Intelligent Systems with Applications 15 (2022) 200091

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

Search strategy formulation for systematic reviews: Issues, challenges and opportunities

Andrew MacFarlane [a,*], Tony Russell-Rose [b], Farhad Shokraneh [c]

[a] Centre for HCI Design, Department of Computer Science, City, University of London, Northampton Square, London EC1V 0HB, UK
[b] Department of Computing, Goldsmiths, University of London, New Cross, London SE14 6NW, UK
[c] Institute of Health Informatics, University College London, 222 Euston Rd, London NW1 2DA, UK

A B S T R A C T

Systematic literature reviews play a vital role in identifying the best available evidence for health and social care research, policy, and practice. The resources required to produce systematic reviews can be significant, and a key to the success of any review is the search strategy used to identify relevant literature. However, the methods used to construct search strategies can be complex, time consuming, resource intensive and error prone. In this review, we examine the state of the art in resolving complex structured information needs, focusing primarily on the healthcare context. We analyse the literature to identify key challenges and issues and explore appropriate solutions and workarounds. From this analysis we propose a way forward to facilitate trust and to aid explainability and transparency, reproducibility and replicability through a set of key design principles for tools to support the development of search strategies in systematic literature reviews.

Introduction

Systematic literature reviews are used to identify and synthesise the best available evidence to support health and social care policies and decisions. The systematic review process starts with specification of the research question and development of a protocol (Hemingway & Brereton, 2009; Higgins et al., 2020). This is followed by a search of the literature to identify relevant studies, screening of these studies, analysis of this evidence and then synthesis in the finished review. Information Retrieval (IR) techniques are essential to the process, particularly in the literature search and study screening phases. A key element is the literature search phase, since errors in this stage can have a significant impact on the overall success of the review and lead to an inaccurate, incomplete, and invalid outcome, e.g., by introducing bias in the synthesis (McGowan & Sampson, 2005). Rectifying a review once it has been completed can be costly; analogous to the problem in software engineering where errors overlooked early in the process (e.g., in the requirements stage) can be significantly more difficult to correct later.

In this review, we focus on literature searching, specifically the development of the search strategies used in systematic reviews. This is a complex process (Cooper et al., 2018; Lefebvre et al., 2020), in which the search methods and choice of databases to be used to identify literature for the systematic review are specified and peer reviewed. It is recommended that the protocol is approved by peer review before searching is carried out i.e., that the process to identify relevant literature is both rigorous and appropriate. The search strategies are then planned and designed. Initial scoping searches may be carried out and the strategy is iteratively developed until the review team is satisfied that the strategy is fit for purpose. It is then applied to gather studies for the screening phase.

This process is time consuming and even a 'rapid review' can take between three to six months (Hemingway & Brereton, 2009), whilst a full-scale review can take as much as 67 weeks on average (Borah et al., 2017). Relevant studies may be published in the meantime (Shojania et al., 2007), rendering some reviews out of date by the time they are published. Consequently, periodic updates of the search strategy may be required (Shokraneh & Russell-Rose, 2020; Thomas et al., 2017). The search strategies used should therefore be published (e.g., as an appendix to the review) and be fully replicable by third parties. They should also be transparent and explainable, so they can be reproduced as and when required. This is not only to minimise bias and errors, but also to reduce costs and hence reduce waste in research (Chalmers et al., 2014).

Automated solutions have the potential to help searchers to design, build and deploy complex search strategies, but they need to be able to trust these solutions provided. The aim of this paper is twofold: 1) to review current and best practice on search strategy formulation from both an evidence synthesis and IR perspective, and to critically analyse the state of the art in terms of reproducibility, replicability, transparency, explainability and trust and 2) to propose a set of design

* Corresponding author.
E-mail addresses: andym@city.ac.uk (A. MacFarlane), T.Russell-Rose@gold.ac.uk (T. Russell-Rose), f.shokraneh@ucl.ac.uk (F. Shokraneh).

https://doi.org/10.1016/j.iswa.2022.200091
Received 10 August 2021; Received in revised form 20 March 2022; Accepted 1 June 2022
Available online 7 June 2022

principles to address identified issues. It should be noted that whilst we focus on healthcare, much of the analysis applies to other areas of professional search including patent search, legal research and recruitment (Russell-Rose et al., 2018).

The review is structured as follows. We provide our sampling process and inclusion criteria for the review. Key terms are then defined and the review scope outlined. Current practice in the development of search strategies in healthcare together with their deficiencies are reviewed. We then review various tools and approaches put forward to augment search strategy development methods and also assess their strengths and weaknesses. The issues identified in both current practice and supporting tools and approaches are used to build a framework of design principles, specifically to address the shortcomings in current approaches in the search strategy formulation process. We conclude with a summary, outlining the steps forward.

Sampling process and inclusion criteria

The sampling process started with a number of known studies as seed documents e.g., Sampson & McGowen (2005) and eHealth evaluation at CLEF (Conference and Labs of the Evaluation forum). Citations from these were used to identify further sources using a citation pearl growing strategy. These documents were used to build a search strategy with two main facets: 'search strategies' and 'systematic reviews'. Searches were carried out in early June 2020 and repeated in March 2022. This search strategy was applied to Google, Google Scholar, and PubMed. These sources were chosen to gather information from the general, computer science and medical literature. Where possible we restricted some terms to title only i.e., the 'systematic review' search term and its synonyms. We identified further synonyms using automated term suggestions. Where sources provided 'filter by date' functionality, we restricted our search to the period 2000 to 2022 to ensure that the most recent studies were included. We extracted citations from found documents to identify more relevant studies. We only considered English language publications and all accessible publications were reviewed (using the institutional subscription to access articles as and when necessary). In sources that focus on medical literature, we repeated the search using the 'search strategy' facet only. Where general issues were addressed, we carried out specific searches to identify literature e.g., reproducibility and replicability. In terms of screening, we sorted results by relevance and viewed all results to the end of the hitlist until no further relevant studies were found. Documents considered relevant were those that focused on methods and algorithms to support search strategies. Documents that focused only on evaluation issues were excluded. Using this multifaceted strategy, we were able to identify all the major articles and studies for this review to the best of our knowledge.

Definition of terms and review scope

A number of key terms have been introduced above that require precise definition. This is particularly important where there is disagreement over what these terms actually mean. As we draw on literature from multiple domains (healthcare, information science, computer science) we need consistent definitions to provide a coherent narrative. These key terms are critical to understanding the role of search strategies in systematic review. First, there is a requirement for search strategies to be *reproducible* and *replicable*. As a consequence, strategies need to be also *transparent* and *explainable*, so that any searcher with the appropriate skills should be able to reuse a given search strategy appropriately. Second, since searchers desire maximal *control* over the system, there should be a high level of *trust* of the techniques and tools used. Searching for evidence to inform systematic reviews in general requires high recall, since omissions can have serious consequences for the validity of the outcome (e.g., clinical guidance may be based on incomplete or biased evidence). We start with *trust* and weave our way through the concepts such as *control, transparency* and *explainability* until we address *reproducibility* and *replicability*. Finally, we define a number of barriers that can prevent the successful achievement of concepts identified thus far.

Trust

This is a user-focused term, and in our context it means that a searcher trusts the system presented and will use it to conduct a search to find information. The concept of 'Humans in the Loop' is important (Grames et al., 2019; Marshall & Wallace, 2019), as users feel that they need to be in *control* of the system, as opposed to the system controlling them. This issue of *control* reoccurs in many contexts, often associated with automated or artificial intelligence (AI) solutions.

Control

Where the user has maximal control of the system, they are able to use it to solve their problem i.e., to formulate a search strategy that is both *transparent* and *explainable*. These two terms are related, but different. Control leads to trust in the system, lack of control leads to distrust (as will be seen below). An example of control in the context of search strategies is turning on/off auto term mapping or query expansion, and informing the user of the impact of any change. In our context this means that the user has maximal control (Paisley & Parker, 1965) over the design and development of the search strategy used in the systematic review.

Transparency and explainability

Transparency and explainability are closely related concepts, and it can be argued that the difference between them is more quantitative rather than qualitative. Both imply a degree of determinism and predictability, whereby the output from a process can be predicted based on the input. A system that is transparent should provide the user with insight into its inner workings so that its behaviour is predictable (Doran et al., 2017). This contrasts with an opaque system, in which the inner workings are not open to scrutiny - which is the default for many machine learning (ML) algorithms.

A system that is explainable should provide additional insight so that its actions are interpretable according to some mental model. In our context this means that a user should understand why a search strategy produced a particular outcome, e.g., by presenting the inner workings and facilitating the acquisition of a clear and accurate mental model of their behaviour (Russell-Rose & MacFarlane, 2020). In this respect, transparency is a necessary (but not sufficient) criterion for explainability. There is evidence that current IR systems are not explainable in this respect (Russell-Rose & MacFarlane, 2020). However, users are more likely to trust a system if its behaviour is transparent and explainable, which in turn facilitates reproducibility and replicability.

Reproducibility and replicability

The issue of reproducibility has become a major issue not only in healthcare sciences (Shokraneh, 2019) but also in IR experimentation. This has been addressed in CLEF track specifically focused on reproducibility and replicability issues (Ferro et al., 2019). We use the ACM definitions (ACM, 2020), which are defined as (Plesser, 2018):

- Reproducibility: a different team uses the same test collection, and the results are directly comparable with another team's results.
- Replicability: a different team uses a different test collection (topics, documents, relevance assessment and results differ).

Barriers

Barriers to success in addressing concepts such as trust must be

addressed, and throughout the review, we highlight these through an analysis of the literature in the field. We identify three broad barriers that are impediments to the successful formulation of a search strategy (note that several barriers can apply any given issue identified):

- Formalism, i.e., the issue is intrinsic to the representation scheme (e.g. Boolean logic and document centric media)
- Platform, i.e., the issue is related to the tooling and technologies used (e.g. command-line query builders, proprietary databases, etc.)
- Community, i.e., the issue is related to the processes, practices and conventions adopted by the community.

We use these terms throughout the review to highlight the many problems that have been identified through the literature analysis, identifying the barriers that emerge through inspection of the literature.

Search strategy development methods: Current practice

Boolean logic is the de facto approach to structured searching in general and for healthcare systematic reviews in particular (Russell--Rose et al., 2018). There are variations on the process (Clarke et al, 2021), but in most cases it involves the following steps:

1 Subdividing a review topic/question into its constituent concepts. This can be facilitated by the use of conceptual frameworks or templates such as PICO (Shokraneh, 2016);
2 Identifying which concepts should be represented as discrete search blocks or 'facets';
3 Collect terms for each facet, based either on natural language keywords or controlled vocabulary terms;
4 Combining the facets using Boolean and other operators to create an initial search strategy;
5 Testing the strategy to determine the number of results and an initial direction for refinement (with limited time and resources it may not be possible to screen every result). At this stage; it may also be appropriate to consider the use of published search templates or 'filters';
6 Sharing the results with team members for comments on addition, deletion or modification of terms;
7 Execution of the search and translation to the syntax of other databases;
8 Reporting and documenting of the search strategy.

The output of this process - namely the search strategy - can be time consuming and difficult to replicate. Moreover, a number of issues have been identified that limit both the transparency and explainability of current approaches. In the following sections, we provide an example Boolean search strategy and then review the shortcomings of this approach.

Boolean search strategies for systematic reviews

The formulation of search strategies (steps 3-7 above) is based on facets that are developed from concepts (steps 1-2 above). Within each facet a set of synonyms or related terms is identified, and these are connected by applying the Boolean OR operator. A composite expression may then be formed by applying the Boolean AND operator across facets. The Boolean AND NOT operator may be used to remove unwanted concepts or terms. Extended Boolean operators can also be applied to narrow or broaden the search. For example, adjacency operators (e.g., NEXT, WITHIN, ADJ) can be applied to multiple terms to identify a collocation, e.g., "randomized" within 2 words of "controlled trial". Truncation or wildcard operators (*,?,&) can be applied to a single term to truncate a term or to identify spelling variations e.g. randomi?ed, to match one or more characters in that string. Some databases allow terms to be restricted to particular fields, e.g., the term "diet" within the title field of an article ("diet[title]"). Restrictions on fields are also available e.g., "pain freq/3" where the search term 'pain' must appear at least 3 times for the document to be retrieved. There are many ways to construct a Boolean query with these operators including single paragraph (Bramer et al., 2018), block by block (Markey & Cochrane, 1981) and line by line. The choice on constructing a Boolean query has a significant impact on transparency and explainability. Consider Fig. 1 that shows an example of a complex and hard to interpret 'line by line' Boolean search strategy (Francis et al., 2015).

```
1. randomized controlled trial.pt.
2. controlled clinical trial.pt.
3. randomized.ab.
4. placebo.ab.
5. clinical trials as topic.sh.
6. randomly.ab.
7. trial.ti.
8. 1 or 2 or 3 or 4 or 5 or 6 or 7
9. (animals not (humans and animals)).sh.
10. 8 not 9
11. exp Child/
12. ADOLESCENT/
13. exp infant/
14. child hospitalized/
15. adolescent hospitalized/
16. (child$ or infant$ or toddler$ or adolescen$ or teenage$).tw.
17. or/11-16
18. Child Nutrition Sciences/
19. exp Dietary Proteins/
20. Dietary Supplements/
21. Dietetics/
22. or/18-21
23. exp Infant, Newborn/
24. exp Overweight/
25. exp Eating Disorders/
26. Athletes/
27. exp Sports/
28. exp Pregnancy/
29. exp Viruses/
30. (newborn$ or obes$ or "eating disorder$" or pregnan$ or childbirth or virus$ or
influenza).tw.
31. or/23-30
32. 10 and 17 and 22
33. 32 not 31
```

Fig. 1. An example search strategy for Review of 'Oral protein calorie supplementation for children with chronic disease'.

This example highlights two shortcomings of the Boolean approach (Russell-Rose & Shokraneh, 2020). First, it is difficult to establish the overall structure of the strategy, i.e. how facets are related and how terms are combined to form those facets. Second, such strategies do not scale well – an information need may require the use of many facets and many terms spread over a number of pages. This is a key *formalism* barrier. Although Boolean logic remains the default approach, this lack of transparency and explainability inhibits reproducibility and replicability. Additional issues regarding development methods, errors and system heterogeneity are outlined below.

Search strategy development

Although the basics of Boolean search can be learnt in relatively short order (MacFarlane & Russell-Rose, 2016), there are as yet no standards for designing and implementing such searches. This means that developing professional search skills entails a steep learning curve (Yoo & Mosa, 2015) and the knowledge required to successfully use Boolean strategies is typically acquired in the workplace (a *community* barrier). It takes many years to acquire the knowledge required to apply Boolean logic effectively. Designing an effective strategy can be particularly difficult when searchers do not know the best tactics to use (e.g., which combinations of keywords and operators), and as a result resort to exploratory searches (Hoang & Schneider, 2018) - a *community* barrier. This entails multiple iterations and can add further costs to the development of a systematic review (Hemmingway and Brereton, 2009; Borah et al., 2017), and it is often hard for the searcher to know when to stop (Booth, 2010) i.e. to recognise when sufficient studies have been identified - both *community* and *platform* barriers. Guidelines such as PRESS 2015 can be used to audit the strategy (McGowan et al., 2016), but these contain many open questions, are subjective (Shokraneh,

2018) and it can be very expensive to revise a strategy once the systematic review has been finished or published - a *community* barrier. Although it is considered good practice to publish the strategy along with the review, this practice is not always observed, further compromising reproducibility (Biocic et al., 2019; Koffel & Rethlefsen, 2016). Lack of documentation for the iterative process of search strategy development (Russell-Rose and Shokraneh, 2020) is both a *formalism* and *platform* barrier. This is compounded by the lack of a central repository for strategies leading to duplication of work (Biocic et al., 2019; Koffel & Rethlefsen, 2016; Shokraneh, 2018), which is both a *community* and *platform* barrier. There is also evidence that conventional professional search systems based on the Boolean model do not fully support searchers' needs (Russell-Rose & Chamberlain, 2017), particularly regarding advanced functionality such as merging search queries and search results and the ability to publish search strategies to assist reproducibility and replicability - key *platform* barriers.

Errors in search strategies

In one study of published MEDLINE search strategies, it was shown that as many as 90% had at least one error and of these 80% were errors that limited the effectiveness of the search (Sampson & McGowan, 2006). In an investigation of Cochrane reviews, 73% were found to have a faulty search strategy design with 53% containing errors limiting the accuracy of the searches (Franco et al., 2018), with a more recent investigation report an error rate of 92.7% in search strategies for systematic reviews (Salvador-Oliván et al., 2019). Analysis of these errors shows they can be split into three types (MacFarlane & Russell-Rose, 2016):

- Strategic Errors: Incorrect line number referrals (see Fig. 1) can lead to invalid intermediate set merges. Using overlapping search elements leading to redundancy and increased run time for searches. Search strategy is not correctly translated to the database's syntax.
- Tactical Errors: Typically are spelling errors or a missed spelling variant by incorrect use of truncation operators. Incorrect or irrelevant use of subject heading terms. Missing synonyms.
- Logical Errors: Incorrect use of operator e.g. using OR when AND was required.

These errors are present in strategies developed by searchers with significant prior knowledge and experience of Boolean methods, which further highlights the lack of transparency and explainability. This is a key *formalism* barrier.

Heterogeneity in systems

In domains such as healthcare, it is common for users to search multiple databases. Consequently, they must translate their strategy between systems as the operators, fields and knowledge organisation schemes can differ (Bramer et al., 2018). The standard Boolean operators (AND, OR) may have the same syntax and semantics, but the NOT operator can be misinterpreted (NOT is a Boolean unary operator but is often implemented as a binary AND NOT operator in search syntax). Moreover, proximity and truncation operators can differ significantly between database search interfaces. The use of double quotes ("") is one well known method, but variations such as WITHIN, NEAR, or ADJ and NEXT are used to specify the number of words between two search terms, e.g., cancer WITHIN/2 treatments. A further complexity is that proximity operators can either be symmetric or non-symmetric e.g., cancer WITHIN/2 treatments may not be the same as treatments WITHIN/2 cancer if the database treats the order on the search terms as significant. Operators with identical syntax can be treated differently e.g., NEAR/5 in Web of Science specifies a maximum of 5 words between search terms, but the same expression in Embase specifies a maximum of 4 words. Some databases support the use of adjacency between bracketed terms and some either do not or malfunction. Furthermore, there can be a bewildering array of truncation or wildcard operators to capture a single concept, which can operate on single characters, e.g. randomi?ed or multiple characters, e.g. random*. Truncation operators can also differ in semantics e.g., '?' in Embase.com references a single character, whilst in Ovid SP this is 0 or 1 characters; confusingly '$' in Embase.com references 0 or 1 characters whilst in Ovid SP it is an unlimited number of characters. Operators can be prefixed (barely provided by existing search systems), infixed or postfixed and may apply zero or more, or one or more characters.

A further issue is that different terms can be used for the same concept in different databases, with 'AND' between fields e.g. Breast Neoplasms/ OR (Breast/ AND Neoplasms/) (Hoang & Schneider, 2018). This adds further difficulty when among vendor databases, meaning that searchers have to learn different variants of operators and knowledge organisation schemes for the same information need. This has significant implications for the reproducibility and replicability of search strategies, and requires extra effort and a complex methodology to translate the strategies between different databases (Bramer et al., 2018). This heterogeneity issue is a key *platform* barrier.

Summary of problems with Boolean strategies

The complexity of Boolean search strategies leads to a lack of transparency, and it is difficult for searchers to develop effective strategies even with significant prior knowledge and experience. Search strategies can be difficult to conceptualise, time consuming to maintain and error-prone. A further complication is caused by system heterogeneity, whereby strategies need to be translated to systems with different underlying syntax, semantics and knowledge organisation schemes. Explainability and transparency are therefore limited, along with reproducibility and replicability. However, alternative methods have not managed to displace the conventional Boolean approach. Hjørland (2015) mounts a strong defence of Boolean logic and the continued role of expert searchers in complex information needs. In particular he highlights the issue of control, with exact match models such as Boolean logic providing much more control and transparency in query formulation than the best match systems such as ranking schemes. Despite their shortcomings, Boolean strategies are well known and trusted and continue to be used as the foundation for systematic reviews for healthcare and for other domains such as law, patents and recruitment. In the next section, we review approaches that have been used to augment current search strategy development to address issues raised above.

Tools and approaches to augment search strategy development

A number of tools and approaches have been proposed to either augment or replace the methods outlined above for the development of search strategies. Whilst they are gaining traction, they have yet to attract widespread use. Attempts to move beyond the Boolean method have been subject to similar issues regarding transparency, and unless significant benefits are demonstrated, searchers will remain reluctant to move away from conventional methods. The CLEF eHealth track is a notable example of this research direction (Goeuriot et al., 2017; Kelly et al., 2019; Suominen et al., 2018; Kalphov et al., 2017), specifically in the technologically assisted reviews in empirical medicine track (Kanoulas et al., 2017, 2018, 2019). In this section, we review a variety of 1) deployed tools and 2) general approaches including text and data extraction, automatic query expansion (AQE), Tool Automation, Hybrid schemes, Ranking and ML as well as potential methods in data modelling/theory and structure/representation.

Tools 1: Text and data extraction

This is an analytic approach in which data or text is extracted

(Hollmann and Eickhoff, 2017) from documents e.g. to inform search strategies that have long been recognised as being useful tools (Thomas et al., 2011). They are used throughout the systematic review production process, but could be used in the scoping search that needs to be carried out for any review. Searchers have expressed concern about the time consuming nature of extracting information and data from documents (Hoang & Schneider, 2018), and these methods could help with some of the complexity associated with the information need and address common errors such as spelling and non-identification of synonyms. Examples include use of the PICO facet analysis scheme as a template for information extraction (Begert et al., 2020; Burri, 2019; Wallace et al., 2016). A gold study or a set of gold studies can be analysed to extract terms, and each element of the scheme is populated using appropriate search terms connected via Boolean OR. Keyword extraction algorithms can be applied to protocols to identify useful terms (Alharbi et al., 2018). A similar approach extracting data such as the given studies data sample size could also be used to populate the search e.g., as a filter (Marshall & Wallace, 2019). Similarly, keyword co-occurrence networks can be generated from analysed documents to select appropriate terms for Boolean searches (Grames et al., 2019), where some level of reproducibility is possible. Topic modelling can be used to support classification (van Atlena and Olabarriaga, 2017) or clustering (Kalphov et al., 2017) of documents. Hausner et al., and Simon (2012) outline a process to develop filters for queries based on terms extracted from cited documents. O'Mara-Eves et al., (2015), reviewed text mining methods used to support systematic reviews and found that there is almost no attempt at replication between studies - both a *formalism* and *platform* barrier. The best approaches for the application of these techniques are therefore unknown. One attempt at reproducibility (Olorisade et al., 2017) demonstrated severe limitations from published work to match study outcomes independently. Whilst the methods can tackle some issues with the identification of search terms, they cannot address the inherent problems in the complexity of search strategies. The methods rely on Natural Language Processing (NLP) technologies that may not be sufficiently transparent or explainable - a barrier to *community* acceptance. In general, however, information professionals have been prepared to adopt solutions developed in collaboration with their community, such as the use of text mining techniques to identify terms to use in search strategy development, e.g. the Yale MeSH Analyzer (Yale, 2021).

Tools 2: Automatic query expansion (AQE)

Tsafnat et al. (2014) claim that decision support systems (DSS) for automating search strategies can suggest tactics such as choice of keywords and operators and their combinations. Extraction methods outlined in above could also be used to populate the DSS using PICO or other frameworks. These technologies have been used in healthcare (O'Sullivan et al., 2010, 2013), but are focused on the needs of the end users (e.g., clinical staff) and are not utilised in the production of systematic reviews. DSS rely on AI technologies that limit their transparency, and are therefore not particularly explainable.

More conventional IR approaches have been attempted. Scells and Zuccon, (2018a; 2018b; 2019; 2020a; 2020c) have carried out work on query support for systematic reviews. This includes the transformation of a query to create a more effective one, using AQE as part of the strategy. Transformations include syntactic methods such as logical operator replacement, MeSH explosion, field restriction and adjacency replacement (Scells and Zuccon, 2018a; Scells et al. (2019), 2020c) built on this with semantic methods, but the underlying technology is still Boolean. Scells et al. (2020a) looked at building a search strategy in an objective way by reformulating the query by using gold studies to optimise the query, which can be used later in the search strategy development lifecycle. Kim et al. (2011) look at the use of pseudo relevance feedback to build a decision tree to generate Boolean queries, effectively learning the best Boolean query based on traversing and testing various candidates in the tree. These schemes all rely on Boolean logic, and some of the same limitations apply together with associated barriers already identified, but with an added issue of lack of transparency. A survey of query expansion methods for professional search is available in Russell-Rose et al. (2021).

Tools 3: Tool automation

A wide variety of tools to automate the process of developing systematic reviews, including the development of search strategies, examples of which are given above (Lau, 2019). A number of key limitations to the deployment of automated tools have been identified. The systematic review production process specifies a rigorous and clearly defined framework that all searchers must adhere to as part of the development team. A key drawback is that automated tools lack compatibility with the systematic review workflow (van Altena et al., 2019; O'Connor, 2019) - a *community* and *platform* barrier. Issues to do with actually applying the tools in practice also demonstrate limitations. There is evidence of resistance to automated tools for developing search strategies when they are poorly supported (van Altena et al., 2019) - a *platform* barrier. Searchers will find it difficult to use an automated tool that has a steep learning curve associated with it (van Altena et al., 2019; Yoo & Mosa, 2015) - a *community* barrier. Usability of automated tools is questionable as they are often difficult to use, and lack the functionality to support complex information needs (van Altena et al., 2019) - a *platform* barrier. Given the above, there is clear evidence that there are considerable barriers to the adoption of automated tools. Any new system must be seamless and easier to use either to replace or augment current methods than one based on Boolean logic. Any new system must at the least be logically equivalent if not provide clear benefits over current approaches. Automated tools lack transparency and hence explainability. This leads to a lack of trust of new automated tools to support the development of search strategies (O'Connor et al., 2019). The evidence is that these tools have all of the drawbacks but none of the benefits of conventional Boolean search strategy development, which in part explains their limited adoption.

Tools 4: Ranking algorithms, machine learning and learning to rank

Ranking algorithms have been around for many years and are routinely used in many contexts such as web search to present relevant results to users e.g., BM25 (Robertson et al., 1995). Simpler schemes using variations of TF/IDF have been proposed in the field (Scells, Azzopardi, Zuccon & Koopman, 2018, Alharbi and Stevenson, 2017) or extended with the use of lexical statistics such as LogLikelihood, Chi-Square and Odds-Ratio (Alharbi and Stevenson, 2019). These are less used in professional search due in part to a perception of reduced transparency in the algorithms (Russell-Rose et al., 2018). However, there has been interest in developing ranking functions further either using a two-dimensional variation of BM25 (Di Nunzio et al., 2017, 2018a; Di Nunzio, 2018, 2019) or in conjunction with a paragraph2vec approach for query expansion (Wu et al., 2018). Ranking algorithms have been extended with classification schemes such as sequential bayes search (Zou et al., 2018; Zou & Kanoulas, 2020), Naive Bayes (Di Nunzio, 2018; Di Nunzio et al., 2018b) and Logistic Regression (Wu et al., 2018).

The standard ranking models have been augmented with learning to rank schemes (Li et al., 2018). ML has been deployed to automatically filter articles for a given review, using NLP methods to identify terms (Burri, 2019) and others have applied learning to rank methods on those terms extracted from a protocol (Minas et al., 2018) or simply to augment the ranking with word2vec (Chen et al., 2017). Other supervised ML methods used to find articles relevant to a systematic review topic include Neural Networks - NNs (Lee, 2017; Marshall & Wallace, 2019; Singh et al., 2017) and Support Vector Machines - SVMs (Anagnostou et al., 2017; Cohen & Smalheiser, 2018; Wallace et al., 2010).

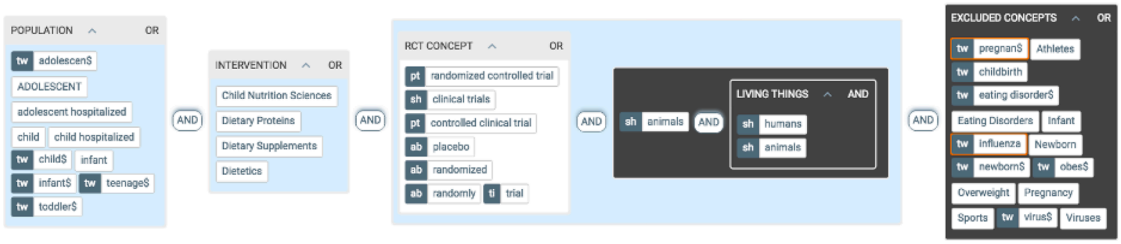

Fig. 2. 2Dsearch visual block representation of line by line strategy from Fig. 1.

Scells et al., and Koopman (2017, 2020b) suggest methods for using Learning to Rank models to rank Boolean queries in much the same vein as the AQE methods reviewed above.

Continuous active ML approaches have also been proposed that have become particularly prominent in the field, that are similar to dynamic ML for IR approaches widely proposed (Sloan & Wang, 2015). This scheme has been used for a wide variety of purposes including Active Ranking (Wallace et al., 2013), Learning to Rank (Li & Kanoulas, 2019) to identify the initial set of documents, classification of documents once identified (Cormack & Grossman, 2016a), or to identify a stopping strategy for article selection (Cormack & Grossman, 2016b; Cormack & Grossman, 2017; Di Nunzio, 2020; Li & Kanoulas, 2020; Hollmann and Eickhoff, 2017). The use of Logistic Regression ML approaches have also been proposed in conjunction with active learning (Cormack & Grossman, 2018; Norman et al., 2017, 2018; Li et al., 2020).

Thomas et al. (2017) review methods for living systematic reviews, using information filtering methods that include ML for classification and information extraction that require periodic updates. These methods can and do show utility, but suffer from a major problem. As with many AI technologies their explainability is limited, and this has undoubtedly inhibited their adoption, and all the barriers identified in the previous sections apply here. Further issues with such technologies include uses on very specific problems e.g., one type of study design or studies within one key topic such as randomised controlled trials. They are often not usable for multi-faceted searches of the type addressed in this review, but can be used as a useful tool for either for query filters ("Hedges") or query reformulation once the query has been developed (Del Fiol et al., 2018; Russell-Rose et al., 2021). Some models such as BioBERT are beginning to have a significant impact in the field (Lee et al., 2020).

Tools 5: Hybrid schemes

Karimi et al., and Zobel (2010) undertook a number of different experiments, in particular a hybrid scheme where Boolean and rankings schemes are used in conjunction. They focus very much on the reproducibility issue and the problems formulating search strategies using Boolean logic, but are realistic as to the downsides of ranking. There is evidence that shows that the hybrid method does push more relevant documents higher up the ranked list, but this still does not address the conceptual problem of building the search strategy in the first place.

Tools 6: Structures/representation

It is possible to move away from the imperative methods used in Boolean search strategies to a more declarative method where relevant documents are specified by what is wanted, rather than how to obtain it. Verberne (2018) argues for transparency in personalised professional search using knowledge graphs to create a representation of the concepts the searcher is looking for. Russell-Rose et al., (2019) and Shokraneh (2019) and Russell-Rose (2019) outline a visual approach where a 2D sketchpad creates a canvas where blocks (concepts) can be created and relationships established which mirrors the Boolean constructs, but with a more direct mapping between the visual representation and the intended semantics (see Fig. 2, a visual representation of the search strategy presented in Fig. 1). Other visual methods have been tried before, which are reviewed by Russell-Rose et al. (2019) e.g., Venn diagrams and 'dust and magnet' representations. These methods still rely on Boolean logic, but provide some level of transparency and better support for concept representation and provide a partial solution to the problem, addressing in part the *formalism* barrier e.g., by supporting the separation of search strategies from the platform when recording and sharing it (Russell-Rose and Gooch, 2018).

Approaches 1: Data modelling/theory

The underlying theory used in systematic reviews is Boolean logic, with extensions to support proximity (a special case of AND), truncation/wildcards (a special case of OR) and search fields (based on metadata). There are many potential ways to deal with this, either replacing the logic with something new, or abstracting the details of the logic away and representing it visually (see above). Ranking as an alternative logic could be used, but is not accepted by the community (see above). There are logics such as 'fuzzy' logic, but these have not gained much headway in the IR field generally (Robertson, 1978), and are likely to face the same resistance to implementation as ranking schemes for many of the same reasons. No serious attempt has been made to use different models and theories in the field apart from those reviewed above to the best of our knowledge.

Approaches 2: Explainability

There is little work on explanations for search strategies used in systematic reviews. Much work concentrates on the screening process e.g. Mi & Jiang (2019) and Thomas et al. (2019), in particular the interpretability of those results. This does not really deal with the essential problem in professional search where the issue is earlier in the process. Thomas et al. (2019) focuses on web searches, not professional ones. Apart from this, there does not appear to be much in the way of primary research being done specifically on professional search, apart from Russell-Rose and MacFarlane (2020) and Verberne (2018), that both identify gaps and issues to be addressed.

Summary of tools and approaches for search strategy development

In summary, there have been attempts to address the shortcomings of conventional approaches or augment them. Such attempts rely either on Boolean approaches and share many of their shortcomings, or on ML and/or information extraction on top of standard IR models (e.g., BM25). Some of the methods can be used to supplement Boolean approaches and improve the search strategy development, but do not address the problem of developing the initial query as they only become useful in a second stage for query reformulation from identified relevant or 'gold standard' studies. Some tools can be used to develop an initial query for non-complex topics i.e., Unsilo (2021) and ResearchRabbit (2021), but deploying the tools to address complex topics given the inherent tacit knowledge required in the same way is a long way off, if possible at all.

Table 1
Summary of problems identified and potential solutions, issue type and process stage.

#	Issue	Citation(s)	Solution (Design principle)	Barrier type	Process stages (section 4)
1	Difficulty in identifying and/or articulating strategy structure.	Russell-Rose and Shokraneh (2020).	Support transparency in mapping (P1).	Formalism	Planning/ designing: 1,2,3.
2	Lack of scalability of Boolean strings.	Russell-Rose and Shokraneh (2020).	Adapt scalable mechanisms (P2).	Formalism	Testing/running: 4,5,7.
3	Steep learning curve.	van Altena et al. (2019), Yoo and Mosa (2015).	Increase transparency, reduce complexity (P1-4).	Platform/ community	All: 1-8.
4	Lack of design standards for search strategy development.	Hoang and Schneider (2018).	Help community develop design standards (P1-5).	Community	Planning/ designing: 1,2,3.
5	Costs of development (time, resources, budgets).	Hemmingway and Brereton, (2009), Borah et al. (2017).	Reduce costs by applying all EARS principles, making the process faster (P1-5).	Platform/ community	All: 1-8.
6	Lack of comprehensive development guidelines.	Shokraneh (2018).	Revise in line with design standards (P1-5).	Community	Planning/ designing: 1,2,3.
7	Lack of a central repository for sharing strategies and/or results (standard templates & format).	Koffel and Rethlefsen (2016); Biocic et al. (2019), Shokraneh (2018).	Provide support for collaboration/team working (P5).	Platform/ community	Reporting/ sharing: 6,8.
8	Limited functionality for search management.	Russell-Rose and Chamberlain (2017), van Altena et al. (2019).	Abstractions are required to provide a variety of functions (P2-4).	Platform	All: 1-8.
9	Prevalence of errors in published strategies.	Sampson and McGowan (2006), Franco et al. (2018).	Eliminate errors via abstractions, provide feedback on user queries (P1-4).	Formalism	Testing/running: 4,5,7.
10	Inefficiencies due to system heterogeneity & need for translation.	Bramer et al. (2018).	Standard search strategy model required (P1-4).	Platform	Testing/running: 4,5,7.
11	Lack of replication in automated systems.	O'Mara-Eves et al. (2015)	Increase transparency, abstract complexity out (P1-4).	Formalism/ platform	Testing/running: 4,5,7.
12	Limited compatibility with existing workflows.	van Altena et al. (2019), O'Connor et al. (2019).	User acceptance of design principles (P1-5).	Platform/ community	All.
13	Poor support for query management.	van Altena et al. (2019).	In built support for users e.g., real time feedback on query effectiveness (P4).	Platform	Testing/running: 4,5,7.
14	Lack of usability & accessibility of existing platforms.	van Altena et al. (2019).	Increase transparency, abstract away complexity (P1-4).	Platform	Testing/running: 4,5,7.
15	Lack of documentation for iterative process of designing the search.	Russell-Rose and Shokraneh (2020).	Provide support for collaboration/team working (P5).	Formalism/ platform	Planning/ designing: 1,2,3.
16	Separation of search strategies from platform during recording, reporting, and sharing.	Russell-Rose et al. (2019).	Provide support for collaboration/team working (P5).	Formalism	Reporting/ sharing: 6,8.
17	Out of date and unreliable search strategies (need for living search strategies).	Thomas at al. (2017), Shokraneh and Russell-Rose 2020, Shojania et al. (2007)	Provide support for collaboration/team working (P5).	Formalism/ platform	Reporting/ sharing: 6,8.

There is little work in actually understanding the underlying problem in terms of frameworks which are explainable/transparent or conceptual ideas which would allow the searcher to better understand their information needs. In some cases, structure (such as the use of PICO, SPIDER, or SPICE) is used. There is resistance to the introduction of automated tools that might assist the searchers (van Altena et al., 2019; O'Connor et al., 2019; Yoo & Mosa, 2015), due to lack of compatibility with workflows and steep learning curves. Approaches examined in this section therefore retain many of the same limitations, and many lack transparency and explainability leading to an erosion of trust. Efforts to address the limitations of Boolean methods have been made, but we need new solutions to ensure that search strategies are transparent, explainable and trustworthy, and facilitate reproducibility and replicability. Some efforts are evident in terms of the development of guidelines e.g. 'The Vienna principles' outlined by Beller et al. (2018), but it is clear that design principles for tools also need to be addressed.

Discussion: A framework for search strategy formulation

Boolean search strategies have been used successfully for many years to resolve complex information needs in healthcare and other domains such as legal research, patent search and recruitment. Professional searchers have built up considerable tacit knowledge to ensure that relevant studies are identified for evidence synthesis, but there are still significant barriers regarding reproducibility and replicability. There is some level of transparency and explainability which whilst limited does facilitate system control and hence trust. However, current methods can be hard to learn, difficult to conceptualise, time consuming to maintain and error-prone, with the heterogeneity of systems adding further complexity. Laudable attempts to address this problem have been made, but many rely on AI and ML technologies that also lack transparency and explainability. In addition, many alternative toolsets and approaches are difficult to use, have a steep learning curve and are not easily integrated into the systematic review production workflow. They therefore demonstrate the same drawbacks as conventional methods, but few of the advantages particularly in terms of trust. The community needs to develop new conceptual frameworks that offer a better way to resolve complex structured information needs in the evidence synthesis process. These frameworks need to help the user better articulate the structure of their information needs, to aid explainability and transparency, and to be more scalable to aid reproducibility and replicability. In this paper we have identified many problems with current methods, and propose a framework that sets down key principles to offer solutions to these problems.

This framework is outlined in the following set of design principles and a mapping from the barriers identified above to those principles. Our aspiration is that this framework will inform the development of a new conceptual framework for search strategy formulation (Russell--Rose & MacFarlane, 2020). The five design principles are as follows:

1 Provide support for transparency in the mapping between logical structure and physical structure, using visual representations to communicate conceptual structure and relationships and encourage exploration [P1].
2 Adopt scalable, declarative formalisms that accommodate complexity and support abstraction and encapsulation, e.g., allowing users to switch between overview and detail views and independently manipulate individual query elements [P2].
3 Delegate lower-level syntactic operators to system functions, e.g., replacing error-prone string manipulation with automated translation to different query syntaxes [P3].
4 Provide real time feedback on query effectiveness, allowing users to perceive the contribution of individual query elements and understand how to make queries more effective [P4].
5 Provide support for collaboration and team working, e.g., through repositories of best practice examples that facilitate versioning, sharing and peer review [P5].

Table 1 provides a mapping between the barriers identified in this review (column 2), their source (column 3), potential solutions in the form of the relevant design principles (column 4) and the barrier category (column 5). In the last column we categorise issues according to the stage in the search strategy formulation process to which they apply (as defined above).

In applying these principles it should be noted that searchers' needs are a continuum from those that are relatively simple and machine applicable (can be automated) such as a scoping review, to those that are relatively complex and require significant human input, such as a full systematic review. Finding the right way to apply the principles to these different types of needs is important to engender trust in any implemented system that embodies such principles. The key to gaining trust is to enhance explainability and transparency for search strategy development, which in turn will enhance replicability and reproducibility.

Summary

In this paper we have used the key concepts of reproducibility, replicability, transparency, explainability and trust to examine the problems faced by professional searchers in developing effective search strategies for evidence synthesis. We have focused on healthcare, but other domains such as patents, law and recruitment stand to gain from any progress that can be made in addressing these concepts. The contribution of this paper is a set of design principles and a conceptual framework which the IR and professional search communities can use to address the many problems that have been identified. It is our aspiration to bring these communities together to promote collaboration to address these problems and apply the framework to generate effective solutions to them.

CRediT authorship contribution statement

Andrew MacFarlane: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Tony Russell-Rose:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Farhad Shokraneh:** Conceptualization, Investigation, Methodology, Validation, Visualization, Writing – review & editing.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

ACM. (2020). Artefact review and badging. Retrieved from https://www.acm.org/publications/policies/artifact-review-and-badging-current.

Alharbi, A., & Stevenson, M. (2017). In *Ranking abstracts to identify relevant evidence for systematic reviews: The University of Sheffield's approach to clef eHealth 2017 task2. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017. CEUR Workshop Proceedings, CEUR-WS.org*. Retrieved from http://ceur-ws.org/Vol-1866/paper_97.pdf.

Alharbi, A., Briggs, W., & Stevenson, M. (2018). Retrieving and ranking studies for systematic reviews: University of Sheffield's approach to CLEF eHealth 2018 Task 2. In CEUR Workshop Proceedings (Vol. 2125). CEUR Workshop Proceedings. Retrieved from http://ceur-ws.org/Vol-2125/paper_169.pdf.

Alharbi, A., & Stevenson, M. (2019). Ranking studies for systematic reviews using query adaptation: University of Sheffield's approach to CLEF eHealth 2019 task 2 working notes for CLEF 2019. In Working Notes of CLEF 2019-Conference and Labs of the Evaluation Forum (Vol. 2380). CEUR Workshop Proceedings. Retrieved from http://ceur-ws.org/Vol-2380/paper_185.pdf.

Anagnostou, A., Lagopoulos, A., Tsoumakas, G., & Vlahavas, I. (2017). Hybridranksvm:A cost-effective hybrid ltr approach for document ranking. In: Working Notesof CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland,September 11-14, 2017. CEUR Workshop Proceedings, Retrieved from http://ceur-ws.org/Vol-1866/paper_89.pdf.

Begert, D., Granek, J., Irwin, B., & Brogly, C. (2020). Towards automating systematic reviews on immunization using an advanced natural language processing–based extraction system. *Canada Communicable Disease Report, 46*(6), 174–179.

Beller, E., Clark, J., Tsafnat, G., Adams, C., Diehl, H., Lund, H., Ouzzani, M., Thayer, K., Thomas, J., Turner, T., & Xia, J. (2018). Making progress with the automation of systematic reviews: Principles of the International Collaboration for the Automation of Systematic Reviews (ICASR). *Systematic Reviews, 7*(1), 1–7. pp.

Biocic, M., Fidahic, M., & Puljak, L. (2019). Reproducibility of search strategies of non-Cochrane systematic reviews published in anaesthesiology journals is suboptimal: Primary methodological study. *British Journal of Anaesthesia, 122*(6), e79–e81.

Booth, A. (2010). How much searching is enough? Comprehensive versus optimal retrieval for technology assessments. *International Journal of Technology Assessment In Health Care, 26*(4), 431–435.

Borah, R., Brown, A. W., Capers, P. L., & Kaiser, K. A. (2017). Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ open, 7*(2), Article e012545.

Bramer, W. M., de Jonge, G. B., Rethlefsen, M. L., Mast, F., & Kleijnen, J. (2018). A systematic approach to searching: An efficient and complete method to develop literature searches. *Journal of the Medical Library Association: JMLA, 106*(4), 531–541.

Burri, K. G. (2019). Intelligent use of artificial intelligence for systematic reviews of medical devices. *Medical Writing, 28*, 22–27.

Chalmers, I., Bracken, M. B., Djulbegovic, B., Garattini, S., Grant, J., Gülmezoglu, A. M., Howells, D. W., Ioannidis, J. P., & Oliver, S. (2014). How to increase value and reduce waste when research priorities are set. *The Lancet, 383*(9912), 156–165.

Chen, J., Chen, S., Song, Y., Liu, H., Wang, Y., Hu, Q., & He, L. (2017). Ecnu at 2017ehealth task 2: Technologically assisted reviews in empirical medicine. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ire-land, September 11-14, 2017. CEUR Workshop Proceedings, Retrieved from http://ceur-ws.org/Vol-1866/paper_99.pdf.

Clark, J. M., Beller, E., Glasziou, P., & Sanders, S. (2021). The decisions and processes involved in a systematic search strategy: A hierarchical framework. *Journal of the Medical Library Association: JMLA, 109*(2), 201. p.

Cohen, A. M., & Smalheiser, N. R. (2018). Ohsu clef 2018 task 2 diagnostic test accuracy ranking using publication type cluster similarity measures. In: Working Notes ofCLEF 2018 - Conference and Labs of the Evaluation forum, Avignon, France, September 10-14, 2018. CEUR Workshop Proceedings, Retrieved from http://ceur-ws.org/Vol-2125/paper_174.pdf.

Cooper, C., Booth, A., Varley-Campbell, J., Britten, N., & Garside, R. (2018). Defining the process to literature searching in systematic reviews: A literature review of guidance and supporting studies. *BMC Medical Research Methodology, 18*(1), 85.

Cormack, G. V., & Grossman, M. R. (2016a). Scalability of continuous active learning for reliable high-recall text classification. In *In Proceedings of the 25th ACM International Conference On Information And Knowledge Management* (pp. 1039–1048).

Cormack, G. V., & Grossman, M. R. (2016b). Engineering quality and reliability in technology-assisted review. In *In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 75–84).

Cormack, Gordon V., & Grossman, Maura R. (2017). Technology-assisted review in empirical medicine: Waterloo participation in CLEF eHealth 2017. In *Proceedings of the Conference and Labs of the Evaluation Forum (CLEF'17)*. Retrieved from http://ceur-ws.org/Vol-1866/paper_51.pdf.

Cormack, G. V., & Grossman, M. R. (2018). Technology-assisted review in empirical medicine: waterloo participation in CLEF eHealth 2018. In CLEF (Working Notes). Retrieved from http://ceur-ws.org/Vol-2125/paper_89.pdf.

Del Fiol, G., Michelson, M., Iorio, A., Cotoi, C., & Haynes, R. B. (2018). A deep learning method to automatically identify reports of scientifically rigorous clinical research from the biomedical literature: Comparative analytic study. *Journal of Medical Internet Research, 20*(6), e10281. p.

Di Nunzio, G. M., Beghini, F., Vezzani, F., & Henrot, G. (2017). An Interactive two-dimensional approach to query aspects rewriting in systematic reviews. IMS Unipd At CLEF eHealth Task 2. In CLEF (working notes).

Di Nunzio, G.M., Ciuffreda, G., & Vezzani, F. (2018a). Interactive Sampling for Systematic Reviews. IMS Unipd At CLEF 2018 eHealth Task 2. In CLEF (working notes).

Di Nunzio, G. M., Maistro, M., & Vezzani, F. (2018b). A gamified approach to NaïVe Bayes classification: A case study for newswires and systematic medical reviews. In *In Companion proceedings of the web conference 2018* (pp. 1139–1146).

Di Nunzio, G. M. (2018). A study of an automatic stopping strategy for technologically assisted medical reviews. In *In European Conference on Information Retrieval* (pp. 672–677). Cham: Springer.

Di Nunzio, G. M. (2019). A Distributed Effort Approach for Systematic Reviews. IMS Unipd at CLEF 2019 eHealth Task 2. In Clef (working notes). Retrieved from http://ceur-ws.org/Vol-2380/paper_205.pdf.

Di Nunzio, G. M. (2020). A Study on a Stopping Strategy for Systematic reviews based on a distributed effort approach. In *International Conference of the Cross-Language Evaluation Forum for European Languages* (pp. 112–123). Cham: Springer.

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. arXiv preprint arXiv:1710.00794. Retrieved from https://arxiv.org/pdf/1710.00794.pdf.

Ferro, N., Fuhr, N., Maistro, M., Sakai, T., & Soboroff, I. (2019). CENTRE@ CLEF2019: Overview of the replicability and reproducibility tasks. *CLEF (Working Notes)*.. Retrieved from http://ceur-ws.org/Vol-2380/paper_258.pdf.

Francis, D. K., Smith, J., Saljuqi, T., & Watling, R. M. (2015). Oral protein calorie supplementation for children with chronic disease. *Cochrane Database of Systematic Reviews*, (5). Retrieved from https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD001914.pub2/full.

Franco, J. V. A., Garrote, V. L., Escobar Liquitay, C. M., & Vietto, V. (2018). Identification of problems in search strategies in Cochrane Reviews. *Research Synthesis Methods, 9*(3), 408–416.

Goeuriot, L., Kelly, L., Suominen, H., Névéol, A., Robert, A., Kanoulas, E., Spijker, R., Palotti, J., & Zuccon, G. (2017). CLEF 2017 eHealth evaluation lab overview. In *International Conference of the Cross-Language Evaluation Forum for European languages* (pp. 291–303). Cham: Springer.

Grames, E. M., Stillman, A. N., Tingley, M. W., & Elphick, C. S. (2019). An automated approach to identifying search terms for systematic reviews using keyword co-occurrence networks. *Methods in Ecology and Evolution, 10*(10), 1645–1654.

Hausner, E., Waffenschmidt, S., Kaiser, T., & Simon, M. (2012). Routine development of objectively derived search strategies. *SYSTEMATIC reviews, 1*(19), 1–10. Retrieved from https://link.springer.com/article/10.1186/2046-4053-1-19.

Hemingway, P., & Brereton, N. (2009). What is a systematic review? *Hayward Medical Communications, 2*, 1–8.

Higgins, J., Thomas, J., Chandler, J., Cumpson, M., Li, T., Page, M., & Welch, V. (2020). *Cochrane handbook for systematic reviews of interventions*. Retrieved from https://training.cochrane.org/handbook/current.

Hjørland, B. (2015). Classical databases and knowledge organization: A case for boolean retrieval and human decision-making during searches. *Journal of the Association for Information Science and Technology, 66*(8), 1559–1575.

Hoang, L., & Schneider, J. (2018). Opportunities for computer support for systematic reviewing-a gap analysis. In: G.Chowdhury, J. McLeod, V. Gillet, & P. Willett. (Eds). In *, 10766. Transforming Digital Worlds. iConference 2018. Lecture Notes in Computer Science* (pp. 367–377). Springer. Cham.

Hollmann, N., & Eickhoff, C. (2017). Relevance-based stopping for recall-centric medical document retrieval. In *In: Working Notes of CLEF 2017 - Conference and Labs ofthe Evaluation forum, Dublin, Ireland, September 11-14, 2017. CEUR Workshop Proceedings*. CEUR-WS.org.

Kanoulas, E., Li, D., Azzopardi, L., & Spijker, R. (2017). CLEF 2017 technologically assisted reviews in empirical medicine overview. In *In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, Dublin, Ireland, 11–14 September 2017*. CEUR-WS.org.

Kanoulas, E., Li, D., Azzopardi, L., & Spijker, R. (2018). CLEF 2018 technologically assisted reviews in empirical medicine overview. In *In: Cappellato, L., Ferro, N., Nie, J.-Y., Soulier, L., eds. CEUR Workshop Proceedings of Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, 10–14 September , vol. 2125*. CEUR-WS.org.

Kanoulas, E., Li, D., Azzopardi, L., & Spijker, R. (2019). CLEF 2019 technology assisted reviews in empirical medicine overview. In *In: CEUR Workshop Proceedings of CLEF Evaluation Labs andWorkshop: OnlineWorking Notes*. CEUR-WS.org (2019).

Karimi, S., Pohl, S., Scholer, F., Cavedon, L., & Zobel, J. (2010). Boolean versus ranked querying for biomedical systematic reviews. *BMC medical informatics and decision making, 10*(1), 58. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-10-58. Retrieved from.

Kalphov, V., Georgiadis, G., & Azzopardi, L., 2017, September. SiS at CLEF 2017 eHealth TAR task. In CEUR workshop proceedings (Vol. 1866, pp. 1–5).

Kelly, L., Suominen, H., Goeuriot, L., Neves, M., Kanoulas, E., Li, D., Azzopardi, L., Spijker, R., Zuccon, G., Scells, H., & Palotti, J. (2019). Overview of the CLEF eHealth evaluation lab 2019. In *International Conference of the Cross-Language Evaluation Forum for European Languages* (pp. 322–339). Cham: Springer.

Kim, Y., Seo, J., & Croft, W. B. (2011). Automatic Boolean query suggestion for professional search. In R. Baeza-Yates, T. Chua & W.B. Croft (Eds.). In *SIGIR 2011: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp.825-834). New York: ACM*.

Koffel, J. B., & Rethlefsen, M. L. (2016). Reproducibility of search strategies is poor in systematic reviews published in high-impact pediatrics, cardiology and surgery journals: A cross-sectional study. *PLoS One, 11*(9), Article e0163309.

Lau, J. (2019). Systematic review automation thematic series. *Systematic Reviews, 8*(70). Retrieved from https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-019-0974-z.

Lefebvre, C., Glanville, J., Briscoe, S., Littlewood, A., Marshall, C., Metzendorf, M. I., Noel-Storr, A., Rader, T., Shokraneh, F., Thomas, J., & Wieland, L. S. (2020). Searching for and selecting studies. *Cochrane Handbook for systematic reviews of interventions* (pp. 67–107). Retrieved from https://training.cochrane.org/handbook/current/chapter-04.

Lee, G. E. (2017). Medical document classification for systematic reviews using convolutional neural networks: Sysreview at clef ehealth 2017. In *In: Working Notes of CLEF2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September11-14, 2017. CEUR Workshop Proceedings, CEUR-WS.org*.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics, 36*(4), 1234–1240.

Li, M., Xu, C., Wei, T., Bao, D., Lu, N., & Yang, J. (2018). ECNU at 2018 eHealth Task1 multilingual information extraction. In CLEF (Working Notes). Retrieved from http://ceur-ws.org/Vol-1866/paper_99.pdf.

Li, D., & Kanoulas, E. (2019). Automatic thresholding by sampling documents and estimating recall. In: Cappellato, L., Ferro, N., Losada, D.E., Muller, H. (eds.). In *CEUR Workshop Proceedings of Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, 9–12 September 2019, vol. 2380*.

Li, D., & Kanoulas, E. (2020). When to stop reviewing in technology-assisted reviews: Sampling from an adaptive distribution to estimate residual relevant documents. *ACM Transactions on Information Systems (TOIS), 38*(4), 1–36.

Li, D., Zafeiriadis, P., & Kanoulas, E. (2020). APS: An active PubMed search system for technology assisted reviews. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 2137–2140).

MacFarlane, A., & Russell-Rose, T. (2016). Search Strategy Formulation: A Framework For Learning. In I. Diaz, J. Fernandez-Luna (Eds.). In *CERI'16 - Proceedings of the 4th Spanish Conference on Information Retrieval. (pp.1-8). New York: ACM*.

Markey, K., & Cochrane, P. (1981). *Online training and practice manual for ERIC database searchers (2nd Edition)*. ERIC Clearinghouse on Information Resources, Syracuse University. Retrieved from https://files.eric.ed.gov/fulltext/ED212296.pdf.

Marshall, I. J., & Wallace, B. C. (2019). Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. *Systematic Reviews, 8*(1). Retrieved from https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-019-1074-9.

McGowan, J., & Sampson, M. (2005). Systematic reviews need systematic searchers. *Journal of the Medical Library Association, 93*(1). Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC545125/.

McGowan, J., Sampson, M., Salzwedel, D. M., Cogo, E., Foerster, V., & Lefebvre, C. (2016). PRESS peer review of electronic search strategies: 2015 guideline statement. *Journal of Clinical Epidemiology, 75*, 40–46.

Minas, A., Lagopoulos, A., & Tsoumakas, G. (2018). Aristotle University's approach to the technologically assisted reviews in empirical medicine task of the 2018 CLEF eHealth Lab. In CLEF (Working Notes). Retrieved from http://ceur-ws.org/Vol-2125/paper_179.pdf.

Norman, C., Leeflang, M., & Neveol, A. (2017). In *Limsi@clef ehealth 2017 task 2: Logistic regression for automatic article ranking. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017.CEUR Workshop Proceedings*. Retrieved from http://ceur-ws.org/Vol-1866/paper_101.pdf.

Norman, C., Leeflang, M., & Neveol, A. (2018). In *Limsi@clef ehealth 2018 task 2: Technology assisted reviews by stacking active and static learning. In: Working Notes of CLEF2018 - Conference and Labs of the Evaluation forum, Avignon, France, September10-14, 2018. CEUR Workshop Proceedings*. Retrieved from http://ceur-ws.org/Vol-2125/paper_198.pdf.

O'Connor, A. M., Tsafnat, G., Thomas, J., Glasziou, P., Gilbert, S. B., & Hutton, B. (2019). A question of trust: Can we build an evidence base to gain trust in systematic review automation technologies? *Systematic Reviews, 8*(1), 143. Retrieved from https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-019-1062-0.

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. Retrieved from https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/2046-4053-4-5.

O'Sullivan, D. M., Wilk, S. A., Michalowski, W. J., & Farion, K. J. (2010). Automatic indexing and retrieval of encounter-specific evidence for point-of-care support. *Journal of Biomedical Informatics, 43*(4), 623–631.

O'Sullivan, D., Wilk, S., Michalowski, W., & Farion, K. (2013). Using PICO to align medical evidence with MDs decision making models. In C.U. Lehman, E. Ammenwerth & C. Nohr (Eds.). In *MEDINFO 2013: Proceedings of the 16th World Congress on Medical Informatics. Amsterdam: IMIA and IOS Press*.

Olorisade, B. K., Brereton, P., & Andras, P. (2017). Reproducibility of studies on text mining for citation screening in systematic reviews: Evaluation and checklist. *Journal of Biomedical Informatics, 73*, 1–13.

Paisley, W. J., & Parker, E. B. (1965). Information retrieval as a receiver-controlled communication system. In Heilprin, L.B. (Ed.). In *Proceedings of the Symposium on Education for Information Science* (pp. 23–31).

Plesser, H. E. (2018). Reproducibility vs. replicability: A brief history of a confused terminology. *Frontiers in Neuroinformatics, 11*, 76. Retrieved from https://www.frontiersin.org/articles/10.3389/fninf.2017.00076/full.

Research Rabbit (2021). ResearchRabbit web site. Retrieved from https://www.researchrabbit.ai/.

Robertson, S. E. (1978). On the nature of fuzz: A diatribe. *Journal of the American Society for Information Science, 29*(6), 304–307.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., & Gatford, M. (1995). Okapi at TREC-3. *Nist Special Publication Sp, 109*, 109.

Russell-Rose, T. (2019). Rethinking 'Advanced Search': An AI-based approach to search strategy formulation. In D. Haynes (Ed). In *The Human Position in an Artificial World: Creativity, Ethics and AI in Knowledge Organization (pp.275-290). Würzburg: Ergon-Verlag*.

Russell-Rose, T., & Chamberlain, J. (2017). Expert search strategies: The information retrieval practices of healthcare information professionals. *JMIR Medical Informatics, 5*(4), e33. Retrieved from https://medinform.jmir.org/2017/4/e33/?utm_source=TrendMD&utm_medium=cpc&utm_campaign=JMIR_TrendMD_0.

Russell-Rose, T., Chamberlain, J., & Azzopardi, L. (2018). Information retrieval in the workplace: A comparison of professional search practices. *Information Processing & Management, 54*(6), 1042–1057.

Russell-Rose, T., Chamberlain, J., & Shokraneh, F. (2019). A visual approach to query formulation for systematic search. In H. Joho, V. Murdock, & P. Qvarfordt (Eds.). In *CHIIR 2019: Proceedings of the 2019 Conference on Human Information Interaction and Retrieval* (pp. 379–383). New York: ACM.

Russell-Rose, T., & Gooch, P. (2018). 2dSearch: A visual approach to search strategy formulation. In *Proceedings of the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems*.Retrieved from http://ceur-ws.org/Vol-2167/paper8.pdf.

Russell-Rose, T., & Shokraneh, F. (2020). Designing the structured search experience: rethinking the query-builder paradigm. *Weave: Journal of Library User Experience, 3*(1).

Russell-Rose, T., & MacFarlane, A. (2020). Toward explainability in professional search. In *Proceedings of the 3rd International Workshop on Explainable Search (EARS 2020). Retrieved from*. https://eprints.goldsmiths.ac.uk/29134/1/EARS%202020.pdf.

Russell-Rose, T., Gooch, P., & Kruschwitz, U. (2021). Interactive query expansion for professional search applications. *Business Information Review*, 1–11.

Salvador-Oliván, J. A., Marco-Cuenca, G., & Arquero-Avilés, R. (2019). Errors in search strategies used in systematic reviews and their effects on information retrieval. *Journal of the Medical Library Association: JMLA, 107*(2), 210. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6466507/.

Sampson, M., & McGowan, J. (2006). Errors in search strategies were identified by type and frequency. *Journal of Clinical Epidemiology, 59*(10), 1057–10e1. Retrieved from https://www.sciencedirect.com/science/article/pii/S0895435606000436.

Scells, H., Zuccon, G., Deacon, A., & Koopman, B. (2017). QUT ielab at CLEF eHealth 2017 technology assisted reviews track: Initial experiments with learning to rank. In *In Working Notes of CLEF 2017-Conference and Labs of the Evaluation Forum [CEUR Workshop Proceedings, Volume 1866] (pp. 1-6). Sun SITE Central Europe*.

Scells, H., & Zuccon, G. (2018a). Generating better queries for systematic reviews. In B. Davidson, Y. Liu & E. Yilmaz (Eds.), SIGIR 2018. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (pp.475-484). New York: ACM*.

Scells, H., & Zuccon, G. (2018b). Searchrefiner: A query visualisation and understanding tool for systematic reviews. In J. Allan & N. Paton (Eds.), CIKM 2018. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (pp.1939-1942). New York: ACM*.

Scells, H., Azzopardi, L., Zuccon, G., & Koopman, B. (2018). Query variation performance prediction for systematic reviews. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 1089–1092).

Scells, H., Zuccon, G., & Koopman, B. (2019). Automatic boolean query refinement for systematic review literature search. In L. Liu, R. White (Eds.). In *WWW'19: Proceedings of The World Wide Web Conference* (pp. 1646–1656). New York: ACM.

Scells, H., Zuccon, G., Koopman, B., & Clark, J. (2020a). A computational approach for objectively derived systematic review search strategies. In: J.M. Jose, E, Ylimaz, J. Magalhaes, P. Castrells, N. Ferro, M.J. Silva, & F. Martins (Eds.). In *ECIR 2020: European Conference on Information Retrieval* (pp. 385–398). Cham: Springer.

Scells, H., Zuccon, G., Sharaf, M. A., & Koopman, B. (2020b). Sampling query variations for learning to rank to improve automatic boolean query generation in systematic reviews. In Y. Huang, I. King T. Liu M. van Steen (Eds.). In *Proceedings of The Web Conference 2020* (pp. 3041–3048). New York: ACM.

Scells, H., Zuccon, G., Koopman, B., & Clark, J. (2020c). Automatic boolean query formulation for systematic review literature search. In Y. Huang, I. King T. Liu M. van Steen (Eds.). In *Proceedings of The Web Conference 2020* (pp. 1071–1081). New York: ACM.

Shokraneh, F. (2016). PICO framework: Two decades of variation and application. Evidence Live. Retrieved from https://10.13140/RG.2.2.35019.23841.

Shokraneh, F. (2018). Evidence-based searching and reporting the search for systematic reviews. BrightTALK, 22 Aug 2018. Retrieved from https://www.brighttalk.com/webcast/16527/328864.

Shokraneh, F. (2019). Reproducibility and replicability of systematic reviews. *World Journal of Meta-Analysis, 7*(3), 66–71.

Shokraneh, F., & Russell-Rose, T. (2020). Lessons from COVID-19 to future evidence synthesis efforts: First living search strategy and out of date scientific publishing and indexing industry (submitted). *Journal of Clinical Epidemiology, 123*, 171–173. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7184009/.

Shojania, K. G., Sampson, M., Ansari, M,T., Ji, J., Doucette, S., & Moher, D. (2007). How quickly do systematic reviews go out of date? *A Survival Analysis. Annals of Internal Medicine, 147*(4), 224–233.

Singh, G., Marshall, I., Thomas, J., & Wallace, B. (2017). Identifying diagnostic test accuracy publications using a deep model. In *In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017. CEUR Workshop Proceedings*. CEUR-WS.org.

Sloan, M., & Wang, J. (2015). Dynamic information retrieval: Theoretical framework and application. In *In Proceedings of the 2015 International Conference on the theory of Information Retrieval* (pp. 61–70).

Suominen, H., Kelly, L., Goeuriot, L., Névéol, A., Ramadier, L., Robert, A., Kanoulas, E., Spijker, R., Azzopardi, L., Li, D., & Palotti, J. (2018). Overview of the CLEF eHealth evaluation lab 2018. In *In International Conference of the Cross-Language Evaluation Forum for European Languages* (pp. 286–301). Cham: Springer.

Thomas, J., McNaught, J., & Ananiadou, S. (2011). Applications of text mining within systematic reviews. *Research Synthesis Methods, 2*(1), 1–14.

Thomas, J., Noel-Storr, A., Marshall, I., Wallace, B., McDonald, S., Mavergames, C., Glasziou, P., Shemilt, I., Synnot, A., Turner, T., & Elliott, J. (2017). Living systematic reviews: 2. Combining human and machine effort. *Journal of Clinical Epidemiology, 91*, 31–37.

Thomas, P., Billerbeck, B., Craswell, N., & White, R. W. (2019). Investigating searchers' mental models to inform search explanations. *ACM Transactions on Information Systems (TOIS), 38*(1), 1–25.

Tsafnat, G., Glasziou, P., Choong, M. K., Dunn, A., Galgani, F., & Coiera, E. (2014). Systematic review automation technologies. *Systematic Reviews, 3*(1), 74. Retrieved from https://link.springer.com/article/10.1186/2046-4053-3-74.

Unsilo (2021). Unsilo web site. Retrieved from https://unsilo.ai/unsilo-manuscript-evaluation/.

van Altena, A. J., & Olabarriaga, S. D. (2017). Predicting publication inclusion for diagnostic accuracy test re-views using random forests and topic modelling. In *In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017. CEUR Workshop Proceedings*. Retrieved from http://ceur-ws.org/Vol-1866/paper_115.pdf.

van Altena, A. J., Spijker, R., & Olabarriaga, S. D. (2019). Usage of automation tools in systematic reviews. *Research Synthesis Methods, 10*(1), 72–82. Retrieved from https://onlinelibrary.wiley.com/doi/pdf/10.1002/jrsm.1335.

Verberne, S. (2018). Explainable IR for personalizing professional search. In ProfS/KG4IR/Data: Search@ SIGIR (pp.35-42). Retrieved from https://pdfs.semanticscholar.org/3bb2/a5e883bb17563f763055e27837e19ae01761.pdf.

Wallace, B. C., Trikalinos, T. A., Lau, J., Brodley, C., & Schmid, C. H. (2010). Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics, 11*(1), 1–11.

Wallace, B. C., Dahabreh, I. J., Moran, K. H., Brodley, C. E., & Trikalinos, T. A. (2013). Active literature discovery for scoping evidence reviews: How many needles are there. In *In KDD Workshop On Data Mining For Healthcare (KDD-DMH)*.

Wallace, B. C., Kuiper, J., Sharma, A., Zhu, M., & Marshall, I. J. (2016). Extracting PICO sentences from clinical trial reports using supervised distant supervision. *The Journal of Machine Learning Research, 17*(1), 4572–4596.

Wu, H., Wang, T., Chen, J., Chen, S., Hu, Q., & He, L. (2018). Ecnu at 2018 ehealth task 2:TEchnologically assisted reviews in empirical medicine. In *In: Working Notes of CLEF2018 - Conference and Labs of the Evaluation forum, Avignon, France, September10-14, 2018*. CEUR Workshop Proceedings, Retrieved from http://ceur-ws.org/Vol-2125/paper_177.pdf.

Yale (2021). Yale Mesh analyser. http://mesh.med.yale.edu/.

Yoo, I., & Mosa, A. S. M. (2015). Analysis of PubMed user sessions using a full-day PubMed Query Log: A comparison of experienced and nonexperienced PubMed users. *JMIR Medical Informatics, 3*(3), e25. Retrieved from https://medinform.jmir.org/2015/3/e25.

Zou, J., Li, D., & Kanoulas, E. (2018). Technology assisted reviews: Finding the last few relevant documents by asking yes/no questions to reviewers. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (pp. 949–952).

Zou, J., & Kanoulas, E. (2020). Towards question-based high-recall information retrieval: Locating the last few relevant documents for technology-assisted reviews. *ACM Transactions on Information Systems (TOIS), 38*(3), 1–35.