CADTH

Text Mining Opportunities: White Paper

University of York

INVESTORS IN PEOPLE

YHEC
York Health Economics Consortium

CADTH

Authors: Julie Glanville, Hannah Wood

Cite As: *Text Mining Opportunities: White Paper*. Ottawa: CADTH; 2018 May.

Acknowledgments: The authors are grateful to David Kaunelis for critically reviewing drafts of the report and for providing editorial support. We are grateful to Quertle and to the Evidence for Policy and Practice Information (EPPI)-Centre for providing access to their resources free of charge to allow us to evaluate them.

About CADTH: CADTH is an independent, not-for-profit organization responsible for providing Canada's health care decision-makers with objective evidence to help make informed decisions about the optimal use of drugs, medical devices, diagnostics, and procedures in our health care system.

Funding: CADTH receives funding from Canada's federal, provincial, and territorial governments, with the exception of Quebec.

CADTH

Table of Contents

Figures

Tables

Section 1: Introduction and Methods

Information specialists undertake a wide variety of literature searches to inform rapid reviews, systematic reviews, health technology assessments (HTAs) and economic evaluations. Searches can range from extensive sensitive searches to inform Health Technology Assessments (HTAs) to more focused and pragmatic rapid searches for products with shorter timeframes. Information specialists also develop search filters and search strings to capture common or frequently recurring topics, issues, and themes. There are many challenges for information specialists when producing efficient search strategies, particularly when time is short and/or search topics are complex or vocabulary is vague.

In recent years, information specialists have identified that text mining applications (TMAs) may offer opportunities to introduce efficiencies into some information retrieval tasks. For example, TMAs can analyze bibliographic citations and provide data about the terms and concepts within those citations, which might help with strategy development, or may be able to automate and/or speed up search strategy development. There is also the perception that TMAs offer objectivity as well as speed because a computer is processing the citations rather than a person. Additionally, TMAs might help with identifying concepts involved in less straightforward searches, such as those to find devices.

Text mining covers a wide variety of techniques that involve using computers to analyze words and their relationships within text. Text mining can range from simple counts of the number of times that words appear in texts (frequency analysis), to machine learning that can distinguish texts by content following a training exercise, and even to semantic analyses that can analyze words according to their meaning within texts.

This white paper explores a range of TMAs to identify whether there are any practical, ready-to-use tools that might help information specialists now with their literature searching tasks. The TMAs have been assessed with a focus on specific health technology assessment products and stages within projects where possible. This white paper also provides some insights into the challenges of using more sophisticated TMAs in daily information retrieval practice. We have been able to test out EPPI-Reviewer 4, which is a commercial systematic review software package that contains text mining features. We present this as one example of how text mining might be integrated into the HTA process and also as an example of one of many commercial text mining packages that are available. These commercial TMAs are likely to offer a more fully featured experience than some of the ready-to-use TMAs we describe. We also make recommendations for future approaches that information specialists may wish to consider.

This white paper has been informed by two reviews of the literature[1,2] and a marketing survey.[3] It is also informed by the authors' experiences as information specialists conducting text mining to inform literature searches and to train other information specialists on using text mining to support literature searches for systematic reviews and technology assessments.

The tools listed in Paynter et al.'s[2] report Appendix E and other tools have all been rapidly assessed for relevance to the needs of information specialists. Text mining is a fast-moving area with many well-established tools, but is also characterized by the availability of many demonstration TMAs that can appear and disappear rapidly. Some quite promising TMAs are PhD projects that are published but then are not maintained or developed further. Many TMAs also use Java, which often requires installation since it is no longer available in all browsers. Finding stable, free TMAs is a challenge. A list of the resources that we have assessed but have not described in detail in this report (because we think other tools achieve the tasks better) can be found in Appendix C.

Section 2: What Is Text Mining?

The UK's National Centre for Text Mining defines text mining as "the process of discovering and extracting knowledge from unstructured data" through identifying relevant texts, identifying and extracting "entities, facts and relationships between them" from the texts, and "data mining to find associations among the pieces of information extracted from many different texts."[4]

This definition highlights some key differences between the way information specialists view their searches and the texts they retrieve from their searches, and the way TMAs are often designed and used. These differences can be characterized as follows:

- unit of analysis (bibliographic records versus texts)
- search structure (Boolean versus non-Boolean)
- form of analysis (coded text versus uncoded text).

We will look at each issue in turn, as these issues go some way in explaining why some TMAs are directly usable by information specialists and others are not.

2.1 The Unit of Analysis

Information specialists focus on the retrieval of bibliographic records or otherwise structured records (such as trial register records) from databases, and, to a lesser extent, the identification of documents such as reports from the Internet. Information specialists have a granular view of the documents they are analyzing to create search strategies and their analysis of document relevance is at the database-record level. Information specialists design search strategies to find database records.

Text mining views any piece of text, of whatever size, as a document, and many TMAs do not distinguish between types of documents. This means that many TMAs are not designed to easily load and process the results of database searches into individual records. Many TMAs would process a file of MEDLINE records, for example, as a single document. This means that developing strategies that focus on retrieving records can be problematic with some TMAs, and some will require the development of import filters to import files of bibliographic records and save each record as an individual document if record-by-record analysis is required. Even with a record view, the TMA may not have options to further distinguish between titles or abstracts and subject heading fields. This relative "blindness" to structure means that some of the valuable features of bibliographic database records that assist with search efficiency, such as publication type or language fields, may not be leveraged in a TMA. Indeed, additional field labels and extraneous text such as address information can add noise to text mining analyses. Often we need to remove the unwanted data from a set of records before loading into TMAs. Where such pre-processing is needed, the time savings anticipated from using a TMA may not be realized. This may be because of the time needed to process records through a separate program, such as EndNote, to export only selected fields to load into a TMA.

One of the factors that has informed our recommendations on easy-to-use TMAs in this report is whether the TMA can easily load and analyze the results of bibliographic database searches.

2.2 The Search Paradigm

Information specialists are experienced at developing search strategies that optimize searches using Boolean operators (and to a lesser extent proximity operators, a subset of the “AND” operator), and typically they expect to be able to use TMAs to inform these searches. Boolean searching is a dichotomous approach to retrieval. Records are retrieved or not by a set of search terms, without any “fuzziness” about records possibly being retrievable. TMAs, however, are based on a variety of approaches, many of which are non-Boolean. The retrieval of documents is determined by algorithms using formulas based on the frequency of occurrence of words and their collocation with other words, and by the probability that words in collocation are meaningfully related. The choice of search terms and their combination may be informed by cut-off values and other decisions that are not typically required in the world of Boolean operator–based interfaces. The retrieval of a record may be determined by a probability calculation, and retrieval relevance is not fixed or dichotomous. When using semantic analysis, TMAs are even further removed from the Boolean experience, since record selection might be informed by the meaning of words within sentences rather than their simple presence or absence within records.

Experiencing this essential difference between database searching using a Boolean interface and document searching using TMAs can be challenging. We might use TMAs to analyze the frequency of words within a set of records, but then deciding which terms resulting from that analysis should be fed back into our Boolean searches might require us to choose cut-off values or make selections based on algorithms that we may not understand or even see. This can feel uncomfortable for an experienced information specialist.

Most TMAs are designed to analyze documents and to help us prioritize the most relevant to our query according to the rules that we program into them. This means that the TMA designer may not have intended the application to be used as a tool to help with other processes, such as searching databases: the TMA might have been designed to be the search environment. Thus, TMAs may neither be optimized to help information specialists in the ways we might hope nor often designed with information specialists in mind. One use of TMAs could be as the final repository and search workbench for the results of many sensitive searches from a range of databases, rather than as the means to inform more precise searches of the original databases.

Often, using TMAs in an optimal way will require expert help to navigate sophisticated software that is not designed for lay users, and/or to apply it to search-related activities. Such experts do not necessarily have a deep understanding of the requirements of information specialists and the issues involved in the retrieval of bibliographic records for systematic reviews and related products. Therefore, significant communication and teamwork may be required to integrate TMAs into information retrieval practice.

2.3 The Form of Analysis

Effective and efficient information retrieval from a bibliographic database requires a detailed knowledge of the database structure and content, and the database’s interface options. The implementation of this knowledge can be seen as a form of programming in that it involves the selection of options such as limits, truncation lengths, and the correct use of operators, as well as the choice of search terms and their combinations into a strategy.

TMAs also require user involvement to ensure that they are working optimally. However, the knowledge required is in terms of the choice of settings for the text analyses, rather than knowledge of the database structure. TMAs are often blind to the document structure, although making use of the document structure from bibliographic records should be an area that text mining can operationalize usefully. The settings that need to be selected in TMAs might include choices about how to analyze the text, how to find terms in collocation, and whether to analyze terms by their role within sentences. These settings are determined by statistics and linguistics. These can all be challenging issues for searchers who do not know about statistics or linguistics in the context of text analysis. The more sophisticated TMAs require expert operators.

TMAs often involve far more human interaction than anticipated. Even though frequency analysis can be relatively low input in terms of expertise, and machine learning can be highly interactive and then tail off as the machine learns to recognize relevant records, the more sophisticated TMAs require substantial user interaction in terms of building work flows, making coding decisions, and analyzing and testing results of decisions.

2.4 Choosing Text Mining Applications

Section 3 looks at a range of relatively easy-to-use, off-the-shelf TMAs. Information specialists may not encounter so many of the issues described in this section when using these TMAs. Section 4 looks at some of the options for using more sophisticated and fully featured TMAs, which are subject to the issues described in this section.

Section 3: Can Easy-To-Use Text Mining Applications Help With Information Retrieval Tasks?

There are a range of information specialists' tasks with which TMAs might help. These have been grouped into the following broad categories:

- term and phrase selection
- search development for vague topics
- concept identification
- relevance ranking to assist with search refinement
- filter development
- peer review.

Although some of these topics inevitably overlap, we have addressed each of these categories individually to assess how much easy-to-use or off-the-shelf TMAs might help with these tasks.

3.1 Term and Phrase Identification

Information specialists anticipate that TMAs could analyze bibliographic citations and present the results of the citation analyses in ways that would be useful to speed up the search process, and possibly to help automate the search process. There is also a perception that TMAs could provide more objective search term selection and could help with identifying subject headings such as EMTREE.

Single-term and phrase identification are tasks that TMAs can support, particularly when identifying words and subject headings in MEDLINE records. Easy-to-use and reliable tools to identify terms in MEDLINE and other databases are shown in Table 1. In this task, TMAs have typically been adapted to the needs of information specialists and can operate on a record-by-record basis, rather than treating search results as a single document. These TMAs typically analyze term frequency in records, and offer lists of terms usually ranging from the most frequently occurring to the least frequently occurring. The TMAs may not, however, always distinguish terms in the title or abstract from those in the subject headings, but may treat them all as single-word terms.

The TMAs that interface to PubMed (such as PubMed PubReMiner) offer the facility to run simple or detailed searches on PubMed and produce an analysis of the frequency of occurrence of single words in the records retrieved. The results will only be as helpful as the search entered, but the usefulness of these TMAs lies in their rapid highlighting of frequent terms, acronyms, and synonyms with total objectivity. The analysis of hundreds of records can take a matter of minutes and offers a more detailed and reliable approach than information specialists' traditional methods of developing searches — screening search results for new terms by eye. Most of the TMAs we have listed will offer a table of results that can be used for discussion with team members and to keep a record of terms reviewed and selected. The TMA results can also help with identifying where to truncate terms.

A variant of these record analysis tools is HelioBLAST (helioblast.heliotext.com). HelioBLAST is a free Internet service that can be used to find PubMed records that are similar to a query. It is best to identify the title and abstract of a highly relevant record and paste it into HelioBLAST. HelioBLAST then analyzes the record and identifies

records in PubMed that are similar to the content of the single abstract entered. As well as providing a list of 50 best PubMed records and displaying a relevance score, it also provides a list of "implicit" keywords. These keywords help to identify concepts that were not originally mentioned in the query abstract, and can provide additional keywords to add to a search. If this process is repeated with a series of relevant abstracts, additional terms might be identified or reassurance provided that no new suggestions were being identified. The MEDLINE analysis is free of charge, but HelioBLAST offers analysis of other databases for a fee.

The AntConc package processes text from any database (laurenceanthony.net/software/antconc/). It is best to process records through Reference Manager or EndNote first, to export only the fields of interest to be analyzed by AntConc. To test out AntConc, we loaded a set of Embase records into EndNote, exported the titles and abstracts out of EndNote into a file, and then loaded that file into AntConc to produce word lists.

SIDER (sideeffects.embl.de) is a very specific but useful tool for identifying side effects by drug, since side effects can be difficult to know in advance.

There are a variety of easy-to-use TMAs that will provide suggestions for MeSH terms to use in searches, and will analyze the MeSH terms in records to help us with developing the MeSH terms in our searches. The simplest to use are probably MeSH on Demand (www.nlm.nih.gov/mesh/MeSHonDemand.html), for suggesting MeSH terms, and the Yale MeSH Analyzer (mesh.med.yale.edu), which shows a table of the MeSH terms that have been assigned to a set of relevant papers identified by entering the PubMed identifiers.

To explore subject headings in databases beyond PubMed/MEDLINE, EndNote offers a very straightforward analysis of subject headings that are either single terms or compound terms (details in Appendix A). The word list option in AntConc can be used to achieve this as well, but since it analyzes words as single terms, subject headings must be preprocessed in EndNote or elsewhere to replace spaces between terms with symbols such as hyphens, and therefore achieve an accurate analysis. This degree of pre-processing suggests that EndNote may be the easiest approach for analyzing the frequencies of subject headings.

Table 1: Text Mining Applications to Assist With Term and Phrase Identification

	PubMed	Embase and Other Bibliographic Software
Identifying frequently occurring terms in the title and abstract	Record by record: • PubMed PubReMiner • EndNote • EPPI-Reviewer 4 By document (sets of records): • AntConc • Voyant	Record by record: • EPPI-Reviewer 4 By document (sets of records): • AntConc • Voyant
Identifying further relevant records from relevant text	• HelioBLAST • EPPI-Reviewer 4	• Other databases can be searched by HelioBLAST on application. • EPPI-Reviewer 4
Identifying frequently occurring subject headings and subheadings	• PubMed PubReMiner • EndNote	• EndNote
Identifying MeSH terms from relevant text such as a protocol	• MeSH on Demand	
Identifying MeSH terms used to index relevant records	• Yale MeSH Analyzer	
Phrases in the title and abstract	By document (sets of records): • Voyant • AntConc • TerMine	By document (sets of records): • Voyant • AntConc • TerMine
Words in proximity in the title and abstract	By document (sets of records): • WriteWords • AntConc	By document (sets of records): • WriteWords • AntConc

Several easy-to-use TMAs identify phrases and words in proximity. These all accept any document (with data from any database) and cannot analyze record by record. It is probably most useful to prepare a file of records comprising just the title and abstract to load into the TMA so that the phrase analysis can be as accurate as possible without "noise" from extraneous information in fields such as the address field.

WriteWords (writewords.org.uk/phrase_count.asp) does a very simple, quick, and clean phrase analysis and allows the searcher to choose how many words are involved in the phrase.

A set of titles and abstracts can be pasted into TerMine (nactem.ac.uk/software/termine), which produces an initial result analysis that shows all phrases highlighted in red. It is often helpful to set a display threshold (e.g., two or three), which reduces the number of phrases in red to those which appear at least two or three times in the text. Since the visual appearance of TerMine can appear cluttered, the phrases can be exported into a sorted table that can then be used for discussions with colleagues or saved for future reference.

A set of titles and abstracts can be pasted into the Voyant (voyant-tools.org) entry screen. Voyant will analyze the set and provide a simple word list with frequency count and a word cloud. It also offers a phrase analysis table and presents words in proximity. Data can be saved to a URL or as tables. Voyant is visually attractive to use and its option to save data sets to a URL is a really helpful feature for sharing and documentation.

AntConc (laurenceanthony.net/software/antconc/) can identify N-gram clusters — for example, two or three words that appear together — effectively finding phrases and

terms in proximity. AntConc will also find terms near a specific term entered by the searcher. Results can be saved to the clipboard, to a text file (.txt) from the file menu, or to a new window by clicking on the “save window” button.

Information specialists will want to minimize file transfers and might want to achieve search result analysis, management of records, and record selection within a single program. To provide insights into one possible program that might fulfill a more integrated role, we have assessed EPPI-Reviewer 4 (eppi.ioe.ac.uk/cms/er4). EPPI-Reviewer is a Web-based software program that is designed to support all stages of the systematic review process, including bibliographic management, screening, risk of bias assessment, data extraction, and synthesis. The most recent release of the software, EPPI-Reviewer 4, has introduced a number of features using text mining technologies. As EPPI-Reviewer is a subscription service, access to the text mining features requires the user or institution to purchase access to the whole software package, including all non-text mining functions. There are a number of options from single-user access to site-wide licensing; a one-month trial access is also available.

EPPI-Reviewer 4 is not particularly intuitive; the functionality and processes required are not clear to an inexperienced user without closely following the user manual (eppi.ioe.ac.uk/cms/Portals/35/Manuals/ER4.5.0 user manuala.pdf?ver=2015-10-12-122019-620). A major rewrite of the application is currently being undertaken. Correspondence with the developers indicates that they will be aiming to make the interface more user-friendly, that the software will be open source, and that it will include data structures suitable for the “next generation” of systematic review methods.

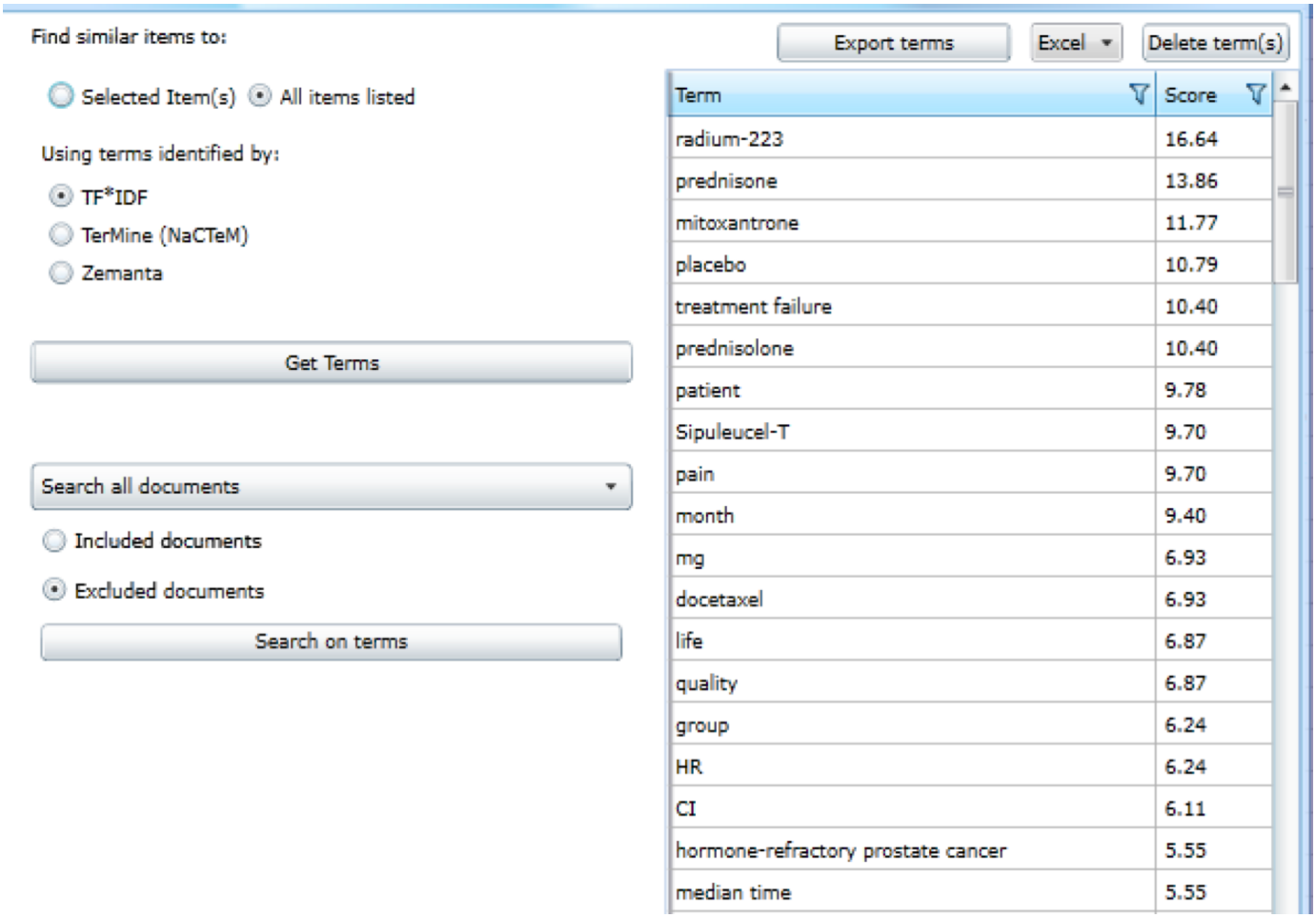
EPPI-Reviewer 4 accepts records in Research Information Systems (RIS) file format, meaning that text mining analysis can be used against outputs from most bibliographic databases and bibliographic management software. It can also process very large volumes of records; developers report that they have reviews with two million records and the machine learning components have been stress tested for up to 200,000 records at a time. We carried out testing using a sample library of 11,000 records.

EPPI-Reviewer 4 provides a number of different term extraction engines that will extract or mine relevant terms from the titles and abstracts of records. However, it cannot interrogate subject indexing fields in a meaningful way.

Although the user manual states that EPPI-Reviewer 4 provides access to four text mining engines (Term Frequency–Inverse Document Frequency [TF*IDF], TerMine, National Centre for Text Mining [NaCTeM], Zemanta, and Yahoo), only the first three of these appear as options on the user interface since Yahoo has disabled its service. Moreover, during our testing, we received an error message each time we tried to use the TerMine and Zemanta options. We note that TerMine can have down periods. We were only able to use TF*IDF as a term extraction engine. TF*IDF is a numerical statistic that reflects how important a word is to a document in a collection or corpus. The value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus. This helps to control for the fact that some words are generally more commonly used than others.

TF*IDF may be applied to all records, or batches of records such as included studies only, to identify and extract the key terms based on their frequency. This may be used to identify new relevant terms to include in search strategies, and to check that no significant terms have been missed. It may also identify potentially “noisy” terms, which are found in a large proportion of the records. The results may be exported so findings are easily shared with the wider review team, or saved so they may inform future searches in similar topic areas.

Figure 1: EPPI-Reviewer 4



3.2 Vague Topics

Information specialists deal with a range of issues concerning vague topics, which we have gathered under this heading. Many HTA products involve searches for difficult-to-define topics such as “implementation” and “ethics.” The topics may be fuzzy, or the vocabulary used to suggest them may be varied and non-standard. TMAs might help to focus such searches. Supplemental questions for drug topics are challenging since they can include issues such as the validity of outcomes and dosage information. TMAs might help with emerging topics where wording is still evolving and where the authors may hint or imply situations such as “second-line” treatment without mentioning the term explicitly. Information specialists also face ad hoc questions that do not necessarily need to be subject to systematic review, and in these cases it would be helpful to be able to automate quick searches. Example questions are shown in Figure 2. Information specialists also seek help with complex topics when trying to focus searches for grey literature.

Figure 2: Examples of Ad Hoc or Vague Questions of Interest

Examples include:

- price transparency in emergency room medication dispensing
- how scales and questionnaires can be found in the literature
- hospital transfer
- lab tests — laboratory-developed tests
- “non”/“not” questions; for example, procedures done by non-doctors, non-drug pain interventions, patients who are not having surgery
- short vs. long-term treatments
- length of stay
- “chemotherapy drugs”
- first responders (could be anyone)
- post disaster planning
- patient preferences.

vs. = versus.

At a very basic level the frequency analysis programs described in Section 3.1 will also help with showing terms which are used in records produced by scoping searches undertaken for some of these vague topics. Visual text presentation tools such as VOSviewer (vosviewer.com) (Section 3.2.1) can also be helpful since they show the concepts present in the search results and can help to identify additional relevant terms that could be added to the strategy to improve search sensitivity. Visual TMAs can also highlight irrelevant concepts. In doing so, TMAs can suggest options for developing strategies to exclude irrelevant topics and reduce noise in search results.

Document clustering tools can provide alternatives to the frequency analysis of database records. Document clustering tools include Carrot2 (Section 3.2.2) or Ultimate Research Assistant (Section 3.2.3). These tools can help to find highly relevant, often grey literature, documents (rather than bibliographic records) that can then be scanned to understand topics and gather relevant terms. For record clustering, rather than document clustering, TMAs like EPPI-Reviewer 4 can also be considered.

Although text frequency analyses and text visual analyses may show ways to focus or broaden strategies, they may still fail to help with some of the real issues involved in searches for vague topics. This is because these searches really require an analysis of word meaning as well as the presence of words. This can be typified by the ‘‘non”/“not” questions that information specialists encounter, where humans reading records can understand the negative setting or use of terms, but may find it difficult to operationalize that knowledge using Boolean search operators. For vague topics or themes that are likely to recur, it might be worthwhile investing in developing text mining solutions using a text mining specialist. The text mining specialist could work with information specialists to develop (within a sophisticated TMA) sets of text mining rules to interrogate the results of searches from a series of database to find likely relevant results. This is discussed in more detail in Section 4.

3.2.1 VOSviewer

VOSviewer (vosviewer.com) is a free-of-charge text visualization tool that can be used directly on the Internet or downloaded onto a PC. Like many TMAs, it requires Java to function. VOSviewer is relatively unusual in that it is designed to accept records easily from bibliographic databases such as PubMed and interfaces such as Web of Science. This makes it a valuable tool for analyzing batches of records from several resources. VOSviewer output is in the form of colour-coded maps showing the themes present in

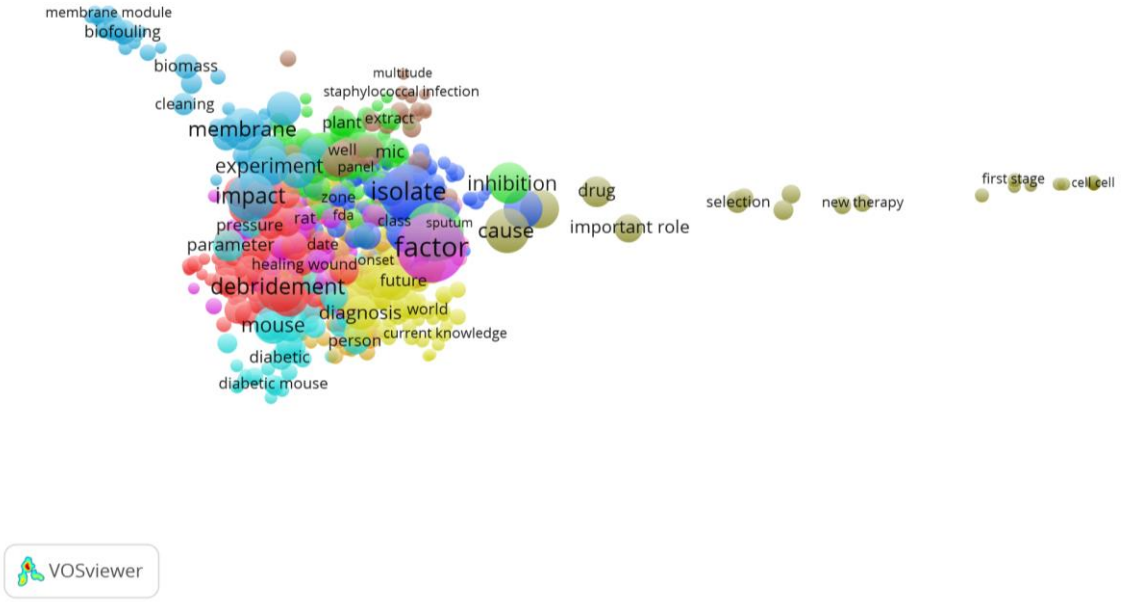
the records that have been analyzed. The value of these heat maps is that they can show how focused the current search has been and can highlight topics that the search might be tailored to avoid. For example, animal or cellular–level studies may show up as discrete coloured groups of terms, and information specialists can consider how to remove these types of studies from their search strategies. The VOSviewer zoom option allows us to consider parts of the heat map in detail to identify more specific terms for the search strategy.

One issue when using VOSviewer is which approach should be used to generate the result sets. VOSviewer can be used to analyze the results of one or more initial focused scoping searches to help develop the search strategy further. But it could also have a role in displaying the results of near-final search strategies visually, to see whether the strategy is optimized or could be improved further.

For the vague searches that information specialists encounter, VOSviewer may be able to show term groupings where search terms are used in senses other than the one of interest. VOSviewer heat maps can be helpful for showing team members the themes emerging from a literature and the maps can be retained to document search development. VOSviewer is quick to learn and results can be obtained rapidly. It also has an excellent and readable manual.

In test searches that we have undertaken to find records about the influence of isoflavones on the development of female cancers, the VOSviewer analysis of a simple scoping search showed three main conceptual groupings: women, rats, and cells. This was a signal to us that we could introduce a focus on women in the strategy and explore whether that would have a major impact in terms of missing relevant studies, and we could also carefully exclude records that focus on animal studies. In another search for biofilms (Figure 3) as a wound treatment, the VOSviewer display of the scoping search showed that biofilms are present in many other contexts than wounds, such as biofouling of equipment. Again, this suggested ways to focus the search strategy.

Figure 3: VOSviewer Presentation of Themes in a Scoping Search of Biofilms

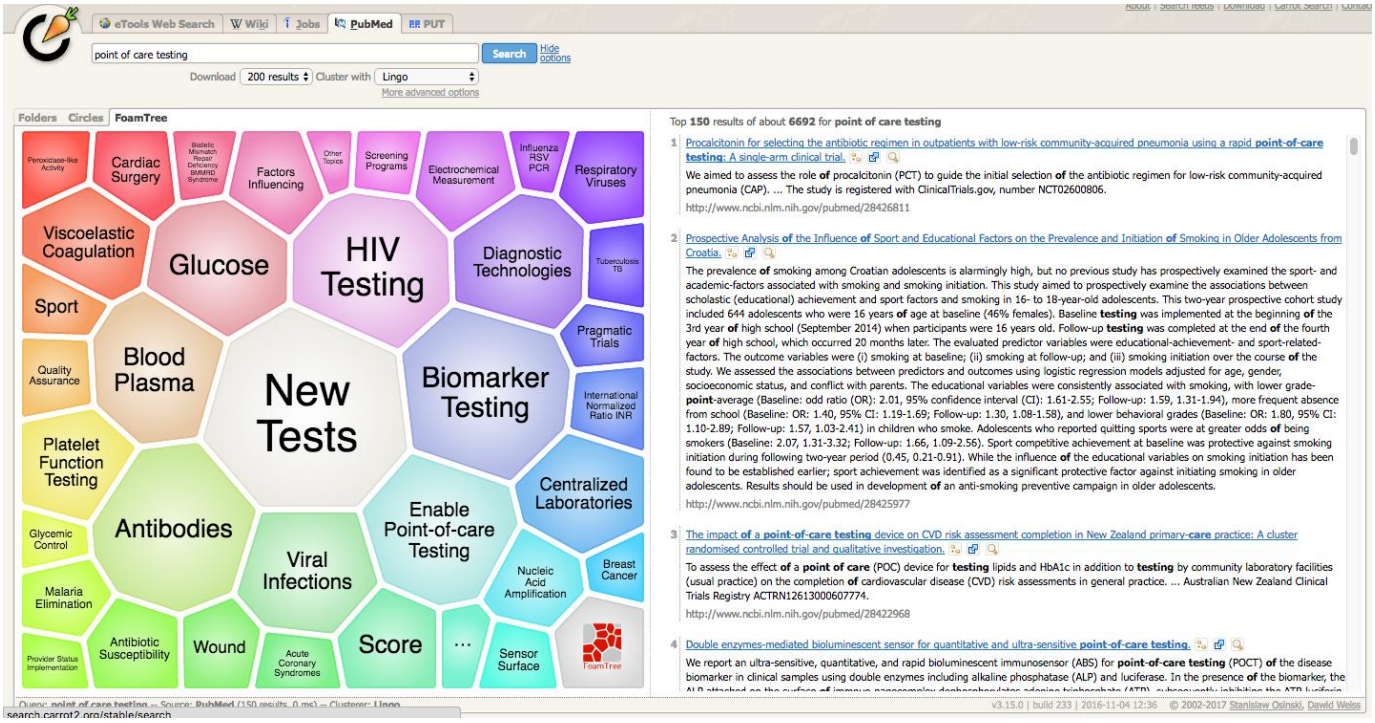


3.2.2 Carrot2

Carrot2 may help when developing searches for vague topics because it finds relevant reports based on simple searches. It is not just focused on database records, but finds guidelines, reviews, and other report literature, which can be useful starting points for researchers and for developing searches. As well as providing grey literature, Carrot2 also searches PubMed. A useful feature of Carrot2 is that it only retrieves and displays titles and abstracts from PubMed. This means it only analyzes the most useful information, and records do not need to be preprocessed through software such as EndNote.

Figure 4 shows a Carrot2 FoamTree visual analysis of the search results from PubMed of a search for "point of care testing." Carrot2 has grouped the results conceptually, which may offer insights into the literature from the perspectives of issues and search terms that are relevant or irrelevant. This visualization is from the free-to-use Carrot2 Internet service, but the fully programmable Carrot2 tools can also be obtained free of charge. The latter offer much more control and sophisticated analyses.

Figure 4: Carrot2 FoamTree Display of Search Results for the Search "Point of Care Testing"



3.2.3 Ultimate Research Assistant

Ultimate Research Assistant is an Internet search tool (ultimate-research-assistant.com/GenerateResearchReport.asp) (yielding a maximum of 50 documents), which carries out searches and generates a report summarizing the search topic. Using "point of care testing" as the search phrase, the program returns a topic summary defining "point of care testing" along with a list of key themes taken from a range of documents, and a listing of frequent words and phrases. These could all be useful for initial scoping of new topics and for identifying relevant and irrelevant search terms and concepts. Ultimate Research Assistant also offers other visualizations, which may be helpful depending on the topic in question, including a "taxonomy" that presents the

concepts in a hierarchy and a mind map of the taxonomy. These visualizations might all be helpful when developing a view of topic extent and relevance, and when trying to decide the boundaries of a topic.

3.2.4 MEDIE and Semantic MEDLINE

Vague topics might be ideal for TMAs that use semantic analysis approaches, interpreting the meaning of words by their role within sentences. Unfortunately, the free-to-use and off-the-shelf semantic analysis TMAs such as MEDIE or Semantic MEDLINE are still highly clinical in their focus, or are designed to find relationships in the literature around proteins and genes. Information specialists may wish to explore these TMAs to see if they assist with some of their search strings. However, for the questions we tested, their application seemed very focused and more suited to topics where clear relationships could be defined rather than for vague topics.

MEDIE (nactem.ac.uk/medie) is an interface to PubMed that runs queries structured in a subject-verb-object framework. For suitable structured questions this might yield results that could help with developing strategies. For example, in a search where “barriers” is used as the subject and “dialysis” is used as the object of the sentence, MEDIE found a series of papers that do deal with barriers to the use of dialysis. MEDIE’s table display option showed that MEDIE made a sensible interpretation of the missing verb, choosing words such as “included” or “prevent” in this example. MEDIE may also have a role in the searches that information specialists undertake for negative issues, since it offers a “not” modifier. This can be used with verbs, for example to search for what does not cause cancer. Sadly, the “not” modifier can only be applied to the verb, rather than the subject or the object, which would be more useful.

Semantic MEDLINE (skr3.nlm.nih.gov/SemMed) offers a PubMed search line and then summarizes the results of the search. The searcher chooses a “summary type” to specify the content focus, e.g., “treatment of disease” or “diagnosis.” The TMA then identifies predications: representation of a relationship expressed as subject and object joined by a relation (or predicate). This TMA is highly structured and is focused on treatment or diagnosis. It seems to have limited application beyond those two relationships, but may be worth testing for device searches that information specialists often find challenging.

3.3 Concept Identification

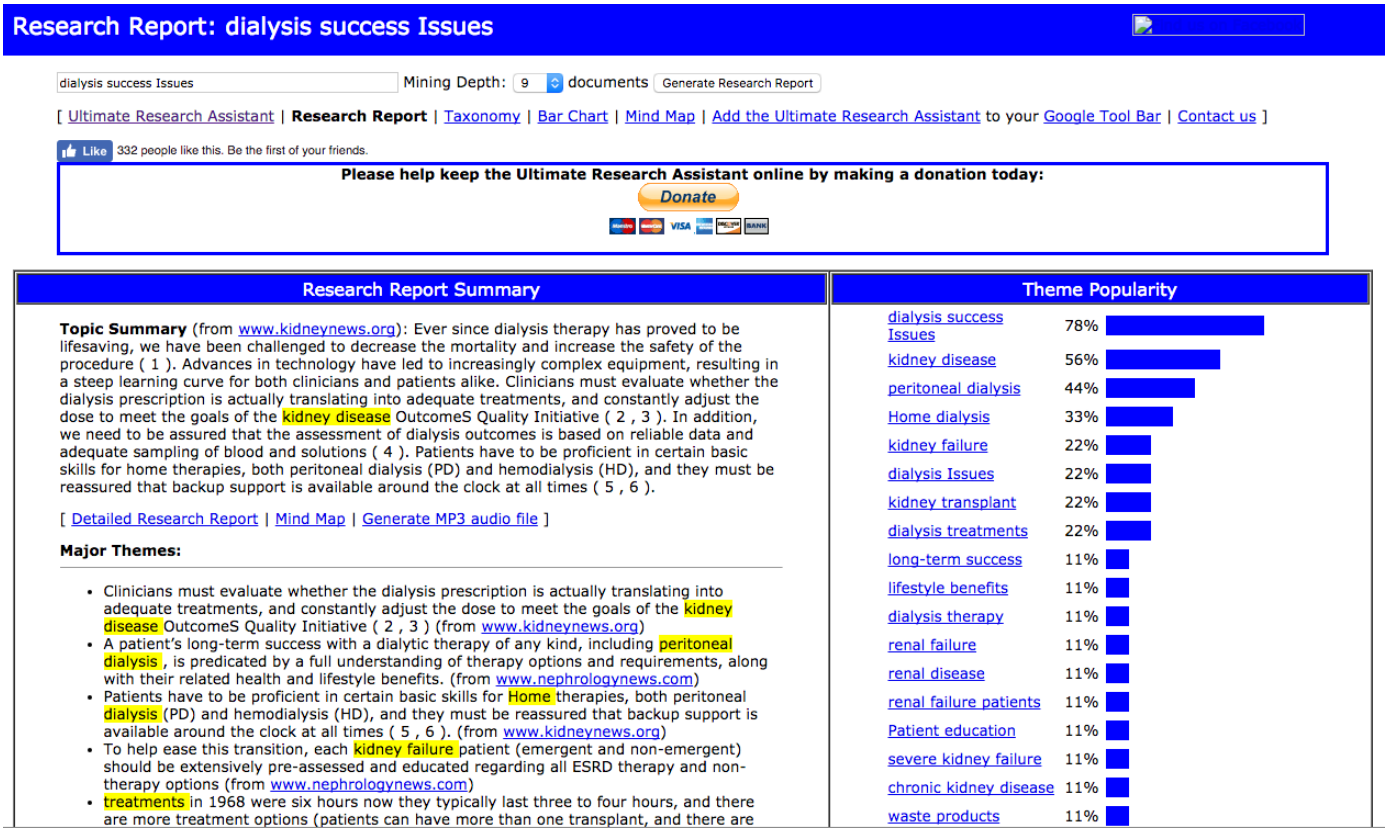
Information specialists face a range of issues that have been grouped under this broad heading. Concept identification for topics beyond searching for drugs might be helpful. Since conceptually difficult searches need to be done in the shortest timelines, useful TMAs need to be quick and easy to learn and use. Specific concepts that have proved difficult to search for and to identify the boundaries for have included issues such as “implementation.” TMAs that help with scoping would be valuable for fleshing out search concept terms such as settings, patient groups, or concepts to exclude. The results of TMAs might suggest how we could approach the search, as well as tools that offer confirmation of the value of known terms, and suggest new terms. TMA tools that provide help with identifying issues within a literature would also be valuable. For example, in a recent review of dialysis, information specialists were asked to identify which issues were important in dialysis uptake. In the case of very large result sets, the usefulness of TMAs to narrow down results safely should be assessed.

Some of the tools already suggested in Section 3.2 will help with concept identification. VOSviewer (Section 3.2.1) in particular is helpful for appreciating the concept groupings in the records retrieved by scoping searches and offers rapid identification of the themes of a literature. PubVenn (pubvenn.appspot.com) is a simple but helpful TMA with which

to show the relative sizes of concepts and the volume impacts of expanding or adding concepts to a search.

Carrot2 and Ultimate Research Assistant (sections 3.2.2 and 3.2.3, respectively) can help with concept identification and clarifying vague concepts. Figure 5 shows the results of a search using Ultimate Research Assistant for "dialysis success issues." The TMA generates a range of publications wherein issues for successful dialysis have been discussed. This could generate ideas for concepts around "success." It would need to be balanced with additional searches such as "dialysis failure issues," "dialysis implementation issues," and "dialysis barriers" to gain additional perspectives and suggestions.

Figure 5: Ultimate Research Assistant: Results for Search for Dialysis Success

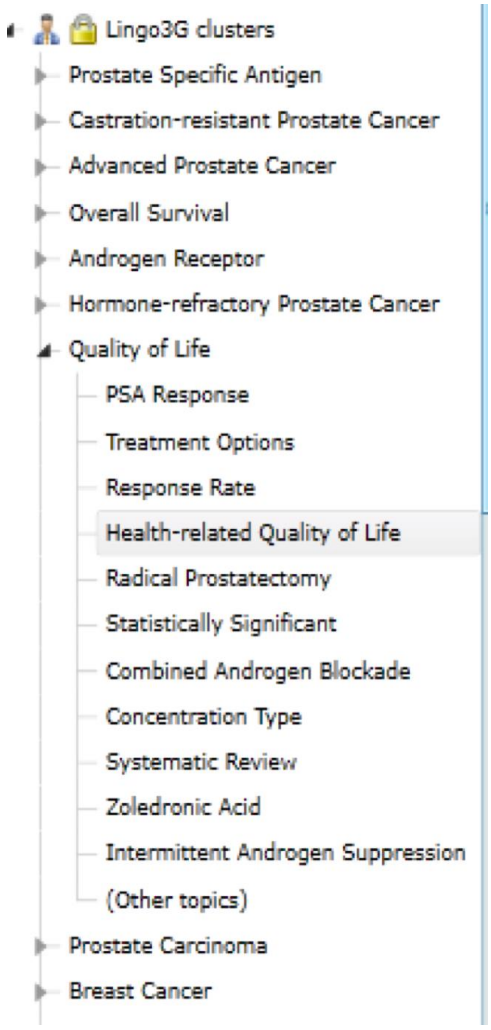


Coremine Medical (coremine.com/medical) may help with some types of concept identification. Typing in a search term generates a breakdown of concepts related to that search term. These concepts can be browsed to identify those which might be relevant or might be excluded. This TMA seems to be most helpful for more concrete searches rather than the vague topics such as "issues," since the analysis seems to be very MeSH-oriented.

EPPI-Reviewer 4 (eppi.ioe.ac.uk/cms/er4) is subscription software for managing the whole systematic review process. It uses the Lingo3G clustering engine to automatically categorize records into clusters based on the terms used in the title and abstract. Significant terms are extracted and used to code the records by theme. Lingo3G arranges these codes in a tree structure, allowing the relationship between them to be investigated. Some of the clusters automatically identified from records retrieved by example searches for a systematic review on enzalutamide for chemotherapy-naive,

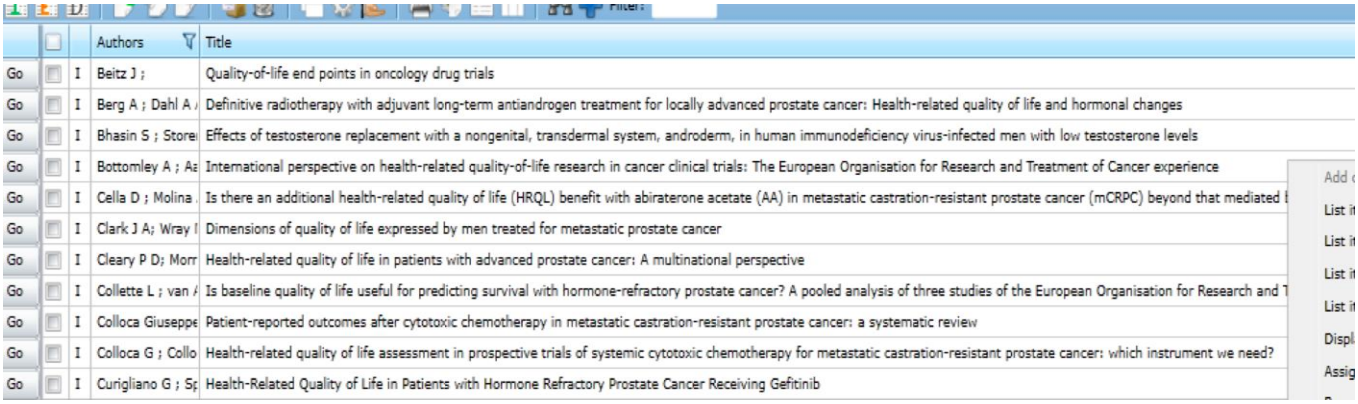
castration-resistant prostate cancer are shown below. EPPI-Reviewer 4 has identified a group of records on quality of life (Figure 6), and this code is expanded to show the "child" codes within this cluster.

Figure 6: EPPI-Reviewer 4: Clusters



The records assigned to each cluster can be viewed; the image in Figure 7 shows the titles of some of the records coded with "Health-related Quality of Life." Automatic clustering using EPPI-Reviewer 4 could potentially identify themes and issues in a large literature or a difficult-to-define literature and therefore aid the development of a search strategy. For example, a search strategy to identify studies evaluating methods of communicating public health messages about sunlight exposure may require the inclusion of terms related to many specific communication techniques that are not necessarily known at the beginning of the project. Applying a clustering tool to the results of a scoping search may identify types of intervention that are prevalent in the literature and can be explicitly built into later search strategies.

Figure 7: EPPI-Reviewer 4: Records Relating to Quality of Life



			Authors	Title
Go		I	Beitz J ;	Quality-of-life end points in oncology drug trials
Go		I	Berg A ; Dahl A /	Definitive radiotherapy with adjuvant long-term antiandrogen treatment for locally advanced prostate cancer: Health-related quality of life and hormonal changes
Go		I	Bhasin S ; Store	Effects of testosterone replacement with a nongenital, transdermal system, androderm, in human immunodeficiency virus-infected men with low testosterone levels
Go		I	Bottomley A ; Az	International perspective on health-related quality-of-life research in cancer clinical trials: The European Organisation for Research and Treatment of Cancer experience
Go		I	Cella D ; Molina	Is there an additional health-related quality of life (HRQL) benefit with abiraterone acetate (AA) in metastatic castration-resistant prostate cancer (mCRPC) beyond that mediated
Go		I	Clark J A; Wray I	Dimensions of quality of life expressed by men treated for metastatic prostate cancer
Go		I	Cleary P D; Morr	Health-related quality of life in patients with advanced prostate cancer: A multinational perspective
Go		I	Collette L ; van /	Is baseline quality of life useful for predicting survival with hormone-refractory prostate cancer? A pooled analysis of three studies of the European Organisation for Research and T
Go		I	Colloca Giuseppe	Patient-reported outcomes after cytotoxic chemotherapy in metastatic castration-resistant prostate cancer: a systematic review
Go		I	Colloca G ; Collo	Health-related quality of life assessment in prospective trials of systemic cytotoxic chemotherapy for metastatic castration-resistant prostate cancer: which instrument we need?
Go		I	Curigliano G ; Sp	Health-Related Quality of Life in Patients with Hormone Refractory Prostate Cancer Receiving Gefitinib

Such clusters may also identify groups of records that are clearly irrelevant. In the example in Figure 7 above, a cluster related to breast cancer was identified despite the relevant population being prostate cancer. The identification of irrelevant clusters may prompt the searcher to consider whether the final strategy could exclude that type of ineligible record without compromising sensitivity.

The type of concept identification shown in EPPI-Reviewer 4 and additional concept identification options are likely to be available in many of the commercial TMAs (Section 4).

3.4 Relevance Ranking to Assist with Search Refinement

Information specialists may be interested in learning whether relevance-ranking TMAs can help with identifying relevant search results rapidly, which can then be screened for additional search terms.

O'Mara-Eves and colleagues[1] have reviewed TMAs for record selection and suggest a range of tools. The techniques involved typically require the TMA to be told which records are relevant and which are not to identify further relevant records. O'Mara-Eves and colleagues note that only six systems were deployed in 2015, so the number of usable off-the-shelf systems is limited. Information specialists would ideally like tools that are quick and easy to use. This suggests that the TMAs need to be set up so that records can be loaded quickly, and then trained rapidly, with a few relevant and irrelevant records to progress quickly in obtaining suggestions for new relevant records. Some potential off-the-shelf options are described in this section, but it is likely that processes tailored to meet the needs of information specialists could be developed using sophisticated TMAs that provide machine learning tools.

Randomized controlled trials (RCTs) are probably the easiest type of record to identify reliably. The RCT Tagger tool (arrowsmith.psych.uic.edu/cgi-bin/arrowsmith_uic/RCT_Tagger.cgi) is rapid and easy to use, and does not require training. Using a population and intervention search strategy or even just the intervention search terms will generate a result set from PubMed in order of probability (on the right hand of the screen) that the record is an RCT. As the RCT Tagger becomes less certain, the probability value decreases.

The nails project (nailsproject.net) may be useful for any topic and any study design. With this tool, a search is undertaken in Web of Science databases, and the results are downloaded, zipped, and then analyzed using the free online Hammer program. The resulting analyses include many that are already easy to achieve with PubReMiner or

similar, but the unique helpful output is the “important papers” listing. This uses three measures to select important papers:

- in-degree in the citation network
- citation count provided by Web of Science (only for papers included in the dataset)
- PageRank score in the citation network.

Within the TMA, the top 25 highest scoring papers are identified using these measures separately, and then the results are combined and de-duplicated. The results are sorted by in-degree, and tied records are first broken by citation count and then by PageRank. Information specialists may wish to test this TMA to explore whether it can offer rapid access to new records from simple searches and potentially lead to suggestions of new search terms.

Medline Ranker (cbdm-01.zdv.uni-mainz.de/~jfontain/cms/?page_id=4) is an easy-to-use, but seemingly highly effective prioritizing tool. This free online resource requires a set of known relevant records with PubMed identifiers and a test set of other records, which might be search results or could be a random set of PubMed records found by Medline Ranker itself. In a test Figure 8 we used a batch of seven records about “barriers [ti] AND dialysis [ti]” as the relevant records, and the PubMed identifiers of 597 records meeting “dialysis [ti] AND 2017,” as the test set of records. Medline Ranker used the seven records to sort the 597 records and presented the records in order of similarity to the seven relevant records. This process was achieved in about 20 minutes from start to finish, including the two searches to find a set of relevant records and a set to mine for new records. Medline Ranker also provides a list of terms that discriminate relevant records from non-relevant records and could be considered for the search strategy. The tool provides tables of the parameters, the PubMed identifiers (PMIDs) of the training set, a table of ranked PMIDs, and discriminating words. These tables make it relatively easy to select PMIDs and search terms to feed into other software such as PubReMiner. It is not clear how large the training collection of relevant records needs to be, but presumably the larger the better.

Sciome Workbench for Interactive computer-Facilitated Text-mining (SWIFT Review) (sciome.com/swift-review) is a more structured TMA. It is free to use and manages many of the systematic review production processes. One aspect of the program is its machine learning option, which orders records by predicted relevance using reviewers’ decisions on a sample of records. Although it only works with PubMed records, this may not be a great disadvantage if it is only being used for identifying additional relevant records for the purpose of developing search strategies, rather than to process the end result of all searches.

Figure 8: Medline Ranker: Display of the Most Relevant Records and Colour-Coded Discriminating Words

Parameters

The training set: 7 / 7 abstracts.
The background set: the profile of the whole Medline database.
The test set: 527 abstracts.
Other: Scoring scheme: Bayes, Min word weight: 0.5, Target database: medline
Initialisation: *2 seconds*.

Top - Results - Discriminative words - Download

Results

Processing 457 abstracts: 0% - 9% - 19% - 29% - 39% - 49% - 59% - 68% - 78% - 88% - 98% - done.

Ranked in less than 1 second.

colors: yellow for training set pmids, green for background set pmids, and brown for *discriminative words* Color code: high low weight.

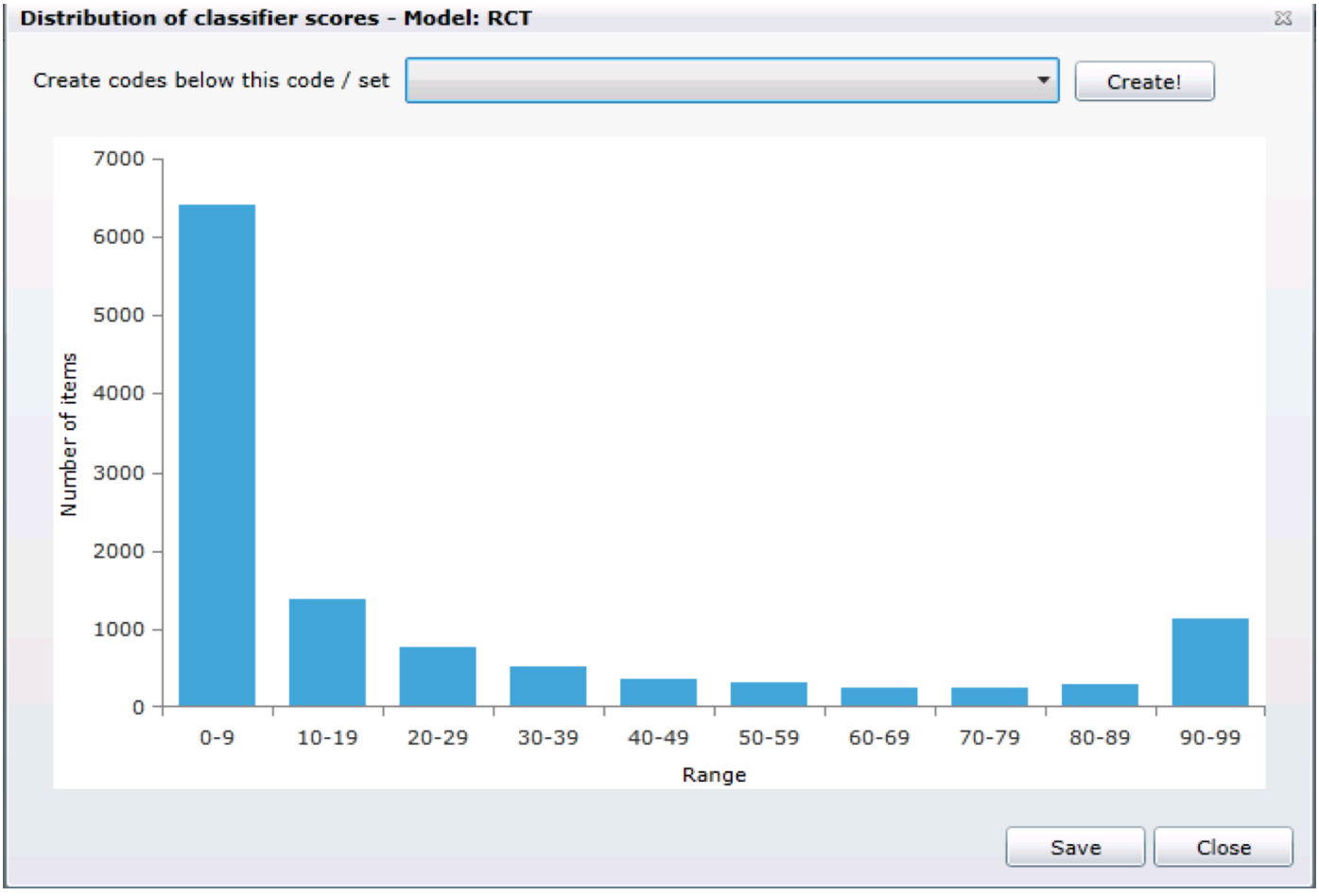
Click on the pmid to read the abstract with highlighted discriminative words

Rank	PMID	Abstract Title	P-value
1	28115864	Continuous ambulatory peritoneal dialysis : perspectives on patient selection in low- to middle-income countries.	6.44e-05
2	28144977	Changing Landscape for Peritoneal Dialysis : Optimizing Utilization .	7.06e-05
3	28242134	System -Level Barriers and Facilitators for Foregoing or Withdrawing Dialysis : A Qualitative Study of Nephrologists in the United States and England .	7.65e-05
4	26673908	Dialysis modality choice in elderly patients with end-stage renal disease : a narrative review of the available evidence .	7.87e-05
5	27555105	Dialysis Patient Perspectives on CKD Advocacy: A Semistructured Interview Study.	8.23e-05
6	27951552	What Is the Best Dialysis Therapy for South Asia: HD or PD?	9.00e-05
7	28201698	Transition of care from pre-dialysis prelude to renal replacement therapy : the blueprints of emerging research in advanced chronic kidney disease .	9.07e-05
8	28404600	Qualitative Interviews Exploring Palliative Care Perspectives of Latinos on Dialysis .	9.18e-05
9	28224937	Telehealth in the Delivery of Home Dialysis Care : Catching up With Technology .	9.41e-05
10	28090226	More Use of Peritoneal Dialysis Gives Significant Savings: A Systematic Review and Health Economic Decision Model .	9.64e-05
11	28417662	Clinician views of patient decisional conflict when deciding between dialysis and conservative management: Qualitative findings from the PAlliative Care in chronic Kidney diSease (PACKS) study.	9.66e-05
12	27831510	Consolidation in the Dialysis Industry, Patient Choice , and Local Market Competition.	9.79e-05
13	28329829	"End-of-Life Care ? I'm not Going to Worry About That Yet." Health Literacy Gaps and End-of-Life Planning Among Elderly Dialysis Patients .	9.95e-05
14	27951551	What Is the Best Dialysis Therapy in Developed and Developing Countries? Peritoneal Dialysis and/or Hemodialysis : The Trend in Korea.	9.97e-05

Abstrackr (abstrackr.cebm.brown.edu) is a record selection and screening program that also has functionality to semi-automate the screening process using active learning principles. Abstrackr's algorithm uses the inclusion and exclusion decisions made by reviewers to predict the likelihood of the remaining records being relevant. Once a reviewer has assessed "enough" records, the remainder can be screened automatically by Abstrackr. Rathbone et al.[5] tested Abstrackr's prediction algorithm using records from four systematic reviews with eligible studies comprising a range of study designs and publication types. The largest systematic review dataset was 1,735 records and the smallest was 517. They undertook sensitivity analysis with a 15,920-record dataset and assessed the precision, false-negative rate, and proportion of studies missed. The record prediction algorithm correctly identified all the relevant citations for two of the four reviews. In the remaining two reviews, it incorrectly predicted an included study was irrelevant in each. Neither of these incorrectly excluded studies had an abstract. The precision varied depending on the size and complexity of the review (16% to 45%). Very large sets of records with very few relevant records proved problematic for record selection, as there were insufficient eligible records on which to effectively "train" the software. However, this test was for full record selection rather than assisting with identifying records to help with scoping searches or developing searches. The prediction function is still being developed and the number of records needed to screen before predictions are made is currently arbitrarily set at approximately "a couple of hundred." This may be too many for most scoping exercises, but might be useful with a challenging topic where other techniques are not proving helpful and identifying more relevant records would be worth the investment of time. It might be that, following investigation and comparison with some of the other machine learning tools listed here, this tool might prove helpful for supporting scoping and strategy-building exercises.

EPPI-Reviewer 4 (eppi.ioe.ac.uk/cms/er4) contains a built-in machine learning classifier; the tool learns which records belong to a particular category as the reviewer classifies or screens a sample, and then applies this to the unscreened records. EPPI-Reviewer 4 is not focused solely on MEDLINE, but can process the results of any loadable databases. EPPI-Reviewer 4 contains one pre-built machine learning model, the "RCT model," which is designed to automatically identify RCTs and has been based on training from more than 280,000 records screened by Cochrane Crowd. The developers report that they plan to cover other study designs in the future. Users can build models to find other types of records that are specific to individual projects, which requires significant time coding and assessing a sample of records. This makes it challenging for projects that require a reasonably rapid turnaround, especially as the learning curve for such unintuitive software is particularly steep. However, we note that EPPI-Reviewer 4 does offer user support. If information specialists require an RCT identifier for records from a range of databases, then the RCT model can be run quickly to classify the records with a score from zero to 99; the higher the score, the more likely the record is an RCT. The records are presented in order of relevance, with the records most likely to be RCTs prioritized. The distribution of the classifier scores can also be visualized (Figure 9). The active learning functionality in EPPI-Reviewer 4, where the machine learns as records are screened, may also be helpful for reviews that are not just focused on RCTs. The machine learning tools are a new feature of EPPI-Reviewer 4 and are not yet featured in the user manual. EPPI-Reviewer 4 support staff are willing to take users through the software options.

Figure 9: EPPI-Reviewer 4: Randomized Controlled Trial Identifier



3.5 Filter Development

Information specialists anticipate that TMAs might help with search filter development in terms of validating the filters they have created previously and developing those filters further. TMAs might help with validation by confirming that all relevant terminology has been included in the filter, as well as perhaps identifying gold standard (GS) sets of records against which to test the filters.

Identifying relevant records to build GSs and reference sets is currently time-consuming, relying either on hand searching or the development of relative recall GS sets of records.[6] Finding more time-efficient but reliable methods to identify GSs would facilitate formal validation of search filters and enable the filters to be published. Publishing filters in peer-reviewed journals typically requires the authors to present information on the testing and validation of filters,[7-9] and many search filters currently lack this data.

Many information specialists produce less formal and more pragmatic search strings to find a wide range of concepts such as opioids, dentistry, and hospitalization. Information specialists anticipate that TMAs could be used to build these strategies and that TMAs might help with identifying whether search terms are changing over time to assist with keeping the search strings and search filters current.

We note that peer reviewers of text mining and search filter design papers tend to be cautious and will expect to see adequate detail and explanation of any techniques used to develop filters. Therefore, any testing and validation projects that information specialists decide to develop should ideally be piloted first to ensure that the method is robust and can deliver results. Any method developed needs to be a process that can be explained in a stepwise way, clearly to peer reviewers, to maximize chances for publication.

3.5.1 Search Term Identification

Assuming a known set of relevant records is available (whether a formally identified GS or partial [quasi] GS [QGS]), many of the TMAs discussed in Section 3.1 will prove useful to analyze the GS/QGS to identify search terms for testing in strategies:

- PubReMiner
- EndNote
- TerMine
- AntConc (for single terms and to identify phrases).

These TMAs will find frequently occurring textwords, subject indexing, and phrases, but then the challenge is selecting which terms to test. Publications have suggested ways that terms can be both identified and selected to test out to develop search filters.

Kok et al.[10] report how they used PubReMiner to identify frequently occurring single textword terms and MeSH headings and phrases (two to five terms) using TerMine, and then applied selection criteria to choose search terms to test in strategies. A search term was selected if it occurred in at least 5% of the articles in their GS and five times more often in the GS records as in non-relevant records. They then ranked the terms by using a “cross-product of both selection criteria.” The authors do not, unfortunately, report the software that they used to make these calculations, since doing so would likely require a table of words and frequencies per record, which is not easy to achieve from either PubReMiner or TerMine but could be achievable using other software such as Provalis Research’s SimStat and WordStat (provalisresearch.com) or AntConc. This is one example of one approach, and there are a range of choices such as the 5% frequency of occurrence that the authors implemented, which could be explored or varied.

Hausner et al.[11] suggest an approach to deriving search strategies using a test set derived from records included in systematic reviews. Hausner conducted term frequency analysis using the text mining applications within the R program.[12] Hausner's frequency selection cut-off is a term that appears in at least 20% of the GS records. These terms may not be discriminatory, so they are measured against a random set of records from a database such as MEDLINE (population set). The most frequent terms in the GS set with a low sensitivity of 2% or less in the population set are then selected. The authors hope that these terms discriminate GS records from non-relevant records. Again, this process of checking frequencies has to be done at the record level, suggesting it is carried out in R or another package. The authors select the candidate-controlled vocabulary using PubReMiner for MeSH or EndNote for other databases. In this example, the cut-off values for frequency are initially much higher, but the technique also uses non-relevant records to try to focus in on terms that are genuinely discriminatory.

Li and Lu also suggest methods of filter development using PubMed and Unified Medical Language System® (UMLS),[13] but their methods seem to be suitable for filters that find broad disease topics (the topics are defined by what searchers select rather than a priori definitions of what the searcher is seeking), and rely on access to the searches run in PubMed.

ProtAnt (laurenceanthony.net/software/protant) and many other TMAs provide options to identify search terms that discriminate GS/QGS records from other records. This requires identification of a set of records that are relevant and a set of irrelevant or less relevant records. ProtAnt analyzes documents (which may contain sets of records) rather than records, whereas other software, such as the Provalis tools, can be used to conduct an analysis on a record-by-record basis.

3.5.2 Search Term Confirmation

TMAs may be able to assist with search term confirmation. In this case they would be used to verify whether search strings or search filters contained terms likely to discriminate relevant records. For example, if it is possible to collect a set of irrelevant records as well as relevant records, then Medline Ranker (cbdm-01.zdv.uni-mainz.de/~jfontain/cms/?page_id=4), discussed in Section 3.4, may be able to show how well the search terms in a particular search string discriminate known relevant records from known irrelevant records (or just a general sample of PubMed records). AntConc and ProtAnt can compare a document of relevant records with a document of irrelevant records and again would demonstrate which are discriminating search terms. This might then provide confirmation of the validity of search strings and filters.

Alternatively, known relevant records could be analyzed and search terms selected using methods described by Kok et al.[10] or Hausner et al.[11] These search terms could then be compared with those in search filters and search strings.

3.5.3 Gold Standard Identification (For Filter Testing and Validation)

Identifying GS sets of RCTs is relatively straightforward and probably does not require a TMA. A robust approach would be to use CENTRAL as a source to identify GS sets of RCTs by topic, just as the National Health Service Economic Evaluation Database can be used as a proxy for historic GS sets of economic evaluations. RCT Tagger, discussed earlier, could be used as a tool for identifying RCTs to form subject-specific GSs, for example RCTs of dialysis. Citation network analysis may be another way to develop GSs, although again does not require a TMA specifically,[14] but does offer an approach that might go beyond RCTs. The process described by Belter[14] will still involve a large degree of record selection to discriminate relevant from irrelevant records in the citation lists that are obtained.

Identifying a GS for studies that are not RCTs is challenging. Again, there are some non-TMA methods for identifying relative recall GSs, such as using Epistemonikos (epistemonikos.org) or PDQ Evidence (pdq-evidence.org). These services can be used to identify reviews of eligible topics, and from these reviews we can gather the eligible studies. There is also the McMaster PLUS resource (hiru.mcmaster.ca/hiru/HIRU_McMaster_PLUS_Projects.aspx), which might also act as a source of records since it is a collection of records that have been collected in a systematic way, which could be used to form QGS.

TMAs that might help with identifying relevant records to form a GS in a semi-automated and rapid way include those listed in Section 3.4, such as Medline Ranker. The challenge in using a TMA is to define the file of records to be assessed for relevance in such a way that it is as robust as possible. So, for example, in the case of a "hospitalization" search string, the file would need to have been found using a sensitive strategy that is ideally broader than the "hospitalization" search string to be tested or validated. Then machine learning approaches could be used, specifying which records are relevant and which are not. Relevance and non-relevance would need to be very clearly defined, which is often challenging when describing vague topics. Once the TMA had been trained to identify relevant records using a sample, the rest of the set of records could be processed by the TMA.

Since one ideal for GS creation is handsearching relevant journals, a method to explore that would be less led (and influenced) by an existing search strategy might be to select a set of key relevant journals, add all their records into Medline Ranker, and then train the TMA to recognize relevant records. This GS would be representative rather than comprehensive, but would provide much less biased test material since the input records will not have been determined by a topic search. Peer reviewers may still criticize this approach, from the perspective that the journals are not representative. However, it might be possible to filter records from many journals if the training progresses well. Obviously, peer reviewers will also potentially query the choice of the TMA and its algorithms.

EPPI-Reviewer 4 offers TF*IDF analysis, which can be applied to "batches" of records indicated by codes. This allows EPPI-Reviewer 4 to use term extraction to identify further records that may be potentially relevant, based on a sample of records that have already been assessed. By screening a random sample and then applying TF*IDF to those that are deemed to be relevant, EPPI-Reviewer 4 learns which terms and ideas are likely to be found within a typical included item. EPPI-Reviewer 4 then provides the functionality to automatically search for the terms most frequently used in the sample of the included records across the remaining unassessed records. This might be a useful tool to test out to see whether it can assist with more rapid identification of sets of relevant records that might form a GS. We note that the EPPI-Reviewer 4 manual suggests that a minimum sample size of 1,000 records is required for the initial identification of included studies. However, for identifying a GS for filter development, this investment of time may seem worthwhile if the filter will have multiple future uses.

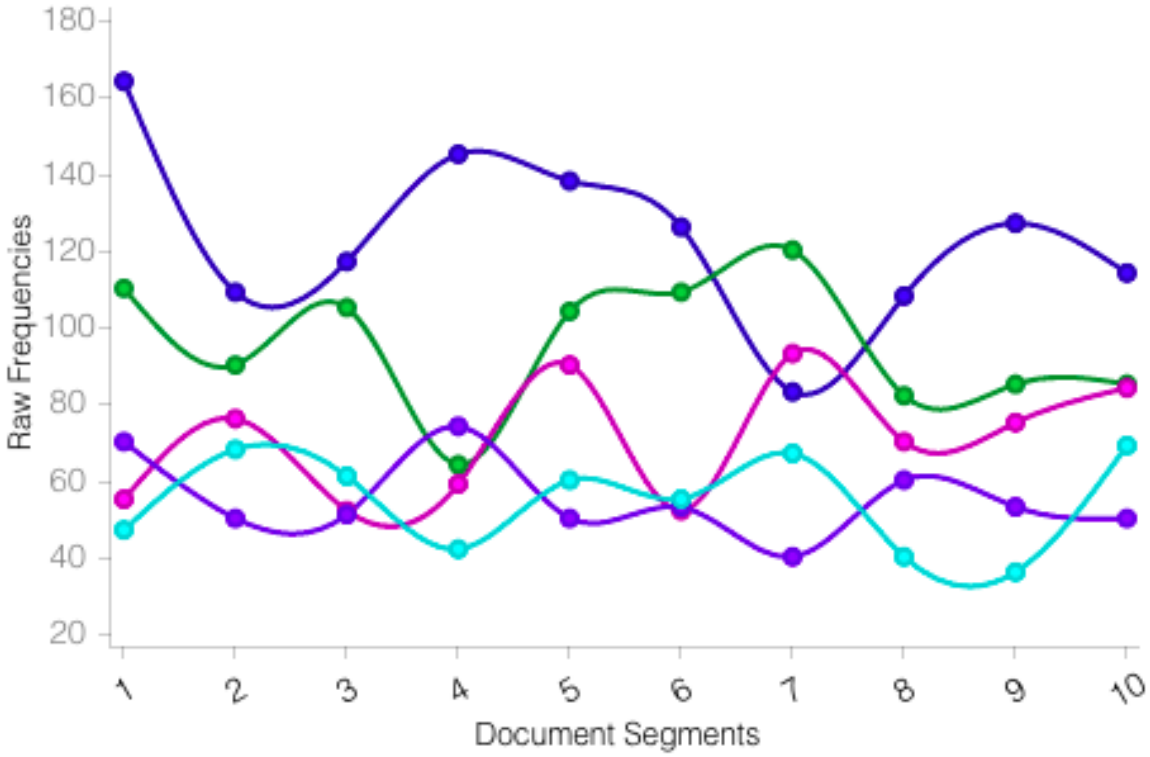
Other machine learning software is available.

3.5.4 Identifying Term Use Over Time

Terminology can change over time — for example, names for syndromes, ethnic groups, and nation groupings, as well as new treatments, techniques, and methods. Identifying whether terminology is changing is part of the process of keeping search strings and search filters current. Many of the TMAs we have listed will be able to show term use over time.

Voyant (voyant-tools.org) offers an option to show word trends over time. If a document with a set of records (from whatever source or sources) containing words of interest are loaded into Voyant in date order, then word frequency can be displayed across the document. This can show how often words are used over time. If a word appears to be trending downwards, this might suggest that it is being replaced with new terminology. Figure 10 shows a Voyant example of the frequency of appearance of five words across a document containing many records.

Figure 10: Voyant: Example of Frequency of Appearance of Five Words Across a Document Containing Records



3.6 Peer Review

TMAs might assist with peer review of search strategies. If a searcher has used TMAs to create the strategy, the TMA results, for example, could form part of the package of information passed from the search designer to the peer reviewer to speed up the assessment process by showing evidence of due diligence.

If the TMAs mentioned already (in Sections 3.1 and 3.5) have been used to develop strategies, then the designer can save screenshots of the results from the TMAs or download tables of search terms. If the strategy designer has used selection rules such

as those described in Section 3.5, these can also be documented as reassurance that a specific approach has been used.[10,11]

To demonstrate the impact of choosing extra concepts to include in a strategy, PubVenn screenshots can prove the relative size of broad concepts. With complex questions, where the concepts have been challenging to identify, it is possible to save the visual maps from TMAs such as VOSviewer and Voyant so that the peer reviewer can load and inspect them. The strategy designer may also wish to show before and after pictures to demonstrate the change in the results pictured between an early search and a fully developed search.

Polyglot search (crebp-sra.com/#/) can help search designers to do search syntax conversion between search interfaces, and reporting the use of this tool might reassure the peer reviewer that searches have been well prepared. This tool converts search syntax from Ovid, MEDLINE, or PubMed to supported databases and interfaces, currently PubMed, Ovid, MEDLINE, CENTRAL, Embase, Web of Science, and Cumulative Index to Nursing and Allied Health Literature. Some simple translation tasks work reasonably well — for example:

- MeSH to MeSH (PubMed to Ovid, Ovid to Cochrane, etc.)
- truncation ($ to *)
- lower case “or/and” in Ovid to upper case “OR/AND” in PubMed
- some aspects of phrase translation (translating breast cancer in Ovid MEDLINE to “breast cancer” in Web of Science).

Other translation tasks (both simple and more complex) do not work — for example:

- field searching (Ovid “ti,ab” to PubMed “[tiab]” to Cochrane “:ti,ab”)
- syntax for combining multiple lines (“or/1-4”)
- searching MeSH as major subject headings (“exp *breast cancer”)
- appropriate subject heading translation between databases.

Used with an awareness of its strengths and weaknesses, this could save the designer time and leave the peer reviewer to concentrate on the issues with which Polyglot search cannot help.

Section 4: How Can Sophisticated Text Mining Applications Help?

The “off-the-shelf” and easy-to-use TMAs described in Section 3 can provide a lot of help for specific search questions and some information retrieval processes. However, some of the specific challenges that many information specialists face, in terms of vague and complex questions, might be better approached using more sophisticated TMAs. This is particularly the case where search questions are about health service organization or services or “issues” that are not well captured in the clinical language systems, such as the UMLS, which underpin some of the TMAs we have reviewed.

There are many free sophisticated text mining packages, such as GATE (gate.ac.uk/overview.html), which can be tailored by experienced users to achieve specific search operations. Many of the free services explored in this document also offer more fully featured options (such as Carrot2), with powerful tools that can be exploited by experienced users. There are also a range of commercial packages whose features would need to be assessed in detail to see whether they offer advantages over free open-source software. There is also not-for-profit software, such as EPPI-Reviewer, where the payment for the software is used to maintain and support the software development. Many of these tools, including Provalis Research’s SimStat and WordStat as well as EPPI-Reviewer 4, are listed in the Agency for Healthcare Research and Quality report, and many more are likely to be available since that report was completed.[2] A list of more sophisticated TMAs that seem to be live, based on our very rapid assessment, is provided in Appendix B. Both free and commercial packages require time to review, learn, explore, and optimize to achieve the operations required by information specialists. Optimal use is likely to require a knowledge of statistics and the differences between algorithms, as well as a knowledge of linguistics.

Fully featured TMAs have the potential to optimize searches for vague or complex topics that might be encountered repeatedly by information specialists. For example, it should be possible to code up a TMA with the indicative terms and semantic relationships that indicate a document refers to a rural setting, non-doctors, or treatments that are not drugs. To achieve this would require an investment of time in identifying the terminology and the typical expressions that suggest a specific word grouping is indicative of a relevant case, and then building the rules that will ensure that these concepts can be retrieved in the future.

Fully featured machine learning environments could be programmed to exploit record prioritization for strategy development and for GS identification and strategy validation. Developing and providing TMA filters that can be published and made available to other users might be a useful external by-product of such exploration. If this is a consideration, then choosing a TMA that other health care information specialists might adopt would be important.

Few “off-the-shelf” packages seem to offer workable semantic analysis from an HTA perspective. Many such tools are designed for very specific tasks, such as identifying genes and proteins. Although semantic analysis can be programmed into TMAs such as GATE, there may be interim solutions where a commercial provider offers sophisticated but user-friendly tools. One such tool, which could be explored further, is Quertle (quertle.com), whose products are called Qinsight and Qexplorer. These packages offer a range of text visualization tools, but in the semantic aspect, offer Power-term searching in which Power terms are preprogrammed concepts encompassing many search terms, such as “$AdverseEvents” (quetzal-search.info/pages/powerterms.shtml). It is then possible to search for a substance or drug name linked by a verb such as “causes” to look for the records that contain the adverse events the substance or drug

might cause. For example, the search phrase “ethanol causes $AdverseEffects” uses the Power term to find records that report on the adverse effects of ethanol. Another Power term with value for economics-related searches might be “$Cost.” Quertle tools are not free to use, but might merit more detailed investigation from the information specialist’s perspective. The authors have had past discussions with Quertle and the company has confirmed in correspondence that they are open to building additional Power terms at customer request, and enriching ones they already offer with additional terms.

A test search in Qinsight using the phrase “causes dialysis uptake” found 20 records mentioning dialysis uptake, which also illustrated how “uptake” has other meanings. The Qinsight option also has tools to refine search results. “Exclude negatives” is used to refine the results to those that have a positive connotation, and “show only negatives” is used to refine the results to those that specifically include a negative connotation. Qinsight goes beyond keyword searching to interpret search terms and provide access to additional records that may contain other synonymous or related terms. For example, a search for “difficult patient” using Qinsight yields records about “complex patients,” “violent patients,” “uncooperative patients,” “problem patients” and “anxious patients.” The terms can also be displayed in a concept cloud. Qinsight accesses PubMed, but also other literature totalling 40 million documents, including more than 10 million full-text documents.

Sophisticated TMAs require specifications of the rules that make a record likely to be relevant or irrelevant. In machine learning applications, this specification is achieved by telling the software whether a record is relevant or irrelevant, and the software gradually recognizes relevance and can begin to grade records against the records it already knows. There are machine learning tools that are relatively quick to learn and use, but whether they can be trained effectively with the vague search topics experienced by many information specialists remains to be investigated. Other TMAs rely on user-built rules. These rules can vary in complexity but are intended to capture vague topics by analyzing word collocation and the semantic relationship between words. These rules may also rely on ontologies where the structure of a topic is captured in a series of formal definitions. We could imagine a vague concept such as “first responders” might be captured by definitions where someone performs CPR or another activity on someone having a heart attack or another illness. In a suitably defined text mining rule, the first responder would be indicated only by the fact that someone is acting on another person experiencing a specific crisis. Clearly the time and skills to develop and code such definitions, and then test them out, is not trivial. But if this is a frequent search, it might be worth the investment of time and effort to develop a set of rules that could be used and reused across records downloaded from any number of databases. The additional benefit of this approach is that the initial search within the bibliographic databases could be as general as “heart attacks or myocardial infarction,” because the focusing would be carried out within the TMA.

Search filters that can run on databases such as MEDLINE and other databases are of importance, but information specialists may also wish to consider whether a TMA environment might be the place to conduct most of the focused searching using filters developed for and built into the TMA. In this scenario, information specialists might design and carry out sensitive searches of a range of databases, de-duplicate the results and load the results into the TMA. The TMA would act as a hopper containing, potentially, many thousands of results that could then be queried by filters and search strategies developed within the TMA. This approach means that the strategies used for bibliographic database searching are optimized for sensitivity, and then focus is introduced within the TMA by a range of purpose-designed filters that make the most of the sophisticated options available within the TMA.

An example of this approach would be an HTA of HER2-positive breast cancer treatments. The searches of bibliographic databases would focus on finding all the variants of HER2-positive breast cancer terminology or the interventions of interest. The results from a range of databases and websites would be downloaded into a TMA and then various preprogrammed filters could be run on the records to find RCTs, and then to find economic evaluations and other topics as required. After, if new questions needed to be asked of the literature, they could be programmed into the TMA.

The challenges of this solution are many, not least finding the right TMA with the necessary flexibility and adopting an approach to information retrieval that is very different to that elsewhere in the HTA community. Potentially, there are also transparency issues in developing such an approach.

The choice of approach and selection of any TMA by information specialists would need to be determined based on objectives captured in a clear, detailed specification of the tasks to be achieved. There are likely to be published selection criteria that can be used to develop the specification, but some features that seem essential to the requirements of information specialists are listed in Table 2. It might be that more than one TMA is required, given the requirements that are agreed. Ideally, the specification for the TMA would be developed in collaboration with a text mining specialist.

Table 2: Selected Requirements for Consideration When Selecting a Text Mining Application

Feature	Details
Input options available or programmable	Since bibliographic records are the unit of analysis for information specialists' tasks, the TMA should ideally be able to import records in RIS format and other common bibliographic formats, or be programmable so that a plug-in can be built to process bibliographic records into individual records. Records from multiple databases can be gathered into a single corpus of records.
Minimal data input/output	It is a major inconvenience to have to pass data in and out of several programs to achieve a result. The optimal TMA would be one in which all the results of searches from a range of databases could be processed and analyzed together, and the results passed to reviewers for assessment, without the involvement of further external processing through additional software. Ideally, the TMA would be able to include many of the tasks in the HTA process.
Incoming fields can be defined and selected for querying and analysis	The ideal analysis would be able to discriminate text from the title, abstract, and keywords, and other imported information would be ignored unless specified. Text in title and abstract fields should be processed as text, but the text in keywords fields should be parsed as phrases, since the phrases are meaningful.
Rules	The TMA should be able to specify the rules for relevance that records must meet. The underlying rules and algorithms within the TMA should be transparent so that these can be referenced in any reporting.
Relationships	The TMA should be able to build descriptions of relationships between concepts, and if necessary to load in ontologies available from elsewhere on the Internet.
Boolean as well as semantic analysis	It would be helpful to be able to achieve Boolean searches as well as rule-based searches and other text mining approaches within the same TMA, so that the impact of different approaches could be benchmarked against the results of a Boolean search (of the title and abstract words).
Comparisons of relevant and non-relevant records /machine learning	This option involves the ability to train the TMA to recognize relevant records and reject irrelevant records.
Documentable	The record flow and the processes that records undergo should be clearly reportable, so that the processes within the text mining package can be reported in journal articles or within reports.

Feature	Details
Output of records	Ideally, the TMA would also be a bibliographic reference tool that would interact with Microsoft Word to permit the production of reference lists. This seems an unlikely possibility at present; therefore, export of bibliographic references into Reference Manager or EndNote via RIS would be essential. Export in a format that permits easy loading into record selection software is also a desirable feature unless record selection is also included within the TMA and is suitable to the information specialist's needs.

HTA = health technology assessment; RIS = research information systems; TMA = text mining application.

Section 5: Discussion and Recommendations

Despite great advances in the use of TMAs in information retrieval for health care research, there is still only piecemeal availability of TMAs for the range of tasks with which information specialists need support. The information retrieval community is interested in increased access to “plug-and-play” tools, rather than tools that require programming skills, although the latter are likely to be the tools that offer the opportunity for tailored solutions.[3]

There are some easy-to-use TMAs for specific tasks that information specialists could use in their daily work to develop search strategies and document their methods. There are also TMAs which may help with identifying relevant records and opportunities to explore the use of TMAs to develop and validate search filters. Many of these tools can be learned and used quite rapidly.

For some of the more challenging issues for information specialists, such as concept identification and record identification for vague topics and the development of robust validated filters, the involvement of a text mining specialist and the acquisition of fully featured sophisticated tools that can achieve the outputs that the information specialists require could be considered for a business plan.

Another option might be for information specialists to identify TMAs that they find appealing and partner with the publishers of those TMAs to purchase bespoke extensions to existing software. Many TMA publishers seem open to collaborative ventures.

If considering options for fully featured software, information specialists may also wish to identify what configuration of TMAs they would prefer:

- Stand-alone tools can be tailored for specific tasks, which may involve transferring files of records between software packages.
- A TMA package that is used to hold and process search results following sensitive searches of a range of databases may be preferred. The TMA would act as a hopper for potentially thousands of results, which could then be queried by filters and search strategies developed within the TMA. This approach would mean that the strategies used for database searching are optimized for sensitivity, and then focus is introduced within the TMA.
- A fully featured search results package where search results are maintained and searched, relevant records are selected, and potentially data extraction and quality assessment and other review activity is achieved, is another option.

Choosing a fully featured tool requires a detailed technical specification based on a clear view of the desired end products or uses. A range of TMAs could then be identified and assessed against the specification to find which are the best fit.

References

1. 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. Syst Rev [Internet]. 2015 Jan 14 [cited 2017 Sep 28];4(1):5. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4320539
2. Paynter R, Banez LL, Berliner E, Erinoff E, Lege-Matsuura J, Potter S, et al. EPC methods: an exploration of the use of text-mining software in systematic reviews [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2016 Apr. [cited 2017 Sep 28]. (AHRQ methods for effective health care). Available from: https://www.ncbi.nlm.nih.gov/books/NBK362044/
3. Maralte BV. Text mining of journal literature 2016: insights from researchers worldwide [Internet]. [London]: Publishing Research Consortium; 2016. [cited 2017 Sep 28]. Available from: http://www.publishingresearchconsortium.com/index.php/prc-documents/prc-research-projects/54-prc-text-mining-of-journal-literature-2016/file
4. Frequently asked questions: what is text mining? [Internet]. Manchester (United Kingdom): The National Centre for Text Mining (NaCTeM), University of Manchester; 2016. [cited 2017 Nov 10]. Available from: http://www.nactem.ac.uk/faq.php?faq=1
5. Rathbone J, Hoffmann T, Glasziou P. Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers. Syst Rev [Internet]. 2015 Jun 15 [cited 2017 Sep 28];4:80. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4472176
6. Sampson M, Zhang L, Morrison A, Barrowman NJ, Clifford TJ, Platt RW, et al. An alternative to the hand searching gold standard: validating methodological search filters using relative recall. BMC Med Res Methodol [Internet]. 2006 Jul 18 [cited 2017 Sep 28];6:33. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1557524
7. Jenkins M. Evaluation of methodological search filters--a review. Health Info Libr J. 2004 Sep;21(3):148-63.
8. Glanville J, Bayliss S, Booth A, Dundar Y, Fernandes H, Fleeman ND, et al. So many filters, so little time: the development of a search filter appraisal checklist. J Med Libr Assoc [Internet]. 2008 Oct [cited 2017 Sep 28];96(4):356-61. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2568852
9. Bak G, Mierzwinski-Urban M, Fitzsimmons H, Morrison A, Maden-Jenkins M. A pragmatic critical appraisal instrument for search filters: introducing the CADTH CAI. Health Info Libr J. 2009 Sep;26(3):211-9.
10. Kok R, Verbeek JA, Faber B, van Dijk FJ, Hoving JL. A search strategy to identify studies on the prognosis of work disability: a diagnostic test framework. BMJ Open [Internet]. 2015 May 19 [cited 2017 Sep 28];5(5):e006315. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4442145
11. Hausner E, Guddat C, Hermanns T, Lampert U, Waffenschmidt S. Development of search strategies for systematic reviews: validation showed the noninferiority of the objective approach. J Clin Epidemiol. 2015 Feb;68(2):191-9.
12. Feinerer I. A framework for text mining applications within R. In: tm - Text Mining Package. Version 0.7-1 [Internet]. Version 0.7-1. Vienna (Austria): The Comprehensive R Archive Network; 2017 [cited 2017 Oct 18]. Available from: https://CRAN.R-project.org/package=tm
13. Li J, Lu Z. Developing topic-specific search filters for PubMed with click-through data. Methods Inf Med [Internet]. 2013 May 13 [cited 2017 Sep 28];52(5):395-402. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3744813
14. Belter CW. Citation analysis as a literature search method for systematic reviews. J Assn Inf Sci Tec. 2016;67:2766-77.

Appendix A: Frequency Analysis in EndNote

EndNote offers simple frequency analysis. All records are indexed as they are loaded into EndNote, and the keywords field is usually indexed automatically to create a term list. Thus, any indexing added to an EndNote keywords field will be indexed and can then be analysed for frequency. Term lists can be used to create new indexes on any field or combined fields e.g., the abstract field, or the title and abstract fields.

To analyse indexing in the keywords field, before loading records into EndNote, decide how to treat the information coming in to the keywords field. You can break up subject index terms by changing the term delimiters; for example, “Pseudomonas infection/dt [Drug Therapy]” could be parsed as:

- separate words, breaking on the space — Pseudomonas Infection dt drug therapy
- two phrases breaking on the “/” — “Pseudomonas Infection” “dt [drug therapy]”
- Two phrases breaking on the “[“— “Pseudomonas Infection/dt” “[drug therapy]”

To analyze subject headings correctly, it is important to keep helpful phrases together and perhaps to divide subheadings from the subject headings. If you want to do frequency analysis of other fields or combinations of fields, do so once the records are loaded.

Ensuring the keywords field is processed correctly, begin with a new empty EndNote library and use the following commands:

- Select “tools,” “define term lists,” and then “keywords.”
- (Change) the delimiters — for example, tick the box next to the “/” symbol to make sure that subheadings will be treated separately from an Emtree heading.
- Then click on “update list”.
- Click “OK.”
- Load your Embase records.

To create the frequency analysis of the keywords field, use the following commands:

- Choose "tools” and then “subject bibliography.”
- Choose “keywords” and click “OK.”
- Choose “select all” and click “OK.”
- Select "layout” to choose the display format.
- Choose “terms” and then “subject terms only.”
- Change the number of lines between entries by removing the “suffix” “^p^p.”
- Change the display order to frequency by selecting “by term count — descending.”
- Click “OK.”
- To print the frequency listing, select “print.”
- To save the listing, select “save.”

To see a title and abstract frequency analysis, first define a term list as follows:

- Select “tools” and then “define term lists.”
- Select “create list.”
- Give the list a helpful name — e.g., “Titleabs.”
- Check the custom delimiters and make sure that “custom delimiter” is ticked, and then add a single space to the box next to it, to ensure that words will be processed individually.
- Select “update list.”
- Then select “add field”, choose the Title field and it will be pasted into the box.
- Then select “add field” again and this time choose the Abstract field. It will be pasted into the box below Title.
- Then select “OK.”

The term list is now linked to two specific fields and we are ready to generate a subject bibliography using the title and abstract field.

To save or print out the title and abstract frequency analysis, use the following commands:

- Select “tools” and then “subject bibliography.”
- Tick the box labelled “In other fields,…”
- Select “title” as well as “abstract” (using the Control key) and click “OK.”
- Choose “select all” and click “OK.”
- Change the display format by selecting “layout.”
- Select “terms” and then “subject terms only.”
- Change the number of lines between entries by removing “^p6p” from the suffix.
- Change the subject term layout by selecting “by term count — descending.”
- Select “OK.”
- To print listing, select “print.”
- To save the listing, select “save.”

Appendix B: Selected Sophisticated Text Mining Applications

Table 3: Selected Sophisticated Text Mining Applications

Text Mining Application	URL
AQUAD (qualitative text analysis)	http://www.aquad.de/en/download/
AQ21 (machine learning)	https://www.mli.gmu.edu/index.php/research/ontology-guided-machine-learning/
Carrot2	http://search.carrot2.org/stable/search
Coding Analysis Toolkit (also DiscoverText)	http://cat.texifter.com/
CATMA	http://catma.de/
EPPI-Reviewer 4	https://eppi.ioe.ac.uk/cms/er4/
GATE	https://gate.ac.uk/overview.html
JMP (SAS add-on)	https://www.jmp.com/en_us/offers/text-analysis.html?utm_campaign=td70114000002KZJq&utm_source=google&utm_medium=cpc&utm_term=text%20mining
KH Coder	http://khc.sourceforge.net/en/
KNIME	https://www.knime.org/
Leximancer	http://info.leximancer.com/
Lingo3G	https://carrotsearch.com/lingo3g/
Linguamatics (i2E)	https://www.linguamatics.com/products-services/ways-deploy/i2e-ondemand
MALLET	http://mallet.cs.umass.edu/
Multiparadigm Indexing and Retrieval MIMIR	https://gate.ac.uk/mimir/
OpenNLP	https://opennlp.apache.org/
PEx	http://vicg.icmc.usp.br/vicg/tool/1/projection-explorer-pex
Provalis Research (SimStat, QDA Miner, WordStat)	https://provalisresearch.com/
PubTator	https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/PubTator/index.cgi?user=User888584375&searchtype=PubMed_Search&query=chlamydia+screening&page=1
Quetzal and Quertle	https://www.quetzal-search.info/
R Simple Learner	http://rutcor.rutgers.edu/pub/rrr/reports2007/5_2007.pdf
RapidMiner	http://docs.rapidminer.com/
TinySVM	http://chasen.org/~taku/software/TinySVM/
Tm in R	http://tm.r-forge.r-project.org/index.html
UIMA framework	https://uima.apache.org/
Weka	http://www.cs.waikato.ac.nz/~ml/weka/

AQUAD = Analysis of qualitative data; CATMA = Computer Assisted Textual Markup and Analysis; EPPI = Evidence for Policy and Practice Information; KNIME = Konstanz Information Miner; MALLET = Machine Learning for language Toolkit; PEx = Projection Explorer; UIMA = Unstructured Information Management Application.

Appendix C: Other Text Mining Applications

The following text mining applications (TMAs) mentioned in Appendix E of Paynter et al, have been rapidly assessed for this report, but are not mentioned in detail in the text because of their focus or because we think other tools are better.

Table 4: Other Text Mining Applications

Resource	URL	Reasons for Non-Selection
AdTAT	http://www.adelaide.edu.au/carst/resources-tools/adtat/	AntConc seems to offer more options.
BioTextQuest	http://bioinformatics.med.uoc.gr/cgi-bin/biotextquest/textQuest.cgi	Creates word clouds and co-occurrence visuals, which can be achieved in Voyant.
Chilibot	https://github.com/chen42/chilibot	Code archived.
Citavi	https://www.citavi.com/en/features.html	Reference management software.
Concordance	http://www.concordancesoftware.co.uk/	Not currently available for Windows.
DBpedia	http://wiki.dbpedia.org/about	DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Internet.
Doctor Evidence	http://drevidence.com/products/doc-data/	Basic search engine: http://drevidence.com/products/doc-library-2/.
G-Bean	http://bioinformatics.clemson.edu:8080/G-Bean/index.jsp (Runs on PubMed — exporting is tricky.)	Seems to collect the “related to” items from PubMed. Rather slow and the export records function is unclear.
GoPubMed	http://www.gopubmed.org/web/gopubmed/	It does not offer title and abstract analysis, unlike PubMed PubReMiner.
Hierarchical Clustering Explorer	http://www.cs.umd.edu/hcil/hce/	Used for genome analysis.
HubMed	http://git.macropus.org/hubmed/	Simple search interface to PubMed — no advantage over PubReMiner.
HuGENet and related tool GAPscreener	https://www.cdc.gov/Genomics/hugenet/default.htm https://phgkb.cdc.gov/HuGENavigator/home.do https://omictools.com/genetic-association-publication-screener-tool	Human genome epidemiology resource.
Import.io	https://www.import.io/	Extracts data out of web pages and into Microsoft Excel.
KH Coder	http://khc.sourceforge.net/en/	It is not clear whether it is analyzing by database record.

Resource	URL	Reasons for Non-Selection
Medline trend	http://dan.corlan.net/medline-trend.html	This page displays the number of entries (articles) in PubMed (MEDLINE) published every year that conform to a search strategy (such as a phrase) entered by searcher.
MEDSUM	http://webtools.mf.uni-lj.si/public/medsum.html#	Word frequency analysis — no obvious advantage over PubReMiner.
MeSHy	http://tools.bat.infspire.org/meshy/	This shows MeSH terms that often occur near to each other. When tested on April 24, 2017, this did not work.
Metta	https://www.cs.uic.edu/~ljia/MetaSearchEngine.HTML	We could not connect to the software.
MyMiner	http://myminer.armi.monash.edu.au/	Very slow and therefore difficult to evaluate.
NLM Medical Text Indexer	https://ii.nlm.nih.gov/MTI/index.shtml	Output is probably useful for inserting into other programs in a text mining workflow.
ParsCit	http://wing.comp.nus.edu.sg/parsCit/	This seems to be a tool to identify references from documents.
Parsifal	https://parsif.al/	Tool designed for software reviews with searching organized to find software topics.
PEx	http://vicg.icmc.usp.br/vicg/tool/1/projection-explorer-pex	Visual presentation of text mining.
PICO	https://pubmedhh.nlm.nih.gov/nlmd/pico/piconew.php	A PICO form which can be used to query PubMed.
PubCrawler	http://pubcrawler.gen.tcd.ie/	A free "alerting" service that scans daily updates to the NCBI MEDLINE (PubMed) and GenBank databases and returns MEDLINE and GenBank database records that match specified research interests.
PubMatrix	https://pubmatrix.irp.nia.nih.gov/cgi-bin/index.pl	This program operates in batch mode — it returns search results but does not seem to offer added value compared with other programs.
PubNet	http://pubnet.gersteinlab.org/	Creates author networks based on a subject search.
PubTator	https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/PubTator/index.cgi?user=User888584375&searchtype=PubMed_Search&query=chlamydia+screening&page=1	PubTator is a Web-based text-mining tool to search articles semantically and to create and export human annotations. Focused on searches of genes and proteins.
PubViz	http://pubviz.fhstp.ac.at/	PubViz is a tool for interactive visualization of publication data: shows amounts of publications along a timeline, uses bar charts for different publication types, word clouds to show co-authors as well as keywords, and a groupable list view.
Qiqqa	http://www.qiqqa.com/	PDF manager but with visual analysis as well. The visual analyses may be a premium feature.

Resource	URL	Reasons for Non-Selection
Reflect	http://reflect.ws/	This tags gene, protein, and small molecule names in any web page.
RetroMine	https://retromine.univ-rennes1.fr/	Not currently live (Oct 2017).
SensPrecOptimizer	http://systematicreviewtools.com/tool.php?ref=SensPrecOptimizer	This claims to optimize searches.
Site Content Analyzer 3	http://www.cleverstat.com/en/sca-website-analysis-software-index.htm#sca_features	Website analysis software for website optimization.
SLRtool	https://github.com/javipeg/SLRtool	A search and download option that downloads BibTex references.
StArt	http://lapes.dc.ufscar.br/tools/start_tool	This is systematic review management tool and runs searches from within the tool.
STRING	http://string-db.org/	Protein-protein interaction network.
Systematic Review Assistant	http://www.datamining.org.uk/sysreview.html	Quality assessment/risk of bias tool.
ReVis	http://www2.ccsl.icmc.usp.br/pt-br/projects/revis	A tool to support the selection and quality evaluation of primary studies in systematic reviews. It claims to provide visual mappings of the set of primary studies to help the user explore the data.
Text Analyzer	https://www.online-utility.org/text/analyzer.jsp	We judged that other tools can do these tasks as well and with additional features.
Unusual Words	https://www.online-utility.org/text/finds_unusual_words.jsp	There are other more sophisticated tools available to help with the identification of discriminating words.
Whatizit	http://www.ebi.ac.uk/webservices/whatizit/info.jsf	Semantic annotator for identifying proteins, etc.
Wordsmith	http://www.lexically.net/wordsmith/version5/index.html	This processes files rather than records, but otherwise it is fully featured.
Xpdf	http://www.foolabs.com/xpdf/	A PDF resource; not a TM resource.

AdTAT = Adelaide Text Analysis Tool; G-Bean = Graph-based Biomedical Search Engine; GAPscreener = Genetic Association Publication screener; MeSH = Medical Subject Headings; NLM = National Library of Medicine; PDF = Portable Document Format; PEx = Projection Explorer; PICO = Patient, Intervention, Comparison, Outcome; SLRtool = Systematic Literature Review tool; StArt = State of the Art through Systematic Review; ReVis = Systematic Review Supported by Visual Analytics; TM = text mining.