed method to identify the proportion of new relevant studies retrieved (if any). Another task to be considered is to automatically further refine derived queries. Scells and colleagues [18,19] have found that automatic generation and refinement techniques can improve the effectiveness of existing Boolean queries.

The computational method presented is envisioned to be integrated into tools for assisting researchers conducting systematic reviews (for example, query suggestion [8]). The aim is not to replace humans constructing search strategies — at the very least, a number of gold standard studies are still needed to seed this approach, (as is typically the case in this context [10]). Query formulation is currently a highly subjective and error-prone process, and reducing subjectivity and mistakes in search strategy construction can only lead to less biased, reproducible, and timely systematic reviews.

**Acknowledgements.** Harrisen is the recipient of a CSIRO PhD Top Up Scholarship. Dr Guido Zuccon is the recipient of an Australian Research Council DE-CRA Research Fellowship (DE180101579) and a Google Faculty Award. This research is supported by the National Health and Medical Research Council Centre of Research Excellence in Informatics and E-Health (1032664).

## References

1. Aronson, A.R.: Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In: Proceedings of the AMIA Symposium. p. 17. American Medical Informatics Association (2001)
2. Clark, J.: Systematic reviewing. In: Suhail A. R. Doi, G.M.W. (ed.) *Methods of Clinical Epidemiology*. Springer (2013)
3. Golder, S., Loke, Y., McIntosh, H.M.: Poor reporting and inadequate searches were apparent in systematic reviews of adverse effects. *Journal of Clinical Epidemiology* (2008)
4. Hausner, E., Waffenschmidt, S., Kaiser, T., Simon, M.: Routine development of objectively derived search strategies. *Systematic reviews* **1**(1), 19 (2012)
5. JPT CHH, G.S.: *Cochrane handbook for systematic reviews of interventions version 5.1.0 [updated march 2011]*. The Cochrane Collaboration (2011)
6. Kanoulas, E., Spijker, R., Li, D., Azzopardi, L.: Clef 2018 technology assisted reviews in empirical medicine overview. In: CLEF 2018 Evaluation Labs and Workshop: Online Working Notes (2018)
7. Karimi, S., Pohl, S., Scholer, F., Cavedon, L., Zobel, J.: Boolean versus ranked querying for biomedical systematic reviews. *BMC MIDM* **10**(1), 1 (2010)
8. Kim, Y., Seo, J., Croft, W.B.: Automatic boolean query suggestion for professional search. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (2011)
9. Lau, J.: *Systematic review automation thematic series* (2019)
10. Lee, G.E., Sun, A.: Seed-driven document ranking for systematic reviews in evidence-based medicine. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 455–464 (2018)
11. Leeftang, M., Deeks, J., Takwoingi, Y., Macaskill, P.: Cochrane diagnostic test accuracy reviews. *systematic reviews*. 2013; 2: 82. pubmed pmid: 24099098. pubmed central pmcid: Pmc3851548. Tech. rep., Epub 2013/10/09. eng
12. Lefebvre, C., Manheimer, E., Glanville, J.: Searching for studies. *Cochrane handbook for systematic reviews of interventions: Cochrane book series* pp. 95–150 (2008)
13. McGowan, J., Sampson, M.: Systematic reviews need systematic searchers (irp). *Journal of the Medical Library Association* **93**(1), 74 (2005)
14. McGraw, K.A., Anderson, M.J., et al.: Analysis of the reporting of search strategies in cochrane systematic reviews. *Journal of the Medical Library Association* **97**(1), 21 (2009)
15. O’Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., Ananiadou, S.: Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews* **4**(1), 5 (2015)
16. Sayers, E.: *A general introduction to the e-utilities*. Entrez Programming Utilities Help [Internet]. Bethesda: National Center for Biotechnology Information (2010)

17. Scells, H., Locke, D., Zuccon, G.: An information retrieval experiment framework for domain specific applications. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (2018)
18. Scells, H., Zuccon, G.: Generating better queries for systematic reviews. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. SIGIR '18 (2018)
19. Scells, H., Zuccon, G., Koopman, B.: Automatic boolean query refinement for systematic review literature search. In: Proceedings of the 2019 World Wide Web Conference (2019)
20. Simon, M., Hausner, E., Klaus, S.F., Dunton, N.E.: Identifying nurse staffing research in medline: development and testing of empirically derived search strategies with the pubmed interface. *BMC medical research methodology* **10**(1), 76 (2010)
21. Tsafnat, G., Glasziou, P., Choong, M.K., Dunn, A., Galgani, F., Coiera, E.: Systematic review automation technologies. *SR* **3**(1), 74 (2014)
22. Uzuner, Ö., South, B.R., Shen, S., DuVall, S.L.: 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association* **18**(5), 552–556 (2011)