Methods Research Report

A Prospective Comparison of Evidence Synthesis Search Strategies Developed With and Without Text-Mining Tools

Methods Research Report

A Prospective Comparison of Evidence Synthesis Search Strategies Developed With and Without Text-Mining Tools

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
5600 Fishers Lane
Rockville, MD 20857
www.ahrq.gov

Contract No. 290-2017-00003C

Prepared by:
Scientific Resource Center
Portland, OR

Investigators:
Robin A. Paynter, M.L.I.S.
Celia Fiordalisi, M.S.
Elizabeth Stoeger, B.S.
Eileen Erinoff, M.S.L.I.S.
Robin Featherstone, M.L.I.S.
Christiane Voisin, M.L.S.
Gaelen P. Adam, M.L.I.S., M.P.H.

AHRQ Publication No. 21-EHC008
March 2021

This report is based on research conducted by the Agency for Healthcare Research and Quality (AHRQ) Scientific Resource Center, funded through the following contracts: Scientific Resource Center III (290-2017-00003C), Brown University EPC (290 2015 00002I), ECRI Institute-Penn University EPC (290 2015 00005I), Alberta University EPC (290-2015-00001I), and RTI-University of North Carolina EPC (290 2015 00011I). The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

The information in this report is intended to help EPCs and AHRQ understand how EPC reports can be improved to benefit health-system decision making. This report is not intended to be a substitute for the application of clinical judgment. Anyone who makes decisions concerning the provision of clinical care should consider this report in the same way as any medical reference and in conjunction with all other pertinent information, i.e., in the context of available resources and circumstances presented by individual patients.

Suggested citation: Paynter RA, Fiordalisi C, Stoeger E, Erinoff E, Featherstone R, Voisin C, Adam GP. A Prospective Comparison of Evidence Synthesis Search Strategies Developed With and Without Text-Mining Tools. Methods Research Report. (Prepared by the Scientific Resource Center under Contract No. 290-2017-00003C). AHRQ Publication No. 21-EHC008. Rockville, MD: Agency for Healthcare Research and Quality. March 2021. Posted final reports are located on the Effective Health Care Program search page.
DOI: https://doi.org/10.23970/AHRQEPCMETHODSPROSPECTIVECOMPARISON.

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of healthcare in the United States. The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new healthcare technologies and strategies. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To improve the scientific rigor of these evidence reports, AHRQ supports empiric research by the EPCs to help understand or improve complex methodologic issues in systematic reviews. These methods research projects are intended to contribute to the research base in and be used to improve the science of systematic reviews. They are not intended to be guidance to the EPC program, although may be considered by EPCs along with other scientific research when determining EPC program methods guidance.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers as well as the healthcare system as a whole by providing important information to help improve healthcare quality. The reports undergo peer review prior to their release as a final report.

If you have comments on this Methods Research Project they may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 5600 Fishers Lane, Rockville, MD 20857, or by email to epc@ahrq.hhs.gov.

David Meyers, M.D.
Acting Director
Agency for Healthcare Research and Quality

Arlene Bierman, M.D., M.S.
Director
Center for Evidence and Practice Improvement
Agency for Healthcare Research and Quality

Christine Chang, M.D., M.P.H.
Acting Director
Evidence-based Practice Center Program
Center for Evidence and Practice Improvement
Agency for Healthcare Research and Quality

Jill Huppert, M.D, M.P.H.
Task Order Officer
Center for Evidence and Practice Improvement
Agency for Healthcare Research and Quality

Acknowledgments

The authors gratefully acknowledge the continuing support of our AHRQ Task Order Officer, Jill Huppert, MD, MPH. In addition, we thankfully acknowledge the following individuals for their contributions to this project: Chelsea Ayers, M.P.H., Tracy Dana, M.L.S., Michele Freeman, M.P.H., Jennifer Lege-Matsuura, M.S.L.I.S., Rebecca Rishar, M.L.I.S., and Ritu Sharma, Bs.C. We would also like to thank the following organizations for their contributions: the Cochrane Collaboration and the U.S. Department of Veterans Affairs Evidence Synthesis Program.

Peer Reviewers

Prior to publication of the final white paper, the EPC sought input from independent Peer Reviewers without financial conflicts of interest. However, the conclusions and synthesis of the scientific literature presented in this report do not necessarily represent the views of individual reviewers.

Peer Reviewers must disclose any financial conflicts of interest greater than $10,000 and any other relevant business or professional conflicts of interest. Because of their unique clinical or content expertise, individuals with potential non-financial conflicts may be retained. The TOO and the EPC work to balance, manage, or mitigate any potential non-financial conflicts of interest identified. The list of Peer Reviewers follows:

Elke Hausner, M.Sc.
Institute for Quality and Efficiency in Health Care
Cologne, Germany

Kanaka Shetty, M.D., M.S.
Southern California/RAND Evidence-based Practice Center
RAND Corporation,
Santa Monica, CA

Claire Stansfield, Ph.D.
Evidence for Policy and Practice Information and Co-ordinating Centre
UCL Social Research Institute
University College London
London, United Kingdom

Siw Waffenschmidt, Dr.
Institute for Quality and Efficiency in Health Care
Cologne, Germany

A Prospective Comparison of Evidence Synthesis Search Strategies Developed With and Without Text-Mining Tools

Structured Abstract

Background: In an era of explosive growth in biomedical evidence, improving systematic review (SR) search processes is increasingly critical. Text-mining tools (TMTs) are a potentially powerful resource to improve and streamline search strategy development. Two types of TMTs are especially of interest to searchers: word frequency (useful for identifying most used keyword terms, e.g., PubReminer) and clustering (visualizing common themes, e.g., Carrot2).

Objectives: The objectives of this study were to compare the benefits and trade-offs of searches with and without the use of TMTs for evidence synthesis products in real world settings. Specific questions included: (1) Do TMTs decrease the time spent developing search strategies? (2) How do TMTs affect the sensitivity and yield of searches? (3) Do TMTs identify groups of records that can be safely excluded in the search evaluation step? (4) Does the complexity of a systematic review topic affect TMT performance? In addition to quantitative data, we collected librarians' comments on their experiences using TMTs to explore when and how these new tools may be useful in systematic review search creation.

Methods: In this prospective comparative study, we included seven SR projects, and classified them into simple or complex topics. The project librarian used conventional “usual practice” (UP) methods to create the MEDLINE search strategy, while a paired TMT librarian simultaneously and independently created a search strategy using a variety of TMTs. TMT librarians could choose one or more freely available TMTs per category from a pre-selected list in each of three categories: (1) keyword/phrase tools: AntConc, PubReMiner; (2) subject term tools: MeSH on Demand, PubReMiner, Yale MeSH Analyzer; and (3) strategy evaluation tools: Carrot2, VOSviewer. We collected results from both MEDLINE searches (with and without TMTs), coded every citation’s origin (UP or TMT respectively), deduplicated them, and then sent the citation library to the review team for screening. When the draft report was submitted, we used the final list of included citations to calculate the sensitivity, precision, and number-needed-to-read for each search (with and without TMTs). Separately, we tracked the time spent on various aspects of search creation by each librarian. Simple and complex topics were analyzed separately to provide insight into whether TMTs could be more useful for one type of topic or another.

Results: Across all reviews, UP searches seemed to perform better than TMT, but because of the small sample size, none of these differences was statistically significant. UP searches were slightly more sensitive (92% [95% confidence intervals (CI) 85–99%]) than TMT searches (84.9% [95% CI 74.4–95.4%]). The mean number-needed-to-read was 83 (SD 34) for UP and 90 (SD 68) for TMT. Keyword and subject term development using TMTs generally took less time than those developed using UP alone. The average total time was 12 hours (SD 8) to create a complete search strategy by UP librarians, and 5 hours (SD 2) for the TMT librarians. TMTs

neither affected search evaluation time nor improved identification of exclusion concepts (irrelevant records) that can be safely removed from the search set.

Conclusion: Across all reviews but one, TMT searches were less sensitive than UP searches. For simple SR topics (i.e., single indication–single drug), TMT searches were slightly less sensitive, but reduced time spent in search design. For complex SR topics (e.g., multicomponent interventions), TMT searches were less sensitive than UP searches; nevertheless, in complex reviews, they identified unique eligible citations not found by the UP searches. TMT searches also reduced time spent in search strategy development. For all evidence synthesis types, TMT searches may be more efficient in reviews where comprehensiveness is not paramount, or as an adjunct to UP for evidence syntheses, because they can identify unique includable citations. If TMTs were easier to learn and use, their utility would be increased.

Key Messages

Purpose of study

The objectives of this study were to compare the benefits and tradeoffs of searches with and without the use of text-mining tools (TMTs) for evidence synthesis products in real world settings. Specific questions included: (1) Do TMTs decrease the time spent developing search strategies? (2) How do TMTs affect the sensitivity and yield of searches? (3) Do TMTs identify groups of records that can be safely excluded in the search evaluation step? (4) Does the complexity of a systematic review topic affect TMTs performance? In addition to quantitative data, we collected librarians' comments on their experiences using TMTs to explore when and how these new tools may be useful in systematic review search creation.

Key messages

- TMTs appear to decrease the time required to develop keyword and subject terms compared to usual practice (UP) search strategy development in six out of seven reports, but the small sample size precludes significance.
- TMTs searches appear less sensitive than UP searches in all but one project, but the small sample size precludes significance.
- Number-needed-to-read (NNR) results were mixed; NNR was lower using TMTs compared with UP in four out of seven reports. Again, the small sample size precludes significance
- TMTs neither affected search evaluation time nor improved identification of exclusion concepts (irrelevant records) that can be safely removed from the search set.
- Across "simple" review topics (i.e., single indication-single drug) TMTs yielded no unique additional relevant citations while missing only one relevant study in three of four reports and reduced time spent on creating searches compared to UP. Thus, TMTs may be useful in simple review search strategy development, and when timeliness is prioritized and comprehensiveness is not critical.
- Across "complex" review topics (e.g., multicomponent interventions) TMTs identified some unique includable citations and reduced time spent in search strategy development but missed more relevant citations compared to UP. TMTs may be more useful as an adjunct to usual practice for complex evidence synthesis reviews (e.g., evidence maps, scoping reviews, systematic reviews, health technology assessments, and update reviews, etc.) especially when comprehensiveness is critical and the review team has adequate time to handle the increased screening burden.

Contents

Tables

Figures

Appendixes

Introduction

Background

Given the explosive growth in biomedical evidence, information retrieval methods research is needed to ensure efficient and effective search processes. Limited investigations to date suggest that the use of text-mining tools (TMTs) in systematic reviews save production time and improve the quality of search results.[1-4] An "objective" approach to developing search strategies using TMTs has been adopted and validated by Germany's Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG).[5-7] Hausner et al. define the objective approach as comprising a set of steps: "generation of a total set (relevant references from systematic review), splitting of the total set into a development set and comparator set, development of the search strategy with references from the development set (analyzing information derived from the titles and abstracts of relevant references with text-mining tools), and validation of the search strategy (checking whether references from the comparator set can be identified with the search strategy developed beforehand)."[6]

To ascertain the applicability of text-mining-based search approaches in a real-world setting, across a variety of review topics and using freely available tools, and specifically inform the practice of evidence synthesis librarians working for the Agency for Healthcare Research and Quality (AHRQ) Evidence-based Practice Center (EPC) Program, a group of EPC librarians and other methodologists conducted a prospective comparative study of search strategy development with and without TMTs.

Search Strategy Development

In their usual practice (UP), evidence synthesis librarians develop search strategies using a sequential process of tasks to identify search terms, generate a logic structure, and evaluate performance (see Figure 1). Many subprocesses focus on achieving an optimal combination of search keywords and subject heading terms (e.g., Medical Subject Headings [MeSH]) to retrieve all relevant citations on a systematic review topic with as few irrelevant citations as possible. Finding an acceptable balance of search sensitivity/recall to precision/number needed to read (NNR) is time consuming and requires analysis of exploratory search strategies and known relevant citations.[8, 9]

For this study, the UP method for developing search strategies was evaluated against a TMT method, specifically in respect to three search subprocesses shown in Figure 1: (1) developing model keyword (title and abstract) terms/phrases, (2) developing model subject heading terms, and (3) evaluating model search strategy, including reviewing relevant citations.

Text-Mining Tools

Many software programs analyze textual documents and bibliographic citations.[10, 11] Some measure the frequency of term/phrase occurrences in a corpus of text, some suggest subject heading terms based on a set of bibliographic citations or a sample of text, and others generate a visual representation of search results to show relationships (e.g., between authors or related topic areas). TMTs may be custom built to support systematic review production or intended for entirely unrelated research tasks. They may be web-based or require downloaded software, either freely available or through a paid license. Resource costs are also associated with the time

needed to learn how to use these tools; while some TMTs are intuitive and simple, others are complex programs that require invested training time.

The variety of available TMTs makes it difficult to decide which tool to use for search strategy development and to determine if the benefit to the systematic review is worth the time needed to learn how to use these tools.[12, 13] This investigation informs on the utility of freely available TMTs for the benefit of EPC evidence synthesis librarians and a broader international community of systematic review producers.

Objectives

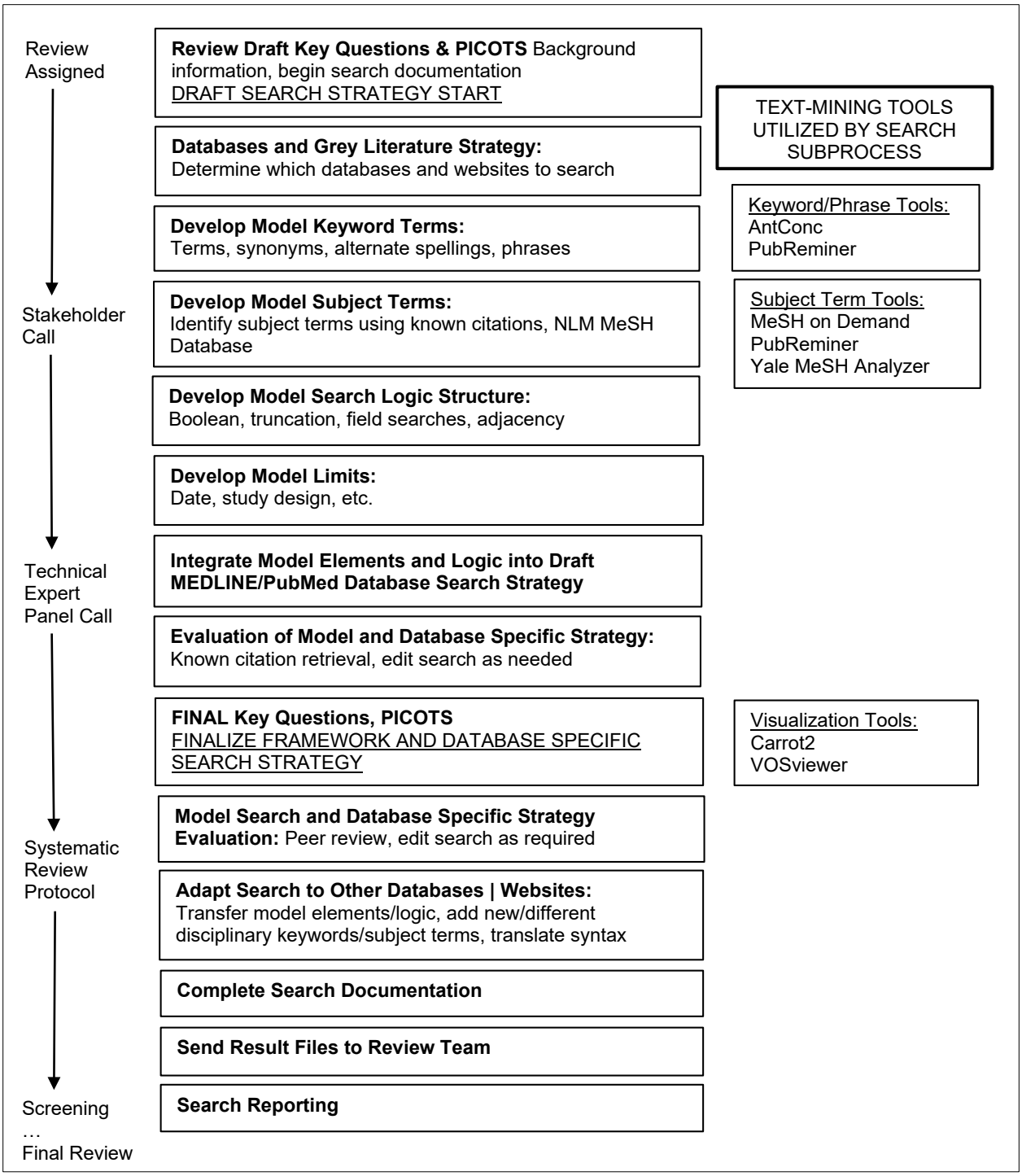
The objectives of this study were to compare the benefits and tradeoffs of searches with and without the use of TMTs for evidence synthesis products in real world settings. Based on prior research[5-7, 12-14] and what we knew about the role of some TMTs in term generation and other TMTs as a way to identify irrelevant term clusters, we hypothesized that TMTs would increase sensitivity and precision (operationalized as NNR [1/precision]), and reduce time needed to develop search strategies.

Key Questions

The following Key Questions (KQs) guided our investigation and the decision to collect both quantitative and qualitative data to inform our results:

1. Do TMTs decrease the time needed to develop keyword and subject term strategies compared to conventional approaches?
2. Does increased search recall from TMTs—
   a. improve the yield of relevant citations that would not have been found using conventional techniques?
   b. result in an unreasonable number of excess citations for screeners given existing staff resources?
3. Does using TMTs in the draft search strategy evaluation step improve the final search by identifying groups of irrelevant records which can safely be removed from the results (improving precision)?
4. Does the type of review topic (complex versus simple) make a difference in the performance of TMTs, according to the criteria evaluated in Questions 1–3?

Figure 1. Usual practice search strategy development process



Abbreviations: MeSH = Medical Subject Headings (Medline), NLM = National Library of Medicine, PICOTS = Population, Intervention, Comparison, Outcome, Timing, Setting.

Methods

Project Identification and Recruitment

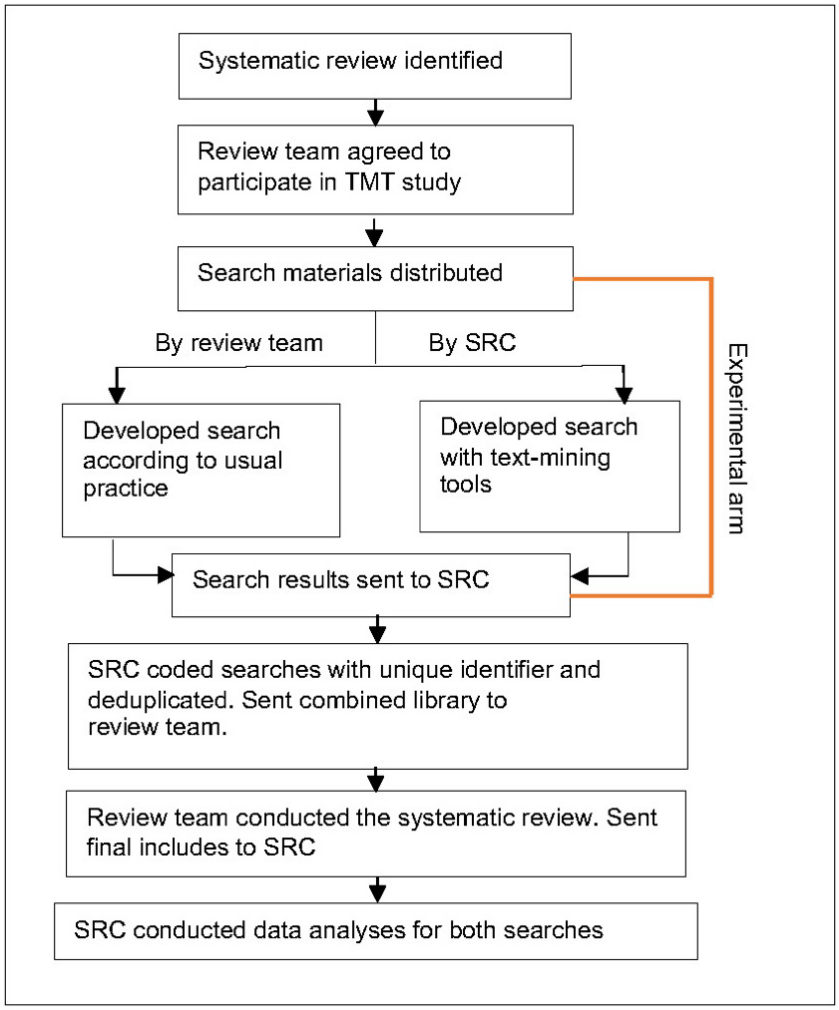
The investigators identified systematic reviews with an approximate completion date of February 2020 that were eligible for inclusion in this project. This date was chosen in an attempt to ensure that all projects were completed before analysis. To ensure a variety of review types, eligible reviews were any systematic review or systematic review update funded by AHRQ or conducted by the U.S. Department of Veterans Affairs Evidence Synthesis Program (VA ESP), or any Cochrane group. These groups were chosen because the project librarians overlapped with our research team of six EPC-affiliated librarians, and were known to follow the methods set out in the EPC Methods Guide[8] and Cochrane Handbook[9] for search strategy design. A target of 10 included reviews was set, including one pilot review to test our process. The target was set to achieve as large a sample as possible within the project timeframe (1 year). We estimated the average number of reviews produced by these groups annually and chose a slightly lower number, anticipating recruitment might be difficult. The six EPC-affiliated librarians discussed each review topic and came to a consensus on classifying it as simple or complex to assess whether adding TMTs was more useful for simple or complex topics.

The Scientific Resource Center (SRC), the methods support center for the EPC program, assisted in coordinating between investigators and review teams. Once a potential review was identified, the SRC contacted the review team, confirmed their participation, and then assigned the study to a TMT search librarian. Three review teams did not respond to our request or declined to participate, due to the extra workload. The study design is summarized in Figure 2.

The MEDLINE search strategy was completed by one UP librarian and one TMT librarian. The UP librarian was the librarian on the team running the chosen review; the TMT librarians were randomly chosen from a group of six EPC-affiliated librarians. All UP and TMT librarians had a master's degree in library science and were affiliated at the beginning of the project with an EPC. More details about the librarians are in the results section. Several librarians worked on more than one project, and a librarian could be a UP librarian on one project and a TMT librarian on another. The UP librarian provided the SRC with the search background materials (the review protocol, minutes of relevant meetings, and other similar documents) and a bibliography or list of PubMed IDs (PMIDs) to be used as seed citations for the TMT librarian's use.

The UP and TMT librarians developed their searches simultaneously and independently using the same platform (Ovid or PubMed), and the TMT librarian search was limited to the same date range as the UP final search. We did not prescribe how TMT librarians were to use the tools, so TMT librarians were free to use UP practices, including determining Boolean structures, ensure all concepts had both MeSH and free-text terms, and checking the results against known citations. TMT and UP librarians were matched by platform because the different interfaces yield different results, thereby reducing potential confounding due to search platform. Librarians remained anonymous throughout the study. All searches were peer reviewed by a second librarian using the Peer Review of Electronic Search Strategies (PRESS) assessment form (http://www.sciencedirect.com/science/article/pii/S0895435616000585)[15] to help ensure that both searches were of high quality.

Figure 2. General overview of study design



Abbreviations: SRC= Scientific Resource Center, TMT= text-mining tools

Selection of Text-Mining Tools

TMT librarians could choose one or more TMTs per category from a preselected list in each of three categories: (1) keyword/phrase tools: AntConc,[16] PubReMiner[17]; (2) subject term tools: MeSH on Demand,[18] PubReMiner,[17] Yale MeSH Analyzer[19]; and (3) strategy evaluation tools: Carrot2,[20] VOSviewer.[21] The TMTs and categories were agreed upon by the investigators before the study began and reflect the free, open-source, or web-based tools, available to most librarians (note that local security firewall issues may preclude use in individual information technology environments) and known to the investigators at the outset of the study. Due to the exploratory nature of research on using text-mining tools in systematic reviews searching, we did not prescribe how librarians used them in our study because the evidence for best practices is still in its infancy. Thus, each librarian had the flexibility to use the tools in the way they considered them to be most helpful. The specific tools and methods each librarian used were captured in the tracking sheet (See Appendix A).

Complex Versus Simple Review Topics

Identifying keyword/phrase and subject terms for a narrowly defined clinical topic (i.e., a single drug for a single indication) is a relatively straightforward process; see for example, "Pharmacotherapy for the Treatment of Cannabis Use Disorder."[22] However, the difficulty of the task is magnified for complex topics, such as: multiple drugs for multiple indications; topics requiring a complicated logic search structure; or diffuse multicomponent interventions (i.e., health services topics); see for example, "Maternal and Fetal Effects of Mental Health Treatments in Pregnant and Breastfeeding Women: A Systematic Review of Pharmacological Interventions."[23] The recognized impact of complexity on the process of conducting systematic review[24] warranted an exploration of variability in TMTs performance for different topic types. The research team prospectively discussed each review as it was identified and decided by consensus if it was simple or complex.

Text Mining Tools Librarian Assignment

TMT librarians were assigned to searches by the SRC to conceal their identities from review teams and other librarians. Not all EPC librarians have access to the Ovid Platform, and some prefer to use PubMed.gov for conducting searches. For the purposes of this study, the content of these two versions of the database were considered equivalent. Thus, when MEDLINE appears in the report text, it indicates the librarian searched either PubMed.gov or Ovid MEDLINE Epub Ahead of Print, In-Process & Other Non-Indexed Citations or Ovid MEDLINE ALL. All PubMed platform searchers conducted their searches in the Legacy PubMed interface.

To avoid confounding due to the differences in these platforms (in terms of search construction and functionality), if the UP librarian searched PubMed, then we chose a TMT librarian who also searched PubMed; similarly, if the UP librarian searched Ovid MEDLINE, then we chose a TMT librarian who also searched Ovid MEDLINE. Some of our chosen TMTs are designed to receive and export information using PubMed syntax only. Because the Ovid platform uses different syntax, Ovid users developed methods to work around this either by inputting a list of PMIDs or recreating a simple PubMed MeSH search in those tools.

Outcome Assessment and Data Analysis

After each librarian developed and conducted her search, the deduplicated retrieved citations were sent to the SRC. The SRC then coded both sets of retrieved citations with unique identifiers to indicate whether a record was unique to the UP search, unique to the TMT search, or retrieved by both. The SRC then sent the review team the combined results of the UP and TMT searches to incorporate into their screening process. After the draft review was completed, the review team sent a list of citations included in the draft report, from any search (UP, TMT, or other sources) for which there was a PMID. Thus, the final report citations may vary from those used in this analysis. The records from this list that had PMIDs were the reference standard included citations and defined how many records from each search were included in the final synthesis. Sensitivity and NNR were calculated against the reference standard for each review for each approach. The complete data tables are in Appendix B.

For each study, both librarians completed a prospectively designed tracking sheet (see Appendix A), indicating the number of hours spent, the MEDLINE platform searched (PubMed or Ovid), and the total number of citations found after deduplication. Time spent was recorded in

total and by specific task: MeSH term generation, keyword phrase generation, and strategy evaluation.

Quantitative Assessment

To operationalize KQ1, we summed the number of hours each librarian spent on each aspect of the search and overall.

To operationalize KQ2a, we calculated sensitivity as identified included / (identified included + not identified included). For KQ2b, we calculated precision as identified included / total citations retrieved; and NNR as 1 / precision. We operationalized these variables as described in Tables 1 (UP searches) and 2 (TMT searches). Because of the small sample size, the analysis was limited to descriptive measures, including means, standard deviations, and 95% confidence intervals, and formal statistical testing was not performed.

Table 1. Usual practice search 2x2 table format for sensitivity and NNR calculations

	Citations Included in Draft Report	All Citations From Both Searches Not Included in Draft Report
Citations identified by UP search	Identified Included: Draft report included citations found by UP search	Identified Excluded: Draft report excluded citations found by UP search
Citations not identified by UP search	Not Identified Included: Draft report included citations not found by UP search	Not Identified Excluded: Draft report excluded citations not found by UP search

Abbreviations: NNR= number needed to read; UP= usual practice.

Table 2. Text-mining search 2x2 table format for sensitivity and NNR calculations

	Citations Included in Draft Report	All Citations From Both Searches Not Included in Draft Report
Citations identified by TMT search	Identified Included: Draft report included citations found by TMT search	Identified Excluded: Draft report excluded citations found by TMT search
Citations not identified by TMT search	Not Identified Included: Draft report included citations not found by TMT search	Not Identified Excluded: Draft report excluded citations not found by TMT search

Abbreviations: NNR= number needed to read; TMT = text-mining tools.

Qualitative Assessment

In the tracking sheet, TMT librarians identified the tool(s) used and answered additional qualitative questions about their process. We elicited TMT librarian comments for some previously unaddressed issues. In order to understand how TMT librarians were using the tools, we asked for a brief description of the methods used in creating the text-mining search. To evaluate whether the seed set of citations used in the TMTs was sufficient and unbiased, we asked for the number of known citations used in the seed set (using predefined response categories) and the TMT librarian's estimation of the representativeness of the seed set. The later response categories included: "overly comprehensive," meaning that it turned up a large number of clearly irrelevant terms; "perfectly balanced," meaning that the terms retrieved appeared sufficient to cover the entire topic but did not include a large number of irrelevant terms; or "subset of vocabulary terms," meaning that the terms returned did not sufficiently cover the topic and the librarian had to add relevant terms identified using different methods. To better understand how the tools could be used, we asked when the software offers multiple types of

analyses, which one(s) were used and why and were any strategies developed to optimize information gleaned from results.

Results

We recruited three organizations to participate in the study: the AHRQ EPC Program, the VA ESP, and the Cochrane Collaboration. We approached 12 review teams, and nine agreed to participate in the study. Two systematic reviews were not included in the final quantitative analyses: one due to protocol violations and the other because our study period ended before the final included citations list was available. The seven evidence syntheses included five de novo systematic reviews, one systematic review update, and one evidence map. These reviews were classified into simple (n=4) or complex topics (n=3). Table 3 lists the review titles and classification, as well as the TMT used. All TMT searches used PubReminer as the keyword/phrase tool. This may result from the comparative ease of using PubReminer. No other tool was consistently used or not used.

In the qualitative section, in addition to the seven reviews in the quantitative analysis, we also included comments from the eighth review whose quantitative results we did not receive before data collection ended.

Table 3. Evidence syntheses included in this study

Review Title	Classification	TMT Used
Aromatherapy and Essential Oils: A Map of the Evidence[25]	Simple	PubReminer, Yale MeSH Analyzer, Carrot2
End-stage Renal Disease and Depression: A Systematic Review*	Simple	PubReminer, Carrot2, VOSviewer
Gulf War Illness - A Systematic Review of Therapeutic Interventions and Management Strategies[26]	Simple	PubReminer, Carrot2
Pharmacotherapy for the Treatment of Cannabis Use Disorder[22]	Simple	PubReminer, Carrot2
Pharmacy Provision of Medical Abortion Care**	Simple**	PubReminer, AntConc MeSH on Demand, VOSviewer
Noninvasive Nonpharmacological Treatment for Chronic Pain: A Systematic Review[27]	Complex	Yale MeSH Analyzer, PubReminer, VOSviewer
Management of Colonic Diverticulitis[28]	Complex	PubReminer, MeSH on Demand, Yale MeSH Analyzer, Carrot2
Maternal and Fetal Effects of Mental Health Treatments in Pregnant and Breastfeeding Women: A Systematic Review of Pharmacological Interventions[23]	Complex	PubReminer, Yale MeSH Analyzer, Carrot2

*Only available on the VA Intranet

**Only qualitative results have been analyzed due to nonavailability of quantitative results at close of data collection.

Participating Librarians

All six participating librarians (UP and TMT) have a master's degree in library science and worked within the EPC Program at the beginning of the study. There was a large overlap, with several librarians serving as both a UP and TMT librarian on different projects. The librarians who served as only UP or TMT do not differ in any substantial way from each other or from the librarians who performed both UP and TMT searches. Table 4 presents summary descriptive data on participating librarians. No TMT librarian was particularly familiar with any topic in a way

that could bias the results. Peer reviews of all the search strategies elicited suggestions for keywords or MeSH terms in one UP search and one TMT search.

Table 4. Participating librarian/information specialist characteristics

Characteristic	Number of Librarian/Information Specialists
Number of Years as an Evidence Synthesis Librarian	Mean 9.8 years (range 6–15 years)
Number of Years as Professional Librarian	Mean 15.6 years (range 9–20 years)
Graduate Degree: MLS	1
Graduate Degree: MLIS	3
Graduate Degree: MSLIS/MSLS	2
Additional Graduate Degree	1 MPH

Abbreviations: MLS= Master of Library Science; MLIS= Master of Library and Information Sciences; MSLIS/MSLS= Master of Science in Library and Information Science/Master of Science in Library Science; MPH= Master of Public Health.

Quantitative Results

The quantitative results are based on seven reviews, four of which were classified as simple and three complex. We present the across-study summary data in Table 5. Please see Appendix B for the data tables upon which Tables 5–6 and Figures 3–8 are based.

Table 5. Comparison of UP and TMT searches by review classification

Review Type	Metric	Time (hours)	Sensitivity (percent)*	NNR (citations)*
All Reviews (N=7): UP search	Mean	12	92.0	83
	SD	8	9.4	34
	95% CI	6.0 to 18.2	85.0 to 99.0	58 to 108
All Reviews (N=7): TMT search	Mean	5	84.9	90
	SD	2	14.2	68
	95% CI	4.0 to 6.7	74.4 to 95.4	39 to 141
Simple Reviews (n=4): UP search	Mean	10	96.0	89
	SD	0.5	4.0	38
	95% CI	9.9 to 10.9	92.1 to 99.9	51 to 126
Simple Reviews (n=4): TMT search	Mean	5	92.3	111
	SD	2	2.5	82
	95% CI	2.9 to 7.1	89.8 to 94.8	31 to 192
Complex Reviews (n=3): UP search	Mean	14	86.7	76
	SD	13	11.6	25
	95% CI	0.5 to 29.3	73.6 to 99.8	48 to 104
Complex Reviews (n=3): TMT search	Mean	6	75.0	62
	SD	1	17.0	21
	95% CI	4.5 to 7.2	55.8 to 94.2	38 to 86

*Average of sensitivity, NNR that was calculated for each review.
Abbreviations: NNR = number-needed-to-read; SD = standard deviation; TMT = text-mining tools; UP = usual practice.

Note: NNR is rounded to the nearest whole number for ease of interpretation.

Subgroup Analysis

We included both de novo reviews and a systematic review update in our study. Update searches often require significant reworking of the search strategy and the original review is used as a source of seed citations. However, in the case of R5 (an update search) the UP librarian reused the original search, leading to discrepant results. We therefore did a subgroup analysis without R5, recalculating the sensitivity, NNR, and time spent to evaluate its effect on the all reviews and complex review results. We present the results in Table 6. Overall, removing that review increased mean sensitivity and NNR for the UP and TMT process. It also increased the average time required for the UP search but did not affect the average time required for the TMT search.

Table 6. Average UP and TMT time, sensitivity, and NNR without R5

Review Type	Metric	Time (hours)	Sensitivity (percent)*	NNR (citations)*
All Reviews (N=6) UP search	Mean	14	95.6	85
	SD	SD= 8	SD= 3.4	SD= 36
	95% CI	5.9 to 22.0	92.1 to 99.2	47 to 123
	Mean Difference	**2 additional hours**	**+3.6**	**2 additional citations**
All Reviews (N=6) TMT search	Mean	5	90.4	93
	SD	SD= 2	SD= 4.7	SD= 73
	95% CI	3.4 to 7.5	85.5 to 95.3	16 to 170
	Mean Difference	**No difference**	**+5.1**	**3 additional citations**
Complex Reviews (n=2) UP search	Mean	21	94.9	77
	SD	SD= 11	SD= 1.2	SD= 30
	95% CI	78.8 to 121.1	83.7 to 100	193 to 347
	Mean Difference	**6 additional hours**	**+8.2**	**1 additional citation**
Complex Reviews (n=2) TMT search	Mean	6	86.6	56
	SD	SD= 1	SD= 5.7	SD= 24
	95% CI	5.0 to 17.5	35.7 to 100	163 to 275
	Mean Difference	**No difference**	**+11.6**	**6 fewer citations**

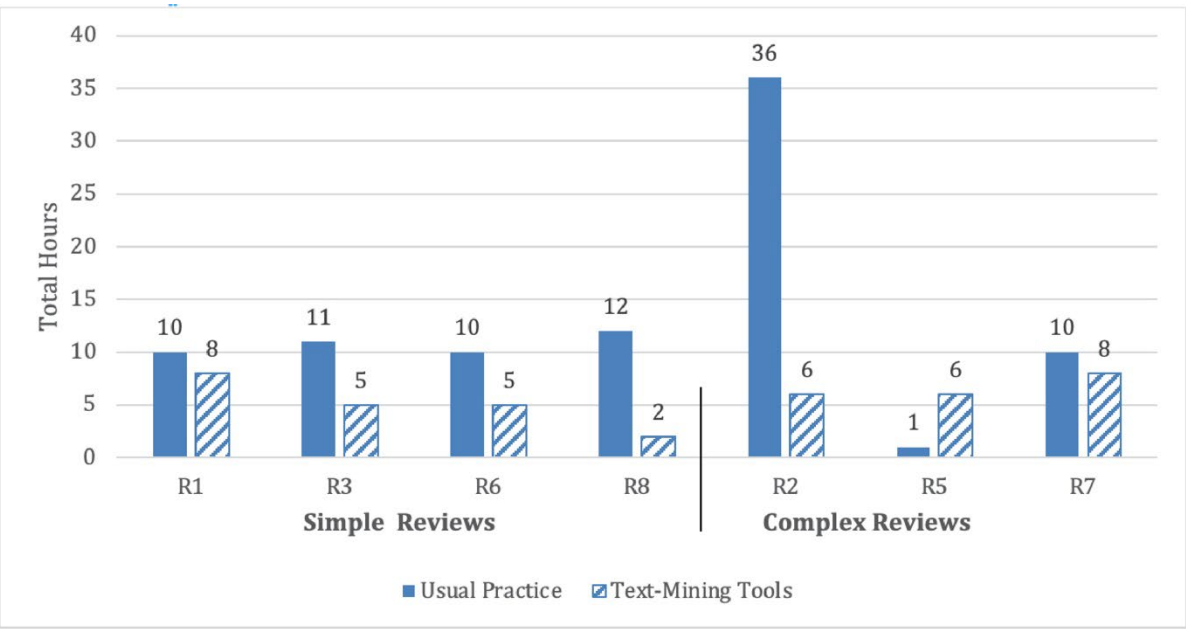
*Average of sensitivity, NNR that was calculated for each review.
Abbreviations: CI = confidence interval; NNR = number-needed-to-read; R = review; SD = standard deviation; TMT = text-mining tools; UP = usual practice.
Note: NNR is rounded to the nearest whole number for ease of interpretation.

Key Question 1: Time Developing Search Strategy (Hours)

The average number of hours to create the search by UP librarians was 12 hours (SD 8 hours), compared to 5 hours (SD 2 hours) for the TMT librarians. Figure 3 shows the total time spent by each librarian on each review, across all three tasks (keyword/phrase, MeSH terms, and strategy evaluation), classified by simple and complex review types. In all but one review (R5), the UP search took more time than the TMT search. The R5 UP librarian reported a very short time because the project was a systematic review update and the librarian ran the existing strategy, while the TMT librarian edited the search with new terms. Figure 4 breaks down time spent across simple and complex reviews by task for each librarian. The time savings does not

come from any single task, as (except for finding MeSH terms for simple reviews) TMT librarians spent less time on each task than UP librarians. The "other" category accounts for time spent in other search activities, such as attending review team meetings for usual practice librarians or learning how to use a text-mining tool for text-mining librarians.

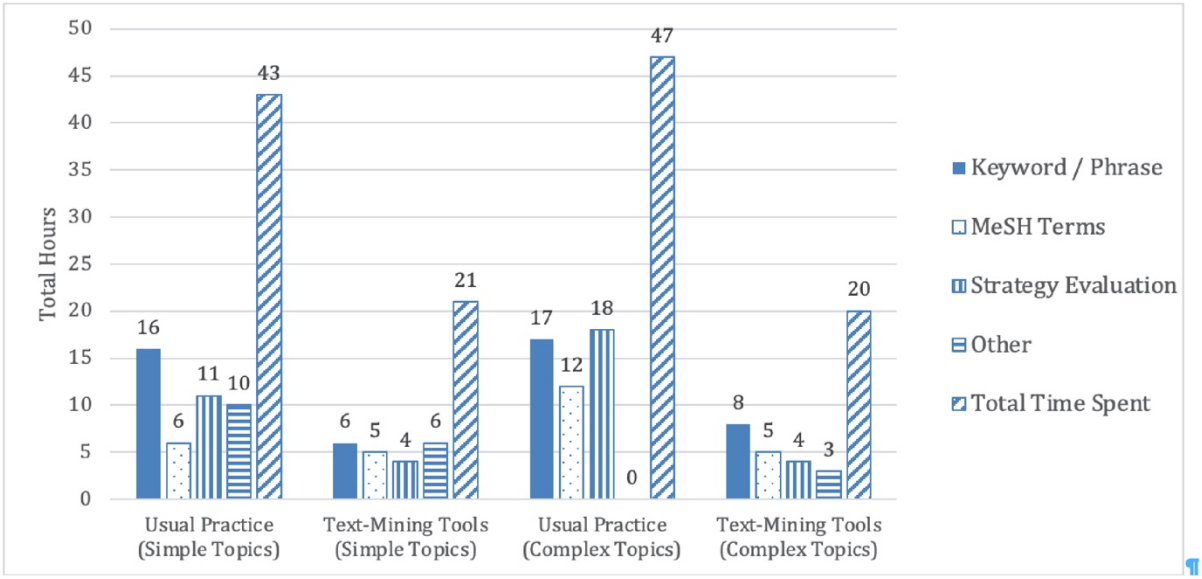
Figure 3. Total number of hours spent on each review by search approach, grouped by classification



Abbreviations: R = review.

Note: Time is rounded to the nearest whole number because the data were imprecise. R4: only qualitative results have been analyzed due to nonavailability of quantitative results at close of data collection.

Figure 4. Total number of hours spent by search step, grouped by type of review and search approach



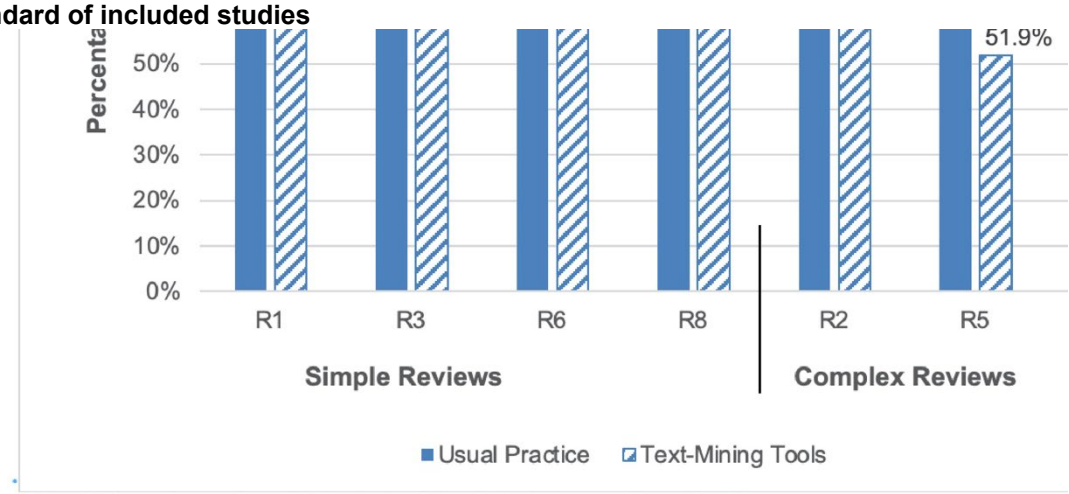
Abbreviations: MeSH = Medical Subject Headings.

Note: Time is rounded to the nearest whole number because the data were imprecise.

Key Question 2a. Yield of Relevant Citations (Sensitivity)

Across all reviews, we found that the UP searches appeared to be more sensitive, with an average sensitivity of 92 percent, while the TMT searches had an average sensitivity of 84.9 percent (calculated on the average of sensitivities, rather than the sensitivity across individual projects). See Figure 5 for results by review, classified by simple and complex review types. Overall, between 5 and 19 unique relevant citations were identified using the UP approach, and between 1 and 4 additional citations were identified using the TMT approach (Figure 8).

Figure 5. Sensitivity of UP and TMT searches by review and classification, compared to the reference standard of included studies



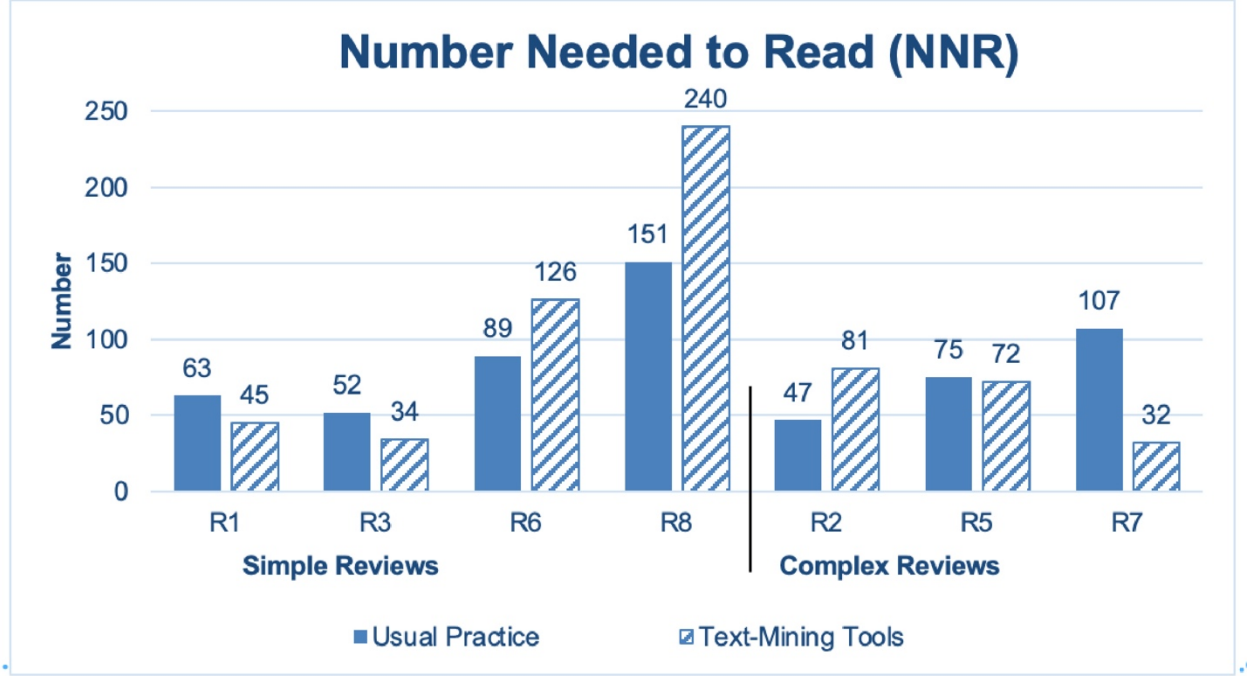
Abbreviations: R = review; TMT = text-mining tools; UP = usual practice.

Note: R4: only qualitative results have been analyzed due to non-availability of quantitative results at close of data collection.

Key Question 2b. Burden of Excess Citations (NNR)

The overall mean NNR results were 83 (SD 34) for the UP librarian and 90 (SD 68) for the TMT librarian, a mean difference of seven more for the TMT librarian (see Figure 6, results grouped by simple and complex review types). NNR was calculated by combining averages, rather than by the average NNR across individual projects.

Figure 6. NNR of UP and TMT searches by review and classification



Abbreviations: TMT = text-mining tools; UP = usual practice.

Note: NNR is rounded to the nearest whole number for ease of interpretation. R4: only qualitative results have been analyzed due to nonavailability of quantitative results at close of data collection.

Key Question 3. Identification of Irrelevant Records During Search Strategy Evaluation

This objective was not addressed in the quantitative results but is discussed below in the qualitative results section.

Key Question 4. Simple Versus Complex Review Topics

Time Developing Search Strategy (Hours)

As shown in Figures 3 and 4, TMT librarians saved more time on complex reviews (average 8 hours saved) than on simple reviews (average 5 hours saved). In complex reviews, for usual practice there was a large range (1 hour in R5, which was the update review, 10 hours in R7, and 32 hours in R2), while TMT times were more consistent (at 6 hours in R2 and R5 and 8 hours in R7). In simple reviews, the times were more consistent across reviews, ranging from 10 to 12 hours for the UP librarian and 2 to 8 hours for the TMT librarian. The average time for simple reviews is not as much less than we had expected, possibly because simple reviews required

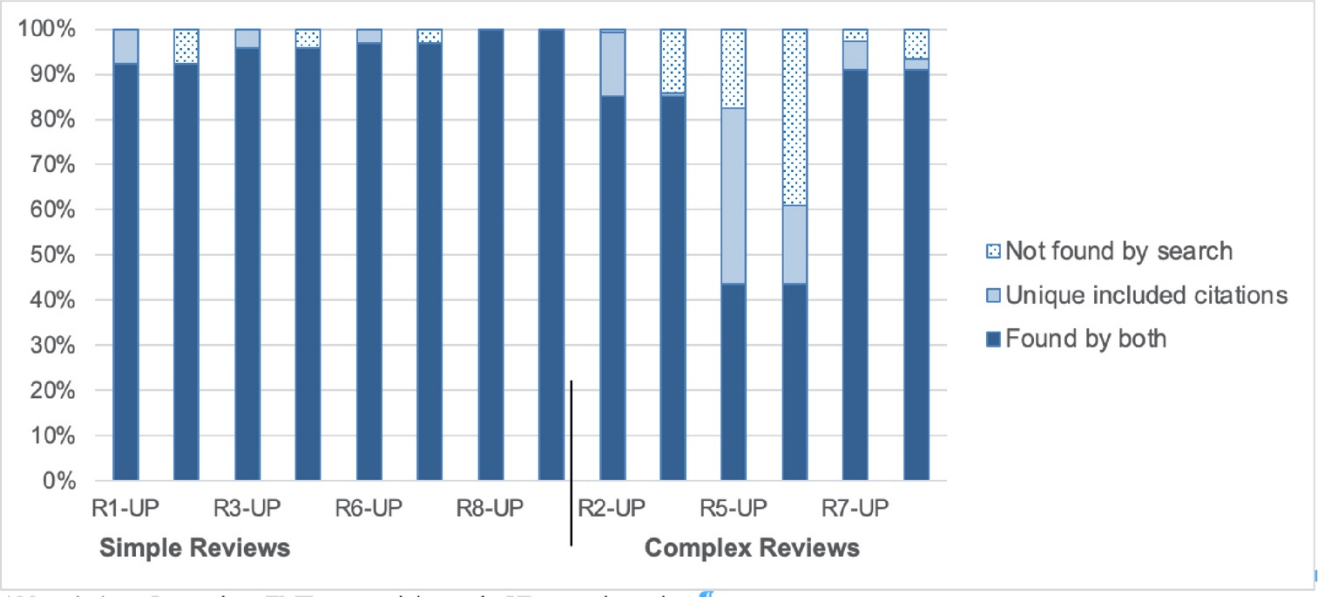
identification of multiple terms for each concept (e.g., a wide variety of aromatherapy interventions) despite having a relatively simple logic structure.

Yield of Relevant Citations (Sensitivity)

In simple reviews, on average, one study identified by the UP search was not identified by the TMT search; the TMT search found no articles not identified by the UP search (a mean sensitivity for UP of 96% compared to 92% for TMT). In complex reviews, the mean sensitivity was lower for both strategies, and the difference in mean sensitivity was greater between the strategies (87% for UP and 75% for TMT) (see Figure 5 for results displayed graphically).

Figure 7 shows that across all reviews, UP searches identified a greater percentage of the relevant (reference standard final included) citations than TMT searches, with the exception of R8 in which both UP and TMT searches identified all included citations. Dark blue bars indicate relevant studies found by both methods, while light blue bars indicate relevant citations found by one method and not the other. In simple reviews, UP searches identified unique included citations in 3 out of 4 reviews (light blue bars) while TMT searches did not retrieve any unique included citations. In complex reviews, UP searches identified more unique included citations than TMT searches across all 3 reviews, but TMT searches did identify at least one unique included citation in each review. The dotted bars indicate the relevant citations not found by a search.

Figure 7. Proportion of reference standard final included citations found by each strategy, by review and classification



Abbreviations: R = review; TMT = text-mining tools; UP = usual practice.

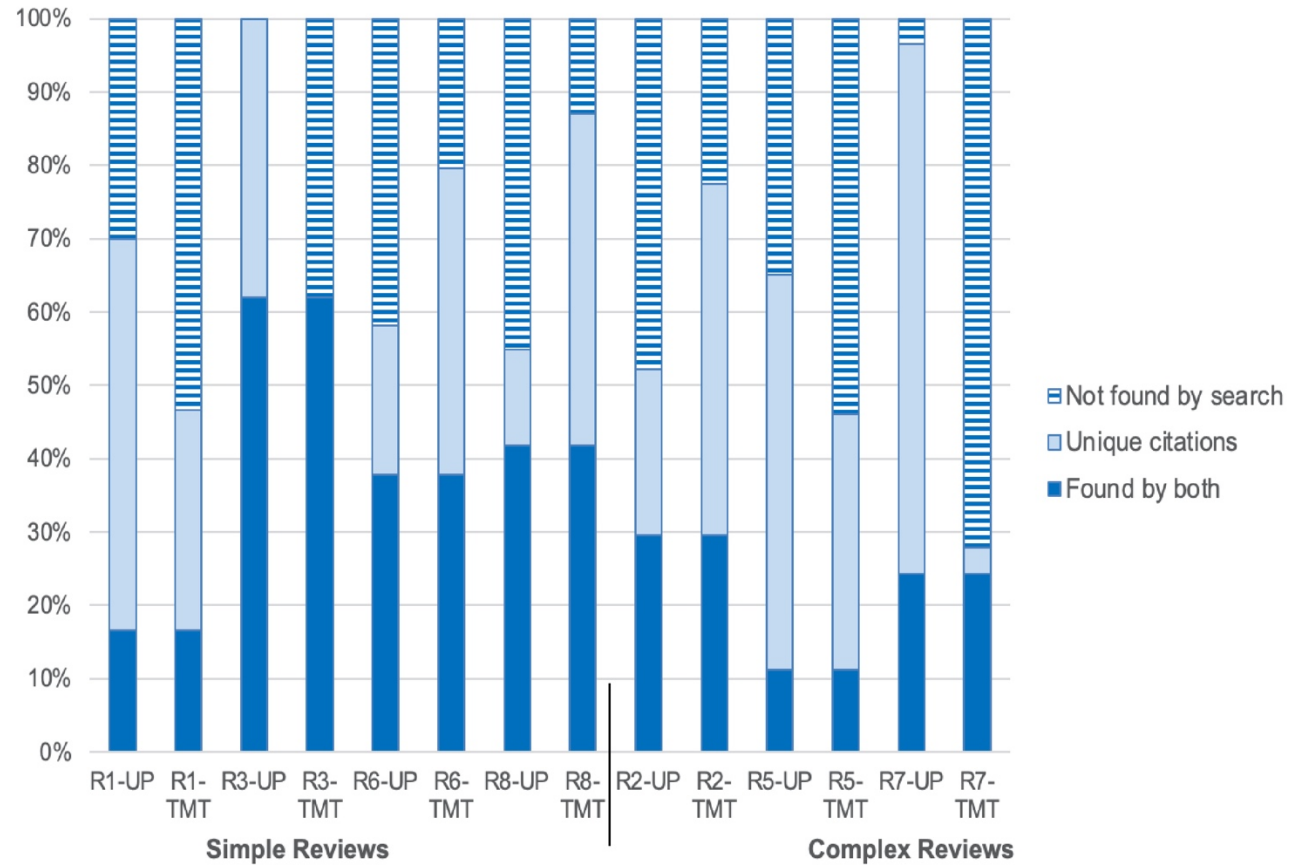
Note: R4: only qualitative results have been analyzed due to non-availability of quantitative results at close of data collection.

Burden of Excess Citations (NNR)

For simple reviews, the TMT search yielded an average 22 more articles per relevant article than the UP search (NNR), but for complex reviews the TMT search yielded an average 14 fewer articles for evidence synthesis team screeners to review per relevant article than the UP search (see Figure 6 for results displayed graphically).

Figure 8 shows that in 4 out of 7 reviews, UP searches retrieved a greater percentage of the total citations (both relevant and not relevant) than TMT searches (dark and light blue bars combined), leading to a greater screening burden. In simple reviews, UP and TMT searches each retrieved a greater percentage of the citations in 2 out of 4 reviews. In complex reviews, UP searches retrieved a greater percentage of the citations in 2 out of 3 reviews. Overlap in citations (dark blue bars) found by both searches ranged from a low of 11.2 percent (R5) to just over 62 percent (R3). This overlap of irrelevant citations retrieved by the different search methods is smaller than we would have expected. The combined mean NNR of UP and TMT searches is 123, indicating that a combined UP and TMT search may lead to greater screening burden. Please see Appendix B for complete data (Figure 7 is column 2 and Figure 8 is column 3).

Figure 8. Proportion of total citations (included and excluded) found by each strategy, by review and classification



Abbreviations: R = review; TMT = text-mining tools; UP = usual practice.

Note: R4: only qualitative results have been analyzed due to non-availability of quantitative results at close of data collection.

Qualitative Results

All the TMT librarians had limited experience using TMTs when the study began. Due to the brevity and small number of comments a full content analysis was not performed; however, we present some topics with selected, edited responses below. Full qualitative comments are available in Appendix C.

Identification of Irrelevant Records During Search Strategy Evaluation Step

The majority of qualitative comments suggest that TMTs most often did not identify irrelevant concepts for removal from the search strategy. However, one TMT librarian reported that they were helpful:

> "Once I generated a map of the results from my MEDLINE strategy (version 1.0) I saw an unrelated cluster of articles (looked like an appendage) about methotrexate. I looked up the term and discovered it is an abortifacient. I removed this MeSH term from the search strategy in version 1.1 which removed about 100 records. I retested the strategy, and the new version still found all of my known items. I exported the revised results to VOSviewer to generate a new map. The new map didn't have this methotrexate cluster. The map looked much more condensed (like a football). I interpreted the new map as showing greater precision in the search results."

Evaluating Seed Set for Bias

As described in the methods section, we developed two questions to evaluate whether the seed set of citations used in the TMTs was sufficient and unbiased (Table 6): (1) the number of known citations used in the seed set and (2) the TMT librarian's estimation of the representativeness of the seed set.

The seed sets for simple topics were evenly split between 11–30 citations (n=2) and 31–60 citations (n=2); whereas, the seed sets for complex topics were mixed, including 31–60 (n=1) and 101+ (n=2). No review topics used fewer than 10 or 61–100 known citations, but this is probably because our sample size was small. In Table 6, we report the number of seed set citations, along with TMT librarian comments by review number. Grouping the results by number of known citations suggests that when there are more known seed set citations the vocabulary terms derived from a TMTs analyses are likely more representative of the review topic. For reviews with fewer than 60 known citations, the variety of responses suggest wide variability in the usefulness of the citations for use as a keyword or MeSH term development set.

Table 7. Number of known citations used in text-mining seed set by review with comments

Review	No. seed set citations*	Representativeness of terms derived from known citations used in TMTs analysis**
R4 (Simple)	11–30	Overly comprehensive (lots of junk terms). **Comment**: It was difficult to determine how representative the known citations were of the topic area.
R6 (Simple)	11–30	The citations may have been a little broad, but generally seemed good.
R8 (Simple)	11–30	Subset of vocabulary terms (had to supplement elsewhere) **Comment**: I had to add a number of terms that were not identified through TMT, probably because there were so few seed citations.
R1 (Simple)	31–60	Perfectly balanced (had all needed terms without a lot of junk).
R3 (Simple)	31–60	This was a very clean search.

Review	No. seed set citations*	Representativeness of terms derived from known citations used in TMTs analysis**
R7 (Complex)	31–60	Subset of vocabulary terms (had to supplement elsewhere). **Comment**: The text-mining tools did help to identify some relevant keywords & MeSH headings however there were many irrelevant results to wade through to find a small number of relevant terms. Seeing the terms out of context in PubReMiner for unfamiliar topic areas was less than helpful and required additional follow-up to determine if keywords were relevant or not. Would have preferred to use a tool where keywords could be viewed in groups (bigrams, trigrams).
R2 (Complex)	101+	Perfectly balanced (had all needed terms without a lot of junk). **Comment**: There was a lot of junk, but this is a complex topic, so I think fewer citations would have led to gaps in the search.
R5 (Complex)	101+	Perfectly balanced (had all needed terms without a lot of junk). Subset of vocabulary terms (had to supplement elsewhere). **Comment**: I would say it was somewhere between these two actually…this was a complex search and it was also an update search, so I had the list of included citations and the existing search to begin. I was also aware of the possibility that exclusively using the existing include list might bias the results (e.g., to vocabulary being used at the time of the previous review (intervening semantic drift) or if the original review search was not inclusive enough), so once a fairly robust search was established, I then re-ran the TMT searches to determine if there were other additional text and MeSH terms to consider. I also experimented with creating an initial search with all the known terms, finding systematic reviews, meta-analyses, trials with those words in the title, taking a sample to plug into PubReMiner and VOSviewer. I thought this approach might work for new review topic searching as well.

*Responses to the question "What number of known citations did you use in the seed set? (up to 10, 11-30, 31-60, 61-100, 101+)"
** Responses to the question "Did the known citations used for text-mining analysis represent the diversity of vocabulary terms or a subset of terms used in this area of research? (overly comprehensive [lots of junk terms], perfectly balanced [had all needed terms without a lot of junk], subset of vocabulary terms [had to supplement elsewhere])"
Abbreviations: R= review; TM= text mining; TMTs= text-mining tools

Developing Methods for Using Text-Mining Tools in Systematic Review Searches

We collected initial experiences of using freely available TMTs to expand our field's real world understanding of how to approach this new class of search tools. Below are three selected and edited quotes on search techniques, followed by comments on specific tools by TMT librarians. See Appendix C for the complete comments.

> "I used the list of known relevant systematic reviews from the provided Excel spreadsheet to create a list of PMIDs to enter into PubReMiner & the Yale MeSH Analyzer. The results were reviewed to identify relevant keywords & MeSH headings. Several of these keywords & MeSH headings were then entered into PubReMiner & MeSH on Demand to identify additional relevant terms. In addition, I also input various portions of text from the Systematic Review Protocol document into MeSH on Demand to identify other relevant headings."

> "The text-mining tools were a great compliment to usual practice and going forward I plan to utilize them more often during the strategy development period. However, I would not

feel comfortable designing a strategy solely using text-mining, as there are many irrelevant results returned and the lack of context for unfamiliar topic areas requires additional follow-up. While working on this project I developed a routine of flagging potentially relevant keywords & headings, which then required me to do additional research to see if they were in fact useful for the strategy."

"I generated my seed set by: (1) using references in the protocol, (2) running a quick PubMed query and looking at related references, (3) identifying review articles on the topic and then adding their included citations. I'm unsure if there are other more effective methods to identify test articles, or if my approach was appropriate?"

Comments on Text-Mining Tools Overall and by Specific Tool

We were interested in gleaning real-life experiences using the study's TMTs (i.e., what works, what were particularly easy/difficult tasks, etc.). Most TMTs are developed to use default PubMed record output, so users of other platforms must create work arounds. One TMT librarian commented that in general the tools increased the complexity of the process:

"Time was added to search process to tweak and troubleshoot issues related to constraints of the TMTs (character limits, output limits, search input formatting issues, etc.). For example, while using some of the tools, searches had to be tweaked several times because the output was too large for the tool to handle. Related to this issue, I had to very narrowly limit the search date range while performing keyword searches related to CT [computed tomography] imaging. Could this very narrow search date window negatively affect process/results?"

AntConc (Keyword/Phrase Tool | Downloadable Software)

AntConc is a linguistic tool useful for identifying high-frequency/occurrence keyword terms and bound phrases. One TMT librarian had the following concern:

"I used [this tool] on a different computer than I had previously (MacBook) and had to override some security settings to get the application to run… (1) Generated a text file from the title and abstracts and imported into AntConc. (2) Analyzed the "Word List" tab, and then (3) selected key terms to see their context via the "Concordance" tab. I found it difficult to determine cut-off threshold for occurrence (selected 4 and did not look below this number). I viewed some "Clusters/N-Grams" for key terms (e.g., abortion). It was difficult to determine when a phrase search for a term should be used instead of a single term (i.e., I knew the phrase "medical abortion" occurred 27 times in my corpus – should I use this phrase in my search or just the term "abortion"?)."

MeSH on Demand (Subject Term Tool | Web-Based Tool)

MeSH on Demand is a National Library of Medicine (NLM) designed to analyze end-user input text and suggest NLM medical subject headings (MeSH). One librarian commented:

"I used a section of text from the protocol to identify subject headings. Note: limited to 10,000 characters – had to select a section of the protocol as full protocol was over 20,000 characters. Very quick to analyze text (took seconds to get my subject headings). No

additional information about term explosions, so I still had to look up each MeSH term. Articles identified as being relevant were not about pharmacists – I did not add any to my seed set of articles."

PubReMiner (Keyword/Phrase and Subject Term Tool | Web-Based Tool)

PubReMiner analyzes end-user defined PubMed records and generates frequency tables of bibliographic record fields (title, abstract, MeSH terms, journal, etc.). Two librarians commented on using PubReMiner:

"Still had to generate a PubMed query. I tried to build a query using my seed set of articles with their PMIDs, but this didn't work (or I couldn't get this to work), so I generated a quick query string: (abortion or mifepristone or misoprostol) AND (pharmacist or pharmacists OR pharmacy OR pharmacies OR chemists). This query resulted in 639 references. Identified additional MeSH terms not found with PubReMiner. Difficult to determine a cut-off threshold for occurrence (selected 10 and did not look below this number)."

"I uploaded the list of PMIDs and ran the search. I then selected the following fields from the right-hand side of the screen to manually adjust the search: MeSH, Substance, and WORD TI_AB. These seemed the most useful ways to focus the search for developing terms. In looking over the results term occurrence was the most important factor in selecting potential terms to test in the strategy."

CARROT2 (Visualization Tool | Downloadable Tool)

Carrot2 is a thematic clustering algorithm for small collections of documents. One librarian opined:

"…I still find Carrot2 more useful for this [strategy evaluation] stage as it clumps the references into topics, but I did like looking at the way terms were connected in VOS."

VOSviewer (Visualization Tool | Downloadable Tool)

VOSviewer visualizes connections in bibliographic networks. One librarian noted:

"For VOSviewer, it took several searches to figure out that the maps I prefer to use are constructed by downloading two Research Information Systems (RIS) files (one including TI, AB, author keywords fields; and the other including MeSH terms and maybe registry name fields). In VOSviewer under file, map, create, create map based on text data, RIS tab (keep ignore fields checked), I could upload each file in turn. Once the map was created, I was looking for non-relevant keywords or MeSH terms that appear as larger bubbles, since these represent the number of occurrences of the term. Sometimes (not always) this has been ideal for finding candidate terms to NOT out to make the search more specific. Of course, one has to test these before removing them from the search..."

"I tried other mapping displays but found the map based on text data the easiest to 'read'. I also found that I had to split the results into two files, one for keywords in the title/abstract

and the other for MeSH terms. When I didn't do that, the MeSH terms dominated the display to the detriment of everything else."

Discussion

Summary of Findings

In this project comparing keyword and subject searches done with and without TMTs, we found that the tools did decrease the time needed to develop keyword and subject term strategies compared to conventional approaches. It took experienced librarians less time to develop the TMT searches than UP searches, and this time savings carried across all tasks. This time savings may have come from the ability of TMT librarians to use word frequency tables rather than having to extract the information from texts, or the fact that TMT conceptual groupings might identify "red herrings" quickly and allow for their elimination. It may also stem from UP librarians involvement in the search process for a longer period of time (i.e., attending team meetings and conducting topic and scope refinement searches).

It is uncertain whether TMTs can improve the evaluation step. Only a single librarian reported that she found concepts that could be eliminated using TMTs, while other librarians reported using these tools but not eliminating any terms based on the results. This suggests that using TMTs in the evaluation step may be worthy of additional study.

In terms of whether TMTs improve search recall (sensitivity), we found that over all the UP search was slightly more sensitive across all projects than the TMT search. Neither UP nor TMT searches were perfectly sensitive across all sample reviews, reinforcing that other supplementary search techniques, including handsearching, are still important for comprehensive systematic review searches.

In regard to whether the type of review topic (complex versus simple) makes a difference in the performance of TMTs, we found that there may be more of a role for TMTs in complex reviews. For simple review topics (i.e., single indication-single drug) TMT searches resulted in no unique relevant citations (and missed one relevant study in three of four reviews), but reduced time spent in search design. For complex review topics (e.g., multicomponent interventions) TMT searches identified some unique includable citations and reduced time spent in search strategy development but missed between four and nine relevant citations identified by the UP search.

Finally, TMT librarians' evaluations of the tools indicate they used a variety of combinations of tools and techniques to complete the searches. The most effective and efficient text-mining methods/processes (aka 'best practices') for searchers are still in a formative stage, and we do not feel that the evidence has reached a point where specific guidance on the use of freely available TMTs is meaningful. The research base on these tools is very small, with very few studies on any specific tool. Additionally, the TMTs are evolving so guidance now may not remain relevant. Best practices in the use of TMTs is an area deserving future research.

Strengths and Limitations

This project is, to our knowledge, the first to use a variety of off-the-shelf, freely available tools to evaluate the contribution of TMTs to search strategy design, using professional librarians with a great deal of experience and peer review to ensure all searches (both UP and TMT) were of high quality.

Nevertheless, this study has limitations. For one, the small sample of reviews makes it hard to draw conclusions across projects, and especially to draw conclusions about its performance in specific types of projects. Future research should evaluate these tools across a larger sample and

variety of reviews, including simple and complex reviews, types of reviews (e.g., rapid reviews, scoping reviews, etc.), and reviews that span different disciplines. Additionally, not every project team we approached agreed to participate in the study. It is possible that the search processes or review types of those who participated differ in a meaningful way from those who did not. Finally, the sample size of librarians was small and all had years of experience, so we were not able to test whether TMTs level the playing field between inexperienced and experienced librarians.

Another limitation comes from our intent to reflect real-world use of freely available tools, as well as our reluctance to pre-emptively establish "best practices" in TMT use. Thus, we gave very little guidance on how specific TMTs should be used or even which tools should be used from our list. This led to a variety of approaches, which affected the quantitative results. For example, TMT librarians combined usual practice with text-mining to varying degrees, making it hard to narrow down the actual effect of text-mining alone. In addition, the peer review process, while ensuring the searches were of comparable quality, may have led to the addition of terms that were not included based on text mining only.

Another limitation has to do with variations in how librarians operationalized time recording and search development. This led to some outliers in the results. It also did not allow us to calculate how much time the UP and TMT librarians spent identifying seed citations. In addition, one study was an update (R5), while the others were de novo reviews, which may limit the utility of the time estimate for that search in the overall time analysis. For this reason, we did a subgroup analysis without R5 to determine its effect on the sensitivity, NNR, and time spent across all reviews and complex reviews.

Finally, the librarians were relatively inexperienced with the tools at the beginning of the study and overall found that the tools took a lot of time to learn and were at times less functional than hoped. The librarians used a variety of tools, so we can't comment on whether a particular tool is better than another. Librarians with more experience and expertise using specific TMTs might achieve better results.

Implications for Practice and Relevance to Existing Literature

This project expands on Hausner et al.'s previous work.[5-7] However, the Wordstat program used in their research is subscription based and is therefore not universally available to librarians in the EPC program. Instead, we used freely available web-based TMTs, which increases the applicability of our findings. Our study looked at the utility of incorporating existing text-mining software (AntConc, PubReMiner, MeSH On Demand, Yale MeSH Analyzer, Carrot2, and VOSviewer) into the process of search strategy design. Overall, we found that these methods were slightly less sensitive, but led to a reduction in time spent developing the search and may reduce the burden on the team in the number of citations that have to be screened.

We were specifically interested to see whether these tools would be helpful in the context of reviews undertaken by the EPC program, which tend to be complex and multicomponent, as compared to simpler one intervention-one indication reviews. We found that incorporating TMTs for complex topics may allow the searcher to find terms that identify citations not found by what Hausner et al. refer to as the conceptual search.[6] Our results are similar to those of Hausner et al., who found that in a sample of Cochrane reviews, an objective search using a text-mining program in Wordstat had similar sensitivity to the original searches.[5] However, unlike Hausner et al., we found that the UP searches appeared to be more sensitive than the TMT searches (but not statistically significantly so), although both UP and TMT searches missed relevant citations. This

may reflect differences in the way that we applied text-mining—via the integration of freely available tools with little guidance, as opposed to via an algorithm and statistical package—or it may have to do with the high rigor of the UP searches in this project. Future research should focus on comparing text-mining tool functionality and usability, as well as establishing guidelines and best practices for librarians using freely available TMTs.

One issue that has not been addressed in the literature to date is how to evaluate known relevant citations used in the seed set for their representativeness of the literature, in terms of vocabulary used, known interventions, MeSH terms assigned, and so on. This type of evaluation is important to prevent what Eustace dubbed, "technology-induced bias" (i.e., if the known citation seed set is biased in one or more ways, it is very likely the TMTs output will also be biased).[29] In addition, the ideal size of the seed sets for systematic review searches is presently unknown, nor have methods been developed to aid in evaluating them for bias. Nevertheless, our qualitative analysis suggests that when there are more known citations as a percentage of the literature base in the seed set, the results of the TMTs are more representative of the breadth of the review topic.

Both the Cochrane Handbook[9] and the AHRQ Methods Guide[8] recommend that systematic review searches be comprehensive, striving to identify all relevant citations. Based on the findings of our study, text-mining technology is not ready to be used as the sole process for developing systematic review searches, but the time savings in search design and relatively high sensitivity for complex reviews suggest that this technology may be useful in reviews that do not require maximum sensitivity, such as rapid or scoping reviews. In addition, TMTs are useful in combination with usual practice to find citations missed by the usual search process. Nevertheless, adding a TMT step to the UP search strategy development process will increase the screening burden (NNR) and time required for search development.

Conclusions

Overall, this study found that incorporating TMTs into search strategy development for systematic review projects may reduce time identifying keyword and subject terms but at the cost of potentially decreased sensitivity. For simple topics, TMT searches seemed to have similar (or slightly lower) sensitivity, higher NNR, and required less time than UP searches, but in all cases with overlapping confidence intervals. For complex topics, TMT searches seemed to have lower sensitivity, lower NNR, and required less time than UP searches. Research is needed to improve the utility of off-the-shelf TMTs for use by systematic review search librarians and examine the various ways librarians are using these tools. In addition, research is needed on how to evaluate the corpus of records used by the tools for representativeness (the seed set).

References

1. Ananiadou S, Rea B, Okazaki N, et al. Supporting systematic reviews using text mining. Social Science Computer Review. 2009;27(4):509-23. doi: 10.1177/0894439309332293.
2. Marcos-Pablos S, García-Peñalvo FJ. Decision support tools for SLR search string construction. Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality. 2018:660-7. doi: 10.1145/3284179.3284292.
3. Mergel GD, Silveira MS, da Silva TS. A method to support search string building in systematic literature reviews through visual text mining. Proceedings of the 30th Annual ACM Symposium on Applied Computing. 2015:1594-601. doi: 10.1145/2695664.2695902
4. Tsafnat G, Glasziou P, Choong MK, et al. Systematic review automation technologies. Syst Rev. 2014;3(1):74. doi: 10.1186/2046-4053-3-74. PMID: 25005128
5. Hausner E, Waffenschmidt S, Kaiser T, et al. Routine development of objectively derived search strategies. Syst Rev. 2012;1(1):19. doi: 10.1186/2046-4053-1-19. PMID: 22587829
6. Hausner E, Guddat C, Hermanns T, et al. Development of search strategies for systematic reviews: validation showed the noninferiority of the objective approach. J Clin Epidemiol. 2015;68(2):191-9. doi: 10.1016/j.jclinepi.2014.09.016. PMID: 25464826
7. Hausner E, Guddat C, Hermanns T, et al. Prospective comparison of search strategies for systematic reviews: an objective approach yielded higher sensitivity than a conceptual one. J Clin Epidemiol. 2016;77:118-24. doi: 10.1016/j.jclinepi.2016.05.002. PMID: 27256930
8. Relevo R, Balshem H. Finding evidence for comparing medical interventions. Methods Guide for Effectiveness and Comparative Effectiveness Reviews 2011. doi: https://www.ncbi.nlm.nih.gov/books/NBK53479/pdf/Bookshelf_NBK53479.pd. PMID: 21433408.
9. Lefebvre C, Glanville J, Briscoe S, et al. Chapter 4: Searching for and selecting studies. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al., eds. Cochrane Handbook for Systematic Reviews of Interventions version 6.0: Cochrane; 2019.
10. Glanville J, Wood H. Text Mining Opportunities: White Paper. Canadian Agency for Drugs and Technologies in Health (CADTH). 2018. doi: https://www.cadth.ca/sites/default/files/pdf/methods/2018-05/MG0013_CADTH_Text-Mining_Opportunites_Final.pdf.
11. Paynter R, Bañez LL, Berliner E, et al. EPC methods: an exploration of the use of text-mining software in systematic reviews. Agency for Healthcare Research and Quality. 2016. doi: https://www.ncbi.nlm.nih.gov/books/NBK362044/pdf/Bookshelf_NBK362044.pdf. PMID: 27195359.
12. Stansfield C, O'Mara-Eves A, Thomas J. Text mining for search term development in systematic reviewing: A discussion of some methods and challenges. Res Syth Methods. 2017;8(3):355-65. doi: 10.1002/jrsm.1250. PMID: 28660680
13. O'Mara-Eves A, Thomas J, McNaught J, et al. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Syst Rev. 2015;4(1):5. doi: 10.1186/2046-4053-4-5. PMID: 25588314
14. Stansfield C, Thomas J, Kavanagh J. 'Clustering' documents automatically to support scoping reviews of research: a case study. Res Synth Methods. 2013;4(3):230-41. doi: 10.1002/jrsm.108. PMID: 26053843.
15. McGowan J, Sampson M, Salzwedel DM, et al. PRESS peer review of electronic search strategies: 2015 guideline statement. J Clin Epidemiol. 2016;75:40-6. doi: 10.1016/j.jclinepi.2016.01.021. PMID: 27005575
16. Anthony L. AntPConc. 1.2.1 ed. Tokyo, Japan: Waseda University; 2017.

17. Slater L. PubMed PubReMiner. Journal of the Canadian Health Libraries Association / Journal de l'Association Des Bibliothèques de La Santé Du Canada. 2014;33:106.

18. Cho D. MeSH on Demand Tool: An Easy Way to Identify Relevant MeSH Terms. NLM Tech Bull. 2014;398.

19. Grossetta Nardini H, Wang L. The Yale MeSH Analyzer. New Haven, CT: Cushing/Whitney Medical Library.

20. Stefanowski J, Weiss D. Carrot2 and Language Properties in Web Search Results Clustering. In: Menasalvas E. SJ, Szczepaniak P.S., ed Advances in Web Intelligence AWIC 2003 Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), . Vol. 2663. Berlin: Heidelberg; 2003.

21. van Eck NJ, Waltman L. Software Survey: VOSviewer, a Computer Program for Bibliometric Mapping. Scientometrics 2010;84:523-38.

22. Kondo K, Morasco BJ, Nugent S, et al. Pharmacotherapy for the Treatment of Cannabis Use Disorder: A Systematic Review. VA Evidence-based Synthesis Program. 2019;VA ESP Project #05-225. doi: https://www.ncbi.nlm.nih.gov/books/NBK555373/pdf/Bookshelf_NBK555373.pdf. PMID: 32227801.

23. Maternal and Fetal Effects of Mental Health Treatments in Pregnant and Breastfeeding Women: A Systematic Review of Pharmacological Interventions. Agency for Healthcare Research and Quality. In progress. doi: https://effectivehealthcare.ahrq.gov/products/mental-health-pregnancy/protocol.

24. Noyes J, Gough D, Lewin S, et al. A research and development agenda for systematic reviews that ask complex questions about complex interventions. J Clin Epidemiol. 2013;66(11):1262-70. doi: 10.1016/j.jclinepi.2013.07.003. PMID: 23953084

25. Freeman M, Ayers C, Peterson C, et al. Aromatherapy and Essential Oils: A Map of the Evidence. VA Evidence-based Synthesis Program 2019. doi: https://www.hsrd.research.va.gov/publications/esp/reports.cfm PMID: 31851445.

26. Freeman M, Nugent S, Ayers C, et al. Gulf War Illness: a systematic review of therapeutic interventions and management strategies. VA Evidence-based Synthesis Program In progress;PROSPERO Number: CRD42019155102. doi: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42019155102.

27. Skelly AC, Chou R, Dettori JR, et al. Noninvasive Nonpharmacological Treatment for Chronic Pain: A Systematic Review Update. Agency for Healthcare Research and Quality. 2020. PMID: 32338846.

28. Management of Colonic Diverticulitis. Agency for Healthcare Research and Quality. In progress. doi: https://effectivehealthcare.ahrq.gov/products/diverticulitis/protocol.

29. Eustace S. Technology-induced bias in the theory of evidence-based medicine. J Eval Clin Pract. 2018;24(5):945-9. doi: 10.1111/jep.12972. PMID: 29998588

Abbreviations and Acronyms

AHRQ	Agency for Healthcare Research and Quality
CT	Computed tomography
EPC	Evidence-based Practice Center
IQWiG	Germany. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen
MeSH	Medical subject headings (National Library of Medicine)
NLM	US. National Library of Medicine
NNR	Number-needed-to-read
PMID	PubMed unique identifier number
PRESS	Peer Review of Electronic Search Strategies
R	Review
RIS	Research Information Systems
SD	Standard deviation
SR	Systematic review
SRC	Scientific Resource Center for the AHRQ EPC Program
TMT	Text-mining tool
UP	Usual practice
VA ESP	US. Department of Veterans Affairs Evidence Synthesis Program

Appendix A. Study Tracking Sheet

Table A-1. Example study tracking sheet

Question	Value	Notes
Review Title		
Review Source (AHRQ, VA ESP):		
Review Topic Type		
Usual Practice Librarian		
UP Anonymization Code		
DATE Topic Brief, KQs/PICOTS received by SRC		
DATE changes to KQs/PICOTS received by SRC		Any modifications to the KQs/PICOTs should be sent to the TMT Librarian immediately
DATE Time Log Received		
Usual Practice Keywords		Report in 15-minute increments in column B, i.e. 1.25, 1.5, 1.75
Usual Practice MeSH		
Usual Practice Strategy evaluation		
TOTAL Time Spent		
DATE UP Search Run		
DATE UP Search Results Received by SRC		
Ovid or PubMed Search?		
Strategy Peer-Reviewed		
UP Search Results Deduped TOTAL:		
UP-SRC EMAIL Notes		
DATE TMT assigned and sent KQs/PICOTS by SRC:		
Usual Practice + Text-mining Tools Librarian		See Line 15 for which group (Ovid or PubMed) to assign from
TMT Anonymization Code		
DATE Time Log Received		
Usual Practice+Text-mining Tool(s) Keywords		Report in 15 minute increments, i.e. 1.25, 1.5, 1.75 Please report the TM tools used here
Usual Practice+Text-mining tool(s) MeSH		Please report the TM tools used here
Usual Practice+Text-mining tool(s) Strategy evaluation		Please report the TM tools used here
TOTAL Time Spent		
TMT Qualitative Comments		
DATE TMT Search Run		
DATE TMT Search Results Received by SRC		
Strategy Peer-Reviewed		
TMT Search Results Deduped TOTAL		
TMT-SRC EMAIL Notes		
Review Team Contact Person		*Instruct Team NOT to delete Custom Field 8 *Set up Outlook reminder to contact Team on anticipated date final include list available *Send thank you note to Team for participating

Question	Value	Notes
Team Contact-SRC EMAIL notes		
TOTAL # of Medline/PubMed Citations Sent to Review Team Contact Person		
EndNote file location		
DATE EndNote Library Sent		
DATE TOTAL Number of Citations Sent		
DATE Anticipated final includes list available		
DATE Draft/Final Includes List Received by SRC from Review Team Contact		
UP coded final includes TOTAL		
TMT coded final includes TOTAL		
UP Recall		
UP Precision		
UP Time		
TMT Recall		
TMT Precision		
TMT Time		

Appendix B. Quantitative Data Tables

Table B-1. UP and TMT search results by individual review

Title Code/Topic Classification	UP or TMT	Total Number of Citations Retrieved*	Total Number of Identified Included Studies (reference standard)	Total Time Spent (Hours)	Sensitivity**	NNR***
R 1 (Simple topic)	**UP**	UP only: 622	UP only: 1	10	100.0% (13/13)	63
	TMT	TMT only: 350	TMT only: 0	8	92.3% (12/13)	45
	Overlap	Overlap: 194	Overlap: 12	N/A	N/A	N/A
	Total	TOTAL: 1,166	TOTAL: 13	N/A	N/A	N/A
R 3 (Simple topic)	**UP**	UP only: 478	UP only: 1	11	100.0% (24/24)	52
	TMT	TMT only: 0	TMT only: 0	5	95.8% (23/24)	34
	Overlap	Overlap: 780	Overlap: 23	N/A	N/A	N/A
	Total	TOTAL: 1,258	TOTAL: 24	N/A	N/A	N/A
R 6 (Simple topic)	**UP**	UP only: 1,035	UP only: 1	10	91.7% (33/36)	89
	TMT	TMT only: 2,116	TMT only: 0	5	88.9% (32/36)	126
	Overlap	Overlap: 1,912	Overlap: 23 (other sources 3)	N/A	N/A	N/A
	Total	TOTAL: 5,063	TOTAL: 36	N/A	N/A	N/A
R 8 (Simple topic)	**UP**	UP only: 431	UP only: 0	11	92.3% (12/13)	151
	TMT	TMT only: 1,493	TMT only: 0	2	92.3% (12/13)	240
	Overlap	Overlap: 1,384	Overlap: 12 (other sources 1)	N/A	N/A	N/A
	Total	TOTAL: 3,308	TOTAL: 13	N/A	N/A	N/A
R 2 (Complex topic)	**UP**	UP only: 2,701	UP only: 19	32	93.7% (133/142)	47
	TMT	TMT only: 5,740	TMT only: 1	5	81.0% (115/142)	81
	Overlap	Overlap: 3,547	Overlap: 114 (other sources 8)	N/A	N/A	N/A
	Total	TOTAL: 11,988	TOTAL: 142	N/A	N/A	N/A
R 5 (Complex topic)	**UP**	UP only: 1,176	UP only: 9	1	70.4% (19/27)	75

Title Code/Topic Classification	UP or TMT	Total Number of Citations Retrieved*	Total Number of Identified Included Studies (reference standard)	Total Time Spent (Hours)	Sensitivity**	NNR***
	TMT	TMT only: 762	TMT only: 4	5	51.9% (14/27)	72
	Overlap	Overlap: 245	Overlap: 10 (other sources 4)	N/A	N/A	N/A
	Total	TOTAL: 2,183	TOTAL: 27	N/A	N/A	N/A
R 7 (Complex topic)	**UP**	UP only: 5,976	UP only: 5	10	96.2% (75/78)	107
	TMT	TMT only: 291	TMT only: 2	8	92.3% (72/78)	32
	Overlap	Overlap: 2,012	Overlap: 70 (other sources 1)	N/A	N/A	N/A
	Total	TOTAL: 8,279	TOTAL: 27	N/A	N/A	N/A

Abbreviations: NNR = number-needed-to-read; R = review; TMT = text-mining tools; UP = usual practice.
*Unique UP and TMT search results; overlap search results; and total number of records reviewed
**Sensitivity = n/N. See Tables 1,2 in the Methods/Data Analysis section for more information on the calculation
***NNR = 1/Precision. See Tables 1,2 Methods/Data Analysis section for more information on the calculation.

Table B-2. Comparison of total UP and TMT librarian time spent by activity (in hours)

Activity	UP librarian Simple Topics N=4)	TMT librarian Simple Topics N=4	UP librarian Complex Topics N=3	TMT librarian Complex Topics N=3
Keyword/ Phrase	16	6	17	8
MeSH Terms	6	5	12	5
Strategy Evaluation	11	4	18	4
Other	10	6	0	3
Total Time Spent	43	21	47	20

Abbreviations: MeSH= medical subheading; TMT= text-mining tools; UP= usual practice.

Appendix C. Qualitative Comments From Text-Mining Librarians

Table C-1. Comments from text-mining librarians

Question	Review Number	Answer
Briefly outline the methods used in creating the TMTs search.	R1	1. Read all background material; created a list of PMIDs from the citations included in the TR document 2. Ran that PMID list through PubReMiner ad exported the results to Excel 3. Got rid of MeSH terms used fewer than 2 times 4. Got rid of free text and substance terms used fewer than 5 times 5. Divided terms up into PI[C]O; kept only P and I terms (C was integrated with I) 6. Ran the previous (2014) Cochrane review search terms through PubReMiner 7. Got rid of MeSH terms used fewer than 5 times 8. Got rid of free text and substance terms used fewer than 10 times 9. Divided terms up into PI[C]O; kept only P and I terms (C was integrated with I) 10. Compared 2 PubReMiner lists and prioritized by number of times used. 11. Noted for each term whether it was relevant for P or I then sorted 12. Removed terms that were irrelevant; combined terms that were better searched together or as phrases. 13. Combined the remaining terms to create a search, added terms with reference to 2014 search, the project PICO, and the MeSH database 14. Ran draft search against PMID list to ensure that it captured all test citations 15. Ran final draft through Carrot2 to look for irrelevant terms that could be excluded 16. Added RCT filter 17. Added date limit
	R3	No response.
	R6	No response.
	R8	There were not all that many citations to go from on this, but I found the TMT particularly straightforward for this on this topic. I used PubReMiner and Carrot2, as I am most familiar with them. I also looked at the AntConc output but didn't do anything based on it.

Question	Review Number	Answer
	R4	Subject Headings MeSH on Demand – I used a section of text from the protocol to identify subject headings. Note: limited to 10,000 characters – had to select a section of the protocol as full protocol was over 20,000 characters. Very quick to analyze text (took seconds to get my subject headings). No additional information about term explosions, so I still have to look up each MeSH term. Articles identified as being relevant were not about pharmacists – I did not add any to my seed set of articles. PubReMiner – Still had to generate a PubMed query. I tried to build a query using my seed set of articles with their PMIDs, but this didn't work (or I couldn't get this to work). So, I generated a quick query string: (abortion or mifepristone or misoprostol) AND (pharmacist or pharmacists OR pharmacy OR pharmacies OR chemists). Query resulted in 639 references. Identified additional MeSH terms not found with PubReMiner. Difficult to determine a cut-off threshold for occurrence (selected 10 and did not look below this number) Key Words AntConc – Used on a different computer than I had previously (MacBook) and had to override some security settings to get the application to run. Exported seed set of 17 articles (below) from PubMed using query: 19442780[uid] or 27770797[uid] or 25702075[uid] or 29508948[uid] or 28823841[uid] or 16291487[uid] or 26604158[uid] or 24268354[uid] or 17046381[uid] or 21757422[uid] or 29752204[uid] or 26869694[uid] or 28673342[uid] or 29351313[uid] or 22402571[uid] or 25702074[uid] or 28935219[uid] Generated a text file from the title and abstracts and imported into AntConc. Analyzed the "Word List" tab and then selected key terms to see their context via the "Concordance" tab. I found it difficult (as per note on PubReMiner above) to determine cut-off threshold for occurrence (selected 4 and did not look below this number). Viewed some "Clusters/N-Grams" for key terms (e.g., abortion). Difficult to determine when a phrase search for a term should be used instead of a single term (i.e., I knew the phrase "medical abortion" occurred 27 times in my corpus – should I use this phrase in my search or just the term "abortion"?). Strategy Evaluation VOSviewer – Never used this tool (only seen it demonstrated). Took me a about an hour to read through the manual and download the program and then download Java to get the program to work on Mac and then generate an RIS file and figure out how to view the file properly! Once I generated a map of the results from my MEDLINE strategy (version 1.0) I saw an unrelated cluster of articles (looked like an appendage) about methotrexate (?). I looked up the term and discovered it is an abortifacient. I removed this MeSH terms from the search strategy in version 1.1 which removed about 100 records. I retested the strategy, and the new version still found all of my known items. I exported the revised results to VOSviewer to generate a new map. The new map didn't have this methotrexate cluster. The map looked much more condensed (like a football). I interpreted the new map as showing greater precision in the search results.
	R2	No response.

Question	Review Number	Answer
	R5	I used Yale MeSH Analyzer and PubReMiner for subject term and keyword text-mining, and VOSviewer for strategy evaluation. I ended up preferring PubReMiner because it presented the results for both keywords (TI, AB) and subjects from my Ovid MEDLINE search in ranked order by the number of times terms were used, which is much easier, faster, and more informative than going through the record layout available in MeSH Analyzer. Once I realized that I could export the PMID list from Ovid in spreadsheet format, and then copy and paste the PMID column from the spreadsheet into PubReMiner it became even easier and faster. For VOSviewer, it took several searches to figure out that the maps I prefer to use are constructed by downloading two RIS files (one including TI, AB, author keywords fields; and the other MeSH terms and maybe registry name fields); then in VOSviewer under file, map, create, create map based on text data, RIS tab (keep ignore fields checked), and then upload each file in turn. Once the map is created, I am looking for non-relevant keywords or MeSH terms that appear as larger bubbles since these represent the number of occurrences of the term. Sometimes (not always) this has been ideal for finding candidate terms to NOT out to make the search more specific. Of course, one has to test these first before removing them from the search...
	R7	For the text-mining searches, I utilized several approaches. I used the list of known relevant systematic reviews from the provided Excel spreadsheet to create a list of PMIDs to enter into PubReMiner & the Yale MeSH Analyzer. The results were reviewed to identify relevant keywords & MeSH headings. Several of these keywords & MeSH headings were then entered into PubReMiner & MeSH on Demand to identify additional relevant terms. In addition, I also inputted various portions of text from the Systematic Review Protocol document into MeSH on Demand to identify other relevant headings.
What number of known citations did you use in the seed set?	R1	31-60
	R3	31-60
	R6	28
	R8	11-30
	R4	17
	R2	101+
	R5	101+
	R7	31-60
Did the known citations used for TMT analysis represent the diversity of vocabulary terms or a subset of terms used in this area of research? (Preventing Garbage in Garbage Out)	R1	Perfectly balanced (had all needed terms without a lot of junk)
	R3	This was a very clean search.
	R6	The citations may have been a little broad, but generally seemed good.
	R8	Subset of vocabulary terms (had to supplement elsewhere) Comment: I had to add a number of terms that were not identified through TM, probably because there were so few seed citations.
	R4	Overly comprehensive (lots of junk terms). It was difficult to determine how representative the known citations were of the topic area.
	R2	Perfectly balanced (had all needed terms without a lot of junk). Comment: There was a lot of junk, but this is a complex topic, so I think fewer citations would have led to gaps in the search

Question	Review Number	Answer
	R5	Perfectly balanced (had all needed terms without a lot of junk). Subset of vocabulary terms (had to supplement elsewhere Comment: I would say it was somewhere between these two actually…this was a complex search and it was also an update search, so I had the list of included citations and the existing search to begin. I was also aware of the possibility that exclusively using the existing include list might bias the results (e.g., to vocabulary being used at the time of the previous review (intervening semantic drift) or if the original review search was not inclusive enough), so once a fairly robust search was established, I then re-ran the text-mining searches to determine if there were other additional text and MeSH terms to consider. I also experimented with creating an initial search with all the known terms, finding systematic reviews, meta-analyses, trials with those words in the title, taking a sample to plug into PubReMiner and VOSviewer. I thought this approach might work for new review topic searching as well.
	R7	Subset of vocabulary terms (had to supplement elsewhere). The text-mining tools did help to identify some relevant keywords & MeSH headings however there were many irrelevant results to wade through to find a small number of relevant terms. Seeing the terms out of context in PubReMiner for unfamiliar topic areas was less than helpful and required additional follow-up to determine if keywords were relevant or not. Would have preferred to use a tool where keywords could be viewed in in groups (bigrams, trigrams).
If the software offers multiple types of analyses, which one(s) did you use and why?	R1	I used the bubbles in Carrot2. I find them most intuitive.
	R3	I use them all, but for Carrot2, I like the bubbles.
	R6	VOSviewer and Carrot2 for strategy evaluation: I was experimenting. I still find Carrot2 more useful for this stage as it clumps the references into topics, but I did like looking at the way terms were connected in VOS. I did not make any changes to the search based on these tools.
	R8	Carrot2 bubbles, also the wheel. Looked at AntConc outputs (primarily the visual network) but didn't do anything with them.
	R4	See above comments on PubReMiner and AntConc. With these tools I used the analyses that calculated occurrence. These seemed logical to me: the more times a term or phrase appeared in my corpus, the more important they were to include in the search strategy. Also, see comments on VOSviewer. I generated a visual map of the MEDLINE search results. This allowed me to identify unrelated clusters of articles and to revise the strategy to include its precision.
	R2	I use them all, but for Carrot2, I like the bubbles.
	R5	For PubReMiner, once I uploaded the list of PMIDs and ran the search, there after selecting the following fields from the right-hand side of the screen to manually adjust the search: MeSH, Substance, and WORD TI_AB. These seemed the most useful ways to focus the search for developing terms. In looking over the results term occurrence was the most important factor in selecting potential terms to test in the strategy. For VOSviewer, I tried other mapping displays but found the map based on text data the easiest to 'read'. I also found that I had to split the results into two files, one for keywords in the title/abstract and the other for MeSH terms, when I didn't do that the MeSH terms dominated the display to the detriment of everything else.
	R7	Typically used default setting initially. For PubReMiner, several searches needed to be tweaked due to large retrieval. While using Carrot2, I found it helpful to view the results in Folders, Circles, & Foam Tree modes.
Do you want to share any strategies you developed to optimize	R1	No response.
	R3	No response.
	R6	No response.

Question	Review Number	Answer
information gleaned from results?	R8	No response.
	R4	A good strategy for me was to increase search precision by using a seed set of articles in conjunction with a visualization tool to evaluate the strategy. Creating a visualization of the search results from Ovid MEDLINE allowed me to identify a MeSH term (abortifacient) that was retrieving unrelated citations about methotrexate. Using the seed set, I could revise the strategy, remove the term and rerun the search to confirm its sensitivity.
	R2	No response.
	R5	See above comments!
	R7	The text-mining tools were a great compliment to usual practice and going forward I plan to utilize them more often during the strategy development period. However, I would not feel comfortable designing a strategy solely using text-mining as there are many irrelevant results returned and the lack of context for unfamiliar topic areas requires additional follow-up. While working on this project I developed a routine of flagging potentially relevant keywords & headings which then required me to do additional research to see if they were in fact useful for the strategy.
Other feedback not covered in this form.	R1	I found PubReMiner to be extremely helpful and relatively easy to use (once I figured out my method) in identifying MeSH terms and keywords. I used Carrot2 to identify unwanted concepts, but in this case, didn't find any.
	R3	I tried to use the same method I used in my other TMT searches for consistency.
	R6	I tried to use the same method I used in my other TMT searches for consistency.
	R8	No response.
	R4	I generated my seed set by: (1) using references in the protocol, (2) running a quick PubMed query and looking at related references, (3) identifying review articles on the topic and then adding their included citations. I'm unsure if there are other more effective methods to identify test articles, or if my approach was appropriate?
	R2	I tried to use the same method used in the cannabis search for consistency.
	R5	Over time I began to see the keyword/MeSH identification step as essentially the sensitivity step in the search strategy development process. Whereas, the strategy evaluation step is more akin to the precision step, increasing the focus of a search.
	R7	Time was added to search process to tweak and troubleshoot issues related to constraints of the TMT tools (character limits, output limits, search input formatting issues, etc.). For example, while using PubReMiner, searches had to be tweaked several times because the output was too large for the tool to handle. Related to this issue, I had to very narrowly limit the search date range while perming keyword searches related to CT imaging. Could this very narrow search date window negatively affect process/results?

Abbreviations: PMID = PubMed Identification; R = review; RCT = randomized controlled trial; TM = text mining.