Advancing the Semi-automation of Literature Identification for Evidence Synthesis: Evaluation Measure Prioritization, Training Data Development, and Language Model Implementation

Gaelen Phyfe Adam, MLIS MPH

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in the department of

Health Services, Policy, and Practice

BROWN

School of Public Health

Brown University Providence, Rhode Island May 2024

© Copyright 2024 by Gaelen P. Adam

This dissertation by Gaelen Phyfe Adam is accepted in its present form by the department of Health Services Research as satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date_____ _____

Ethan M. Balk, Advisor

Recommended to the Graduate Council

Date_____ _____

Thomas A. Trikalinos, Reader

Date_____ _____

Byron Wallace, Reader

Date_____ _____

Ian J. Saldanha, Reader

Approved by the Graduate Council

Date_____ _____

Thomas A. Lewis, Dean of the Graduate School

CURRICULUM VITAE

Gaelen P. Adam earned a Bachelor of Arts in Folklore and Mythology from Harvard University in 1996. She then worked for several years as the production and then managing editor at a series of trade magazines. In 2011, she earned a Master's in Library and Information Science (MLIS) from Simmons University, and subsequently worked as a clinical librarian, providing relevant and high-quality research to front-line clinicians and hospital policy developers. In 2013, she began working for the Center for Evidence Synthesis in Health (CESH) at Brown University, one of nine Agency for Healthcare Research and Quality-funded Evidence-based Practice Centers (EPCs) as a librarian, editor, and research associate. In 2019, she earned a Master's in Public Health (MPH) from Brown University's School of Public Health.

Over the course of her career, she has been involved in all steps of evidence synthesis and related projects and has developed a deep understanding of the methods and tools used in evidence synthesis research. As a research associate and the program manager for the Brown EPC, she has contributed to the production of over 30 evidence synthesis products (systematic reviews, technology assessments, and other similar products) on a wide variety of clinical and public-health topics.

A list of her published work is available in My bibliography:
https://www.ncbi.nlm.nih.gov/myncbi/1zoccHw-NQ1kW/bibliography/public/

ACKNOWLEDGMENTS

I am grateful to my committee, Dr. Ethan Balk, Dr. Thomas Trikalinos, Dr. Byron Wallace, and Dr. Ian Saldanha, for the time and effort they have dedicated to this project and to my development as a researcher. They are not just mentors in my Ph.D. process but have been colleagues and friends for over a decade.

If you want to know how to do evidence synthesis right, participate in any project lead by Dr. Ethan Balk. I have been lucky enough to have participate in dozens. Dr. Balk has been a mentor in every sense of the word. He is a friend, a guide, a constructive critic, and a source of unwavering support. I could not have done this without him. Dr. Trikalinos has fostered my interests in the intersection of computers and systematic review methodology. He has been generous with his time, support, and guidance, while encouraging me to take leadership roles in his projects. Dr. Byron Wallace has brought consistent wit and expertise to my work with him. His immense understanding of the computer science side of this project was absolutely critical to is success. Dr. Ian Saldanha's work in systematic review and core outcomes sets was crucial to the first chapter of this dissertation. He has been generous with his time in reading and commenting on my writing, listening to presentations, and answering what must feel like unending questions.

I would also like to thank: (1) Jay DeYoung, my partner in Chapter 2. Jay not only programmed the language models I used but met with me weekly to explain the computer science behind the code. (2) Dr. Alice Paul, without whose help I would still be cleaning searches. Dr. Paul was immensely helpful in my learning about language models and how they work and has generously helped me write and edit code time and again. (3) Dr. Elyse Couch came to CESH toward the end of this process and not only supported the final chapter with her expertise in qualitative methods, but also gave generously of her time and expertise in reading and commenting on drafts. (4) Valerie Rofberg checked the equations and definitions of all of the measures discussed in Specific Aim 1. (5) Finally, a huge thanks to all of the librarians I have worked with and who contributed to this research, to all of the CESH staff over the years, and to the folk in the AHRQ EPC Program and SRC.

I cannot express how much it meant to me to be able to do this degree while working at Brown. And I would like to thank everyone in the HSPP program who went to bat for me, especially Dr. Amal Trivedi

for successfully making my case to the graduate school and Dr. David Meyers for navigating the ins and outs of the various requirements, so that I could be here today. I would also like to thank my cohort and all of the current and past students who have shared their wisdom borne of experience, particularly Courtney Baird and Jasmine Agostino.

Finally, I would like to thank my family for fostering a love of learning and intellectual curiosity, my daughters, Brooke and Maia, for their endless support and encouragement. And above all, and my husband, Paul, who has literally driven the boat with me through almost 20 years of thick and thin and what must have felt like unending school. I would be lost without you.

TABLE OF CONTENTS

List of Tables

List of Figures

CHAPTER 1. INTRODUCTION

1.1 Background

1.1.1 Systematic Review Process

Systematic reviews and related evidence synthesis products (e.g., rapid reviews, scoping reviews) form the basis of evidence-based healthcare. According to the Cochrane Collaboration website, "A systematic review attempts to identify, appraise, and synthesize all the empirical evidence that meets pre-specified eligibility criteria to answer a specific research question."[1] The systematic review process includes the application of scientific methods to identify and evaluate the studies they summarize, yielding rigorous appraisals of the evidence for a given topic.[2-4] From summarizing and evaluating evidence for providers and patients to informing guidelines that direct individual patient and provider decisions, systematic reviews form the basis of evidence-based healthcare and health policy.[5]

The steps in the evidence synthesis process include the following:

Topic development. In this step, researchers define the specific question(s) they want to answer, develop an analytic framework that illustrates the mechanism by which the Population of interest may be affected by the Intervention or Exposure of interest in comparison with relevant Comparators to impact the Outcomes of interest. This constitutes the widely used PICO (Population, Intervention/Exposure, Comparator, Outcome) formulation, which defines the selection of studies for inclusion in a systematic review and the data to be extracted and summarized.

Literature identification. In this step, the PICO elements are translated into a search query by identifying synonyms and subject terms and combining them with the appropriate Boolean operators (OR, AND, or NOT). The query is typically developed for one database (e.g., Medline) and then translated for other relevant databases (e.g., Embase), and the titles and abstracts of the identified citations are then screened for relevance. Citations deemed relevant during abstract screening are screened in full text, and a final corpus of studies for inclusion in the review is defined.

Data extraction. In this step, relevant data from each included study are extracted from study reports, and the risk of bias in the study is assessed.

Analysis and synthesis. In this step, the data are collected into tables, and narrative analyses, and, when appropriate, meta-analyses summarize the results across studies. The review team then makes conclusions

regarding outcomes. Often, the review team goes through the additional step of determining the strength (or certainty) of evidence across studies for each conclusion (outcome).

Reporting. The systematic review is written up as a final report and shared with relevant stakeholders and, ideally, the public.

This dissertation focuses on ways to improve the quality and reduce the time and resources spent during the literature identification step, specifically in the process of search query development. This step is fundamental to a systematic review because it generates the potentially relevant evidence for screening. Missed studies compromise the comprehensiveness of the systematic review process, while, on the other hand, searches with poor specificity compromise the efficiency of the process.

1.1.2 Too much literature, too little time

The Cochrane Collaboration, an international network of independent researcher volunteers who conduct systematic reviews and develop methodologic guidance for their conduct, is perhaps the best-known source of systematic reviews. Over the past 30-plus years, the Cochrane Collaboration has built and sustained a reputation for providing high-quality and unbiased systematic review evidence.[6-8] It has codified its recommended methods in The Cochrane Handbook for Systematic Reviews of Interventions ("Cochrane Handbook"), a highly regarded and comprehensive manual for conducting evidence syntheses.

The Cochrane Handbook recommends identifying a corpus of potentially relevant articles using "thorough, objective and reproducible" search queries, and then screening this set in duplicate.[9] This methodology leads to queries implemented in multiple medical literature electronic databases, which are designed to cast a wide net, resulting in review teams commonly needing to screen tens of thousands of records (typically titles, abstracts, and publication data) to identify potentially relevant articles (often hundreds), which must be further screened in full-text. Literature identification, including searching for studies and screening the resulting citations to identify the relevant literature, plus data extraction, account for approximately two-thirds of the time required to conduct a systematic review.[10] Thus, the desire to identify every relevant article must be balanced against time, resource, and budget constraints.[11]

The balance between expansive, comprehensive queries and resource limitations may introduce a potentially significant bias: queries are commonly artificially limited to return only the number of citations considered feasible for the team to screen. If the articles missed have similar results to those found, the only consequence is a lower certainty of evidence (largely stemming from lower precision) for any conclusions drawn from the systematic review synthesis. However, if the search query systematically misses a specific subset of studies, the review results could be biased.[12 13] For example, hypothetically, if

the clinical trials that report the effectiveness of an intervention are identified but the cohort studies of harms are missed, the report could mistakenly conclude that a given intervention is more effective and less harmful than it really is. Medical librarians play an important role in developing unbiased, comprehensive search queries that appropriately balance the requirement to identify all relevant studies with the constraints of the team's time and budget.[14] However, not every review team has access to an experienced librarian, and, even for trained librarians, the development of queries is time-consuming. One study that surveyed 185 librarians reported that the mean time it took to develop and translate search queries for a single review was 26.9 hours (median 18.5 hours).[15] Recent studies, evaluating hundreds of reviews, have shown that queries developed without the inclusion of a medical librarian were less likely to appropriately incorporate key aspects, such as subject headings, adequate natural language terms and truncation, translation of the base query into other databases, and documentation of the search process, including the specific query, in a reproducible way.[14 16] Thus, there is a clear need for tools that assist in the design and development of high-quality search queries.

1.1.3 Semi-automation in evidence synthesis

Many tools to **semi-automate search query development and abstract screening** use *natural language processing*, a combination of text mining and machine learning. *Text mining* has been characterized as "a variety of processes that enable discovery of words and patterns in collections of text."[17] The primary goals of text mining are to retrieve and extract information from unstructured text, and then to look for patterns. *Machine learning* models, such as support vector machines, random forests, and neural networks, can be used to find relevant documents or extract data. Natural language processing combines text-mining tools and machine learning models to make sense of unstructured human language for a specific objective, such as classification, ranking, retrieval of relevant documents, text generation, text extraction, and relation extraction. Recent advances in natural language processing have yielded language models that are pretrained on textual data from large amounts of text (usually from the Internet) to predict which words generally follow specific sequences of other words, then these models are retrained (fine-tuned) on a smaller dataset that is specific to the desired task. At a basic level, transformer-based language models take an input sequence and translate it to an output sequence. One of the first broadly used transformer-based language models the Bidirectional Encoder Representations from Transformers (BERT) model.[18 19] BERT produces a representation for every token (word part) in context, using tokens both before and after the token of interest (hence bidirectional) for prediction. The success of BERT has spawned the development of increasingly large pretrained decoder-only large language models, such as T5, ChatGPT, and Claude.ai. Unlike bidirectional models, decoder-only models only use tokens before the token of interest for prediction of the next token, this makes them exceptionally good at text generation.[20]

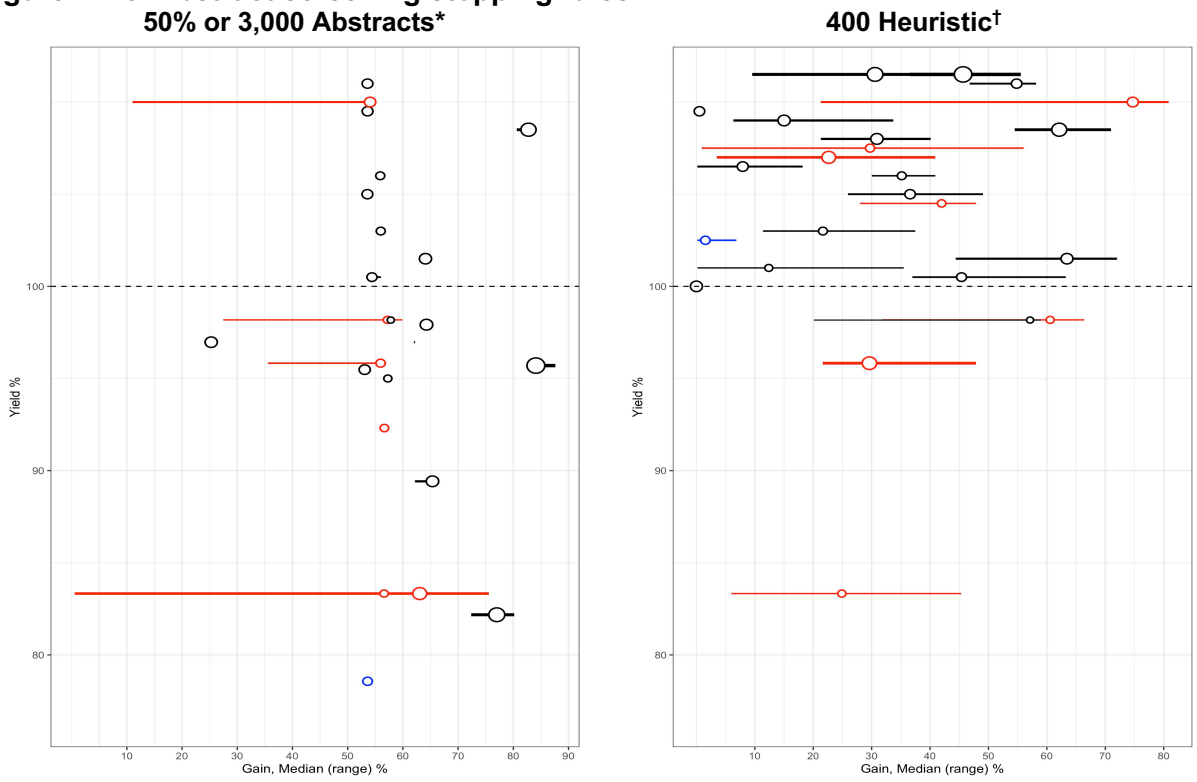
Text mining and machine learning during abstract screening. Traditional abstract-screening tools that use text mining and machine learning to assist in the screening of records identified from the search query have been extensively used and studied.[11] These tools use a series of human labels (e.g., "accept" or reject") to estimate the probability that each of the abstracts that have not yet labeled by a human are relevant. These predicted probabilities allow each abstract to be ranked and classified as likely relevant or likely irrelevant based on an assigned threshold. Training is done either using a single large training set (in, for example, an update of a systematic review update where the original review serves as the training set) or iteratively over several training sets developed during the screening for an ongoing review,[21] a strategy known as *active learning*. A large number of studies have shown that machine-learning algorithms are able to rank and classify records accurately, with adequate training.[11 22-24] One example of such a tool is Abstrackr, which relies on natural language processing, in which a set of 11 support vector machines rank each abstract as to the likelihood of relevance of each abstract based on the abstracts that have already been screened. Then a "vote" is taken of each model's prediction, and the prediction of the majority of models is used.[22]

Research to identify indicators for when the likelihood of missing any relevant studies is sufficiently low to justify when a team could stop human screening and rely on the model's labels for the unscreened abstracts is ongoing.[25 26] In unpublished prior work on Abstrackr, we simulated the active learning algorithm over approximately 50 screening iterations of 24 completed review projects, based on random sets of the initial 100 abstracts screened. We evaluated the *yield* (i.e., the percentage of included citations identified by a semi-automated screening process) and *gain* (i.e., the percentage of citations that are rejected by the classifier in the prediction phase and are never read by a human) across the screening process at a series of *stopping-rule heuristics* (e.g., stop after screening half of the citations, stop after a specific prediction level is reached, stop after a specific prediction level is reached and a set number of sequential abstracts are rejected).

Figure 1.1.3 shows the results of two possible stopping rules. With one heuristic, shown in the left panel, the review team stops human screening at 50% of the corpus or when 3,000 abstracts have been screened, whichever comes first (regardless of the size of the corpus). Based on this heuristic, across evaluated reviews, a median of 97.8% of relevant abstracts were found (full range 70.8% to 100%) and a median of 56.6% of citations were not screened by humans (full range 25.3% to 84.1%). In the other heuristic, shown in the right panel, the review team stops human screening when the maximum prediction score is less than 0.4 and the team has rejected an additional 400 consecutive citations. The cutoff of 0.4 was based on our experience screening several dozen projects. The choice of 400 additional rejections was the lowest of several tested and is based on calculations that show that the upper 97.5% confidence interval

bound for a proportion of 0/400 is less than 1%. We also tested thresholds of 600, 800, and 1000 citations after the 0.4 prediction score was reached. Based on this heuristic (maximum prediction score less than 0.4 and 400 subsequent rejections), across evaluated reviews, a median of 100% of relevant abstracts were found (full range 83.3% to 100%) and a median of 30.7% of citations were not screened by humans (full range 0% to 74.7%). Although no stopping heuristic had perfect performance over all 50 iterations, using random starting sets of citations, heuristics that combine a maximum prediction score threshold and a series of negative labels may perform well enough to be practical to implement in most systematic reviews. This work is being updated on a new version of the Abstrackr program that runs a BERT-based language model pretrained on PubMed abstracts[27] and has been implemented in the Systematic Review Data Repository Plus (SRDR+) software's abstract screening tool.

Figure 1.1.3. Abstract screening stopping rules



* Stop human screening after 3,000 abstracts or 50% of dataset has been labeled, whichever comes first.
† Stop human screening when the maximum prediction score is less than 0.4 and we have rejected an additional 400 consecutive citations.
Each project is represented as a dot and line, showing the median and range of abstracts not screened at the median iteration yield (percent relevant articles found; y axis). All projects above the dashed lines had perfect yield in the median iteration. Smaller circles indicate smaller projects. We ran 50 iterations for each project with different random starting sets. Black lines indicate projects in which there was little variability between iterations; red lines indicate projects in which there was substantial variability between iterations, with some iterations achieving perfect yield much later than others, leading to variable gain as well; the blue dot is the project that achieved high yield only after all abstracts were screened.

Text mining and machine learning in literature search query formulation. There is less evidence for the role of machine learning or artificial intelligence in literature search query *formulation* than there is for its role in abstract screening. Most available tools employ text mining to automatically generate terms for query expansion, identify related articles through citation analyses, or translate queries for use in different databases. A recent evaluation of 21 available tools for query development in the Systematic Review Toolbox[28] (an online catalogue of tools that support various tasks within the systematic review process) concluded that although useful tools are available for most steps of the process, all currently available tools struggle with long or complex queries, requiring the user to break the query into its component parts (e.g., condition of interest, interventions of interest, eligible study designs).[29]

Machine learning models have been used to combine automatic term generation and citation ranking, leveraging thesauruses (e.g., MetaMap, MeSH) and human identification of relevant articles.[30 31] In one example, our team developed Pythia, a tool that uses machine-learning algorithms to rank all of the citations in Medline that are available through the PubMed interface based on relevance to a natural language prompt, returning the 100 top-ranked citations for human screening.[32] The labeled citations are then iteratively used by Pythia to refine the query and re-rank the citations, thus alternating between searching and screening. When tested against seven completed projects (with an arbitrary stopping criterion of 10 rounds of screening or approximately 1,000 abstracts screened), we found that while the burden was decreased in most reviews, the sensitivity of Pythia to retrieve the relevant abstracts was low, ranging from 8% to 58%. We concluded that the poor sensitivity was, at least in part, because the natural language questions have large amounts of implicit domain knowledge (by the humans) that traditional machine learning models struggled to encode and parse. For example, in the Nonmelanoma Skin Cancer report had two key questions, which translated into two natural language questions: (1) For adult patients with basal cell and squamous cell carcinoma of the skin, what is the comparative effectiveness of various interventions, overall and in subgroups of interest? And (2) For adult patients with basal cell and squamous cell carcinoma of the skin, how do the adverse events associated with the various interventions compare overall and in subgroups of interest? However, these seemingly clear questions are translated by humans into a complex network of concepts and terms, pertaining to various aspects of the population (disease concepts, location concepts, and population concepts), intervention, outcomes, and study design. It is likely that Pythia would have performed better with modern pretrained models, such as the BERT-based models discussed earlier.

With the rise of artificial intelligence, there is increasing interest in “teaching” large pretrained models to parse natural language questions into queries. Wang et al. investigated whether the large general language model ChatGPT could be used to formulate Boolean queries. They found that queries developed by this

model had extremely poor sensitivity (between 4% and 13% of relevant citations correctly identified), albeit with a large reduction in the number of irrelevant citations identified.[33] There is a need to improve the performance of such models before they can be incorporated into the systematic review process.

1.1.4 Methods for the evaluation of tools

With the ongoing development of tools to semi-automate aspects of the query development, it is increasingly important to find meaningful ways of evaluating their performance. Hirschman and Blashke discuss performance evaluation in depth in their chapter in Text Mining for Biology and Biomedicine.[34] They point out that one must first identify the relevant stakeholders. For systematic review literature identification methods and tools, this includes the users (librarians and other systematic reviewers), developers (computer scientists and software engineers), and statistical experts. Hirschman and Blashke further note that "it is critical to select clear, reproducible, and easily understood evaluation measures." [34] Yet, in a 2015 systematic review of studies using tools to semi-automate query development and screening, O'Mara-Eves et al. identified 19 evaluations measures.[11] My recent (2022) update to this review identified an additional 18 evaluation measures, bringing the current total to 37. Some evaluation measures have been used frequently, but almost two-thirds were used in fewer than five studies, and 41% were used in only a single study. On the one hand, rarely used measures may be less useful. On the other hand, they may be used less because they are novel, and this field is evolving quickly. Thus, there is a need to identify what existing measures are most useful, as well as the attributes that make them most useful.

Many evaluation measures are based on a standard diagnostic test study two-by-two table, as shown in Table 1.1.4. These measures require a reference ("gold") standard set of accepted and rejected articles. However, they are constrained by the fact that the true negative group can be very large. With few accepts and many rejects, measures whose workload calculations include true negatives will not change as rapidly as their true performance changes. The reference standard set generally comprises the final articles included in a well-run systematic review that incorporated multiple literature sources, broad queries, full manual screening, etcetera. However, not every topic has a comprehensive or well-run systematic review, and in these cases the reference standard may not, in fact, give a full list of included studies. Thus, another option is to create a reference standard set by combining the relevant articles found by either of the method or tools being compared. O'Mara-Eves and colleagues note that the appropriate evaluation measure will depend on the study design and method or tool being evaluated. Some measures may be meaningless in measuring the performance of some types of tools. For example, measures designed to assess the ability of machine learning tools to adequately rank unscreened abstracts, such as work saved over sampling at 95% sensitivity, are not meaningful in the context where a human screens all records and

sensitivity may not reach 95%. They also note that these measures are subject to research decisions on the cutoff value used, the definition of true negatives, and other similar subjective decisions; thus, a given measure may not be directly comparable across studies.[11]

Table 1.1.4. Example two-by-two table

	Records Included in Reference Standard	Records Not Included in Reference Standard
Records identified by Tested Tool/Methodology	True Positives (TP)	False Positives (FP)
Records not identified by Tested Tool/Methodology	False Negatives (FN)	True Negatives (TN)

1.2 Specific Aims

This dissertation aims to move the field of literature identification automation forward, first by establishing what evaluation measures (or types of measures) are most valuable to different systematic review stakeholders and then by using these measures to empirically evaluate a novel language model. The model is fine-tuned on a set of more than 10,000 manually curated titles, research questions, and search queries, then it is evaluated using a separate dataset of 57 reviews, which includes the titles, key questions, search queries, and the PubMed identifiers (PMIDs) of the final included studies. Finally, we use qualitative methods to evaluate the utility of the model for medical librarians.

The specific aims are:

Specific Aim 1: Establish which attributes of measures that quantitatively evaluate the performance of literature identification tools or methods for systematic reviews are important to stakeholders, through a discrete choice experiment.

Specific Aim 2: Develop a dataset of systematic review search queries and use this dataset to fine-tune a large language model that develops de novo search queries. Develop a separate evaluation dataset and empirically evaluate this model.

Specific Aim 3: Use semi-structured interviews with practicing medical librarians to assess how they would use the fine-tuned large language model and to guide its further development.

1.3 Dissertation Structure

This mixed-methods dissertation is divided into three chapters, corresponding to the three specific aims. In the first chapter, we discuss a discrete choice experiment, in which we seek to establish the characteristics of measures that are meaningful to evaluate literature identification tools/methods. Based on these characteristics, we suggest measures that should be used in reports of the results of new tools so

that they can be compared with each other, giving users the information they need (i.e., common set of measures) to choose a tool and developers the information they need for future development.

In the second chapter, we use these measures in the empirical evaluation of a tool that we developed to generate search queries from free text. To create this tool, we first cleaned more than 10,000 search queries published in the PROSPERO database, and then we used those data to fine-tune a language model.

In the final chapter, we discuss interviews with eight medical librarians about the tool's performance, how it could be useful, and how they would like to see it evolve.

CHAPTER 2. A DISCRETE CHOICE EXPERIMENT TO ESTABLISH THE ATTRIBUTES OF MEASURES THAT ARE USEFUL FOR EMPIRICALLY EVALUATING NOVEL LITERATURE IDENTIFICATION TOOLS

2.1 Introduction

Systematic reviews and related evidence synthesis products (e.g., rapid reviews, scoping reviews) form the basis of evidence-based healthcare.[5] Yet they are expensive and time-consuming to produce, often requiring more than 1,000 person-hours of labor.[10] Literature identification, including searching for records and screening the resulting records to identify the truly relevant records, plus data extraction, account for approximately two-thirds of the time required to conduct a systematic review.[10] Automation of these time-consuming processes has the potential to reduce the time and costs of systematic reviews, helping ensure that these important products will not be avoided or abandoned due to their expense.[35]

Despite the wide variety of tools that have been developed over the past 15 years to assist in the literature identification process,[36] the uptake of existing tools by systematic reviewers has been low. The low uptake has been attributed to two major factors. First, there is a lack of trust in existing tools, and second, many tools are not easily integrated into current workflows or are difficult for individuals without the relevant technical expertise to learn and use.[37] A major challenge to establishing trust in new literature identification tools is the wide variety of measures that have been employed to evaluate their efficacy. This heterogeneity in measures used makes it difficult for the user to compare tools across studies of tool performance, thus precluding the comparative analysis required to choose a tool or know how to best implement one in the review workflow. In addition, many of the measures used (e.g., F-score) are not immediately intuitive and others suffer from ambiguity (e.g., recall/sensitivity, burden). In their 2015 review, O'Mara-Eves et al. identified 19 measures.[11] In 2022, we updated the search and identified 18 new measures, which brings the total to 37. A few measures were used very commonly, but almost two-thirds were used in fewer than five studies and 41% were used in only a single study.

2.1.1 Motivation for a discrete choice experiment

To establish which measures a variety of stakeholders prefer, and thus which measures to recommend that developers use consistently to make tool/method comparison easier, we decided to conduct a survey. We initially thought that we would simply ask which measures a variety of stakeholders (e.g., librarians,

systematic review methodologists, software developers) would like to see in the evaluation of novel tools or methods of literature identification. However, we quickly established that many stakeholders are not explicitly familiar with at least some of the measures that have been used to date, and when asked which they liked best, they expressed a lack of sufficient knowledge to choose. Thus, we decided to take a step back and see if there were certain attributes of each measure (such as whether it represented a count or a ratio, how it changed across data sets, and so on) that would allow us to categorize the types of measures preferred and thus recommend specific measures to be reported more consistently. Again, we were concerned that explicitly eliciting this information would not be useful because the stakeholders do not necessarily consciously conceptualize measures by their attributes, though they may still map to those attributes in deciding which measure they prefer.

We chose to use a discrete choice experiment, where users are given a series of questions in which they choose between two or more alternatives. Each alternative has (or does not have) a series of attributes, and the choices made by respondents between the option sets that differ in one or more attributes can be used to learn about overall preferences for these attributes (i.e., which attributes are most important to stakeholders). This, in turn, can be used to predict their choices for measures with similar or different combinations of attributes.[38] For this experiment, we collected the measures that have been used in published studies, identified a set of five attributes that could be used to describe the measures at a broad level, and performed a discrete choice experiment via a survey to elicit stakeholders' beliefs about the relative importance of each attribute.

Because the terminology in this report may not be immediately intuitive, Table 2.1.1 lists the terms used in this chapter, their definitions, and an example of each.

Table 2.1.1. Terms used and their definitions

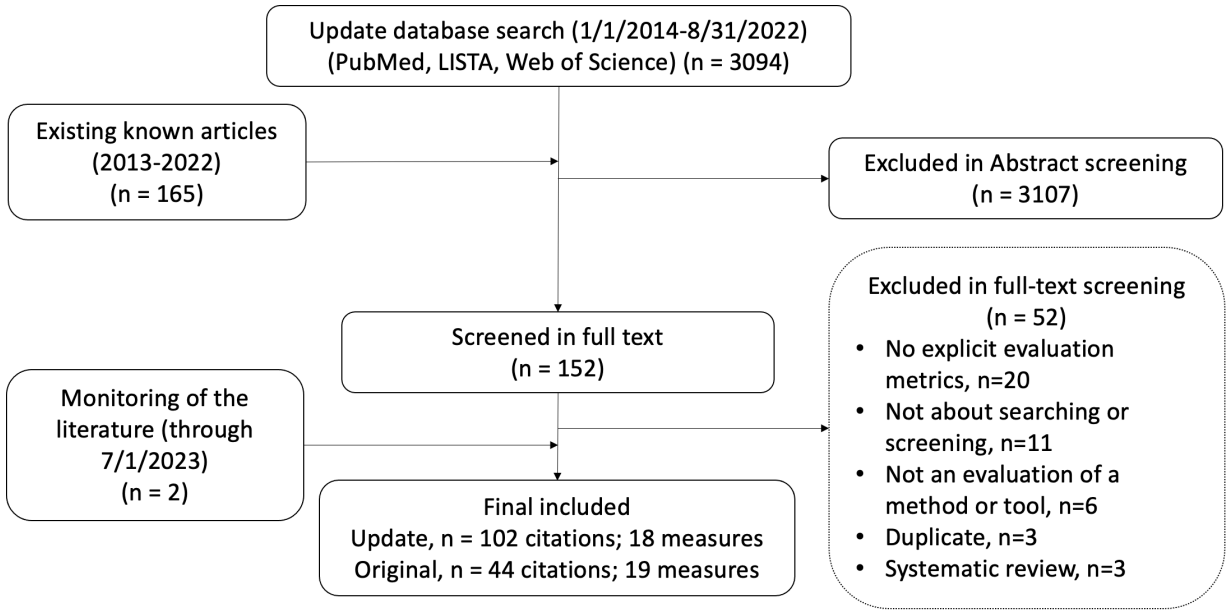
Term	Definition	Examples
Measure	A way of describing the performance of a literature identification tool	Sensitivity, precision, work saved over sampling
Set	Group of measures that together indicate the performance of a literature identification tool	True positives + precision + total number identified
Pair	Two measure sets that are compared with the respondent choosing the one they prefer	True positives + precision + total number identified vs. Sensitivity + accuracy + total number identified
Attribute	Abstract characteristic of a measure that may define why the measure is or is not preferred	Relevant records correctly identified as a number or a ratio
Scenario	Hypothetical use case that is intended to help establish which measures are to be compared	Scenario 1: Comparison of two queries/database search strategies. Scenario 2: Comparison of two computer-aided screening processes (machine-learning models) that have been trained in different ways.

2.2 Methods

2.2.1 Collecting the measures

In their 2015 review, O'Mara-Eves et al. identified 44 records that used 19 measures at least once in assessing the performance of a novel tool or method to identify (search or screen) literature for evidence synthesis.[11] We updated of their search in September 2022, and identified 100 new relevant records (flow diagram in Figure 2.2.1) and 16 new measures since 2015. Our ongoing monitoring of the literature in Medline via PubMed through July 2023 identified 2 more records with measures not previously identified. Thus, we identified a total of 37 measures used in 102 records. A few measures were used very commonly, but almost two-thirds were used in fewer than five studies and 41% were used in only a single study.

Figure 2.2.1. Flow diagram



For this evaluation, we used only the 16 measures whose calculations are based on two-by-two tables. Table 2.2.1-1 illustrates the two-by-two tables used in calculating the measures. Table 2.2.1-2 lists the measures, their definitions, and their formulas. To provide context, we divided the measures into two scenarios:

Scenario 1: Comparison of two queries/database search strategies: "You are reading/reporting a study that contrasts two queries representing different literature identification methods, such as database queries written by a human vs. ChatGPT or a database query vs. citation tracing strategy." The 11 measures included in this scenario all assume that the entire corpus is manually screened and therefore rely on a single two-by-two table.

Scenario 2: Comparison of computer-aided screening processes (machine-learning models) that have been trained in different ways: "You are reading/reporting a study that contrasts two literature-identification methods." The five measures included in this scenario are calculated using two two-by-two tables: one for records that are manually screened and one for those screened only by the tool.

Table 2.2.1-1. Two-by-two tables

Scenario 1

	Records Included in Reference Standard	Records Not Included in Reference Standard
Records identified by Tested Tool/Methodology	TP	FP
Records not identified by Tested Tool/Methodology	FN	TN

Scenario 2

	Records Screened by tool + humans		Records screened by tool only	
	Records Included in Reference Standard	Records Not Included in Reference Standard	Records Included in Reference Standard	Records Not Included in Reference Standard
Records identified by Tested Tool/ Methodology	TP_L	FP_L	TP_U	FP_U
Records not identified by Tested Tool/Methodology	FN_L	TN_L	FN_U	TN_U

Notes: *L* = human-labeled records; *U* = unlabled records; TP = true positive; FN = false negative; FP = false positive; TN = true negative; P = positive; N = negative.

Table 2.2.1-2. Measures, definitions, and formulas for each scenario

Scenario	Measure	Definition	Formula
Scenario 1	Sensitivity/ recall	Number of relevant articles identified among all relevant articles	$\frac{TP}{TP + FN}$
	True Positives	Number of relevant articles identified	TP
	Accuracy	Total number of correctly classified items divided by the total number of items	$\frac{\beta TP + TN}{\beta(TP + FN) + FP + TN}$
	Error	Total number of falsely classified items divided by the total number of items	$\frac{\beta FP + FN}{\beta(TP + FN) + FP + TN}$

Scenario	Measure	Definition	Formula
	Precision/Positive predictive value (PPV)	The proportion of articles identified as being relevant by the tool that are truly relevant	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$
	Negative Predictive Value (NPV)	The proportion of articles identified as being not relevant by the tool that are truly not relevant	$\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}}$
	Specificity	Number of articles identified as irrelevant among all irrelevant articles	$\frac{\mathrm{TN}}{\mathrm{FP}+\mathrm{TN}}$
	Number needed to read (NNR)	The number of articles that must be screened to find one relevant article	$\frac{\mathrm{TP}+\mathrm{FP}}{\mathrm{TP}}$
	Positive likelihood ratio (LR+)	Indicates the increase in odds favoring the likelihood of inclusion given identification by the tool/methodology	$\frac{\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}}{\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}}$
	Correlation coefficient (CC)	Measures the differences between actual values and predicted values and is equivalent to the chi-square statistic for a 2 x 2 contingency table	$\frac{(\mathrm{TP}\cdot\mathrm{TN})-(\mathrm{FP}\cdot\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$
	F-score family of measures	A weighted harmonic mean of Sensitivity/Recall and Precision	$\frac{(\beta^2+1)\mathrm{TP}}{(\beta^2+1)(\mathrm{TP}+\mathrm{FP}+\beta^2\mathrm{FN})}$
Scenario 2	Absolute screening reduction	Number of items rejected by the classifier	$\mathrm{FN}_U + \mathrm{TN}_U$
	Sensitivity of the process (Yield)	The proportion of included records identified by a semi-automated screening process	$\frac{(\mathrm{TP}_L+\mathrm{TP}_U)}{(\mathrm{TP}_L+\mathrm{TP}_U+\mathrm{FN}_L+\mathrm{FN}_U)}$
	Burden	The proportion of records that a systematic reviewer has to manually screen to identify all relevant articles	$\frac{(\mathrm{N}_L+\ \mathrm{TP}_L+\mathrm{TP}_U)}{(\mathrm{N})}$
	Gain	The proportion of records rejected by the classifier in the prediction phase and never read by a human	$\frac{(\mathrm{N}_U)}{(\mathrm{N})}$
	Combined Sensitivity of the process (Yield) and Burden	Relative measure of burden and yield that takes into account preferences for weighting these two concepts	$\frac{(\beta\cdot\mathrm{yield})+(\mathrm{gain})}{\beta+1}$
	Work saved over sampling (WSS)	Proportion of papers that reviewers do not have to read because they have been screened out by the classifier	$\frac{(\mathrm{TN}_U+\mathrm{FN}_U)}{\mathrm{N}-(1-\beta)}$

Notes: Scenario 1 in purple and scenario 2 in green. *L* = human-labeled records; *U* = unlabled records; TP = true positive; FN = false negative; FP = false positive; TN = true negative; N = total number

We used dummy data to calculate plausible values for each measure: one for a tool that performed well and a second for a tool that performed less well. We then put the measures into sets that included all measures required to define a tool. Each set included (1) a measure to establish whether the tool identifies all relevant records (sensitivity), (2) either a measure assessing workload alone or a combined measure assessing sensitivity and workload, and (3) the total number of records identified.

2.2.2 Establishing the attributes

Over a series of meetings, two members of the team looked closely at the 16 measures to establish meaningful attributes. The goal of this project is to evaluate the attributes of the measures rather than measures themselves. Thus, there must be enough distinguishing features that no profile of attributes contains more than two measures for each scenario. In the end, we established five attributes that are not colinear (for scenario 1) and are meaningful in terms of the measure's use and interpretation.

For measures that establish whether all relevant records are identified:

Attribute 1: Number vs. ratio of relevant records correctly identified: For the two measures used to establish whether the tool identifies all relevant records, the attribute was simply whether the measure represented a ratio or not. Thus, whether the measure was (i) simply the number of True Positives or (ii) a ratio of the number of True Positives divided by all relevant records, alternatively referred to as sensitivity, recall, or yield (in the case of scenario 2).

For measures that target workload reduction:

Attribute 2: Dependent on prevalence. Whether the measure is affected by the prevalence of included records in the overall population (N measures with this attribute = 7). This is important because prevalence will change between evaluation sets and, in active learning scenarios, over time. Operationally, this was determined by whether the measure changed as the prevalence of included records in the report changed.

Attribute 3: Explicitly weights sensitivity and workload. Whether the measure is designed with a valuation between sensitivity and workload (N measures with this attribute = 3), indicated by a β in the formula (either explicit or implicit). These measures can be used to create a single number that describes performance in terms of both sensitivity and workload. For some (e.g., the F1 score and accuracy) the valuation is implied to be equal weight; for others (other F scores and Work Saved over Sampling), the user is expected to specify a weight for the relative importance of sensitivity.

Attribute 4: Increases linearly with workload (N measures with this attribute = 3). These measures increase in a linear fashion as workload (True Positives + False Negatives) increases. They can, thus, be easily interpreted in terms of the workload reduction. Operationally, this was determined by whether the derivative of the measure divided by the derivative of the total workload is greater than 0.

Attribute 5: Changes slowly under class imbalance. Whether the workload calculations include true negatives (N measures with this attribute = 8). Because these datasets tend to be highly imbalanced, with few included records and many rejected records, measures whose workload calculations that include true negatives will not change as rapidly as screening burden changes. For example, when looking at database search strategies where the pool of potentially relevant records is in the millions, specificity (True Negatives/Negatives) tends to be nearly perfect, even in searches that are not very precise.

A requirement for the attributes used in a discrete choice experiment is that they can be used to predict responder preferences for a new measure with the same combination of attributes.[38] For example, in this case, we could predict the desirability of the Diagnostic Odds Ratio, calculated as (True Positives*True Negatives)/(False Positives*False Negatives) as a measure. The attributes of the Diagnostic Odds Ratio are that it is not dependent on prevalence, includes no explicit valuation between sensitivity and workload, is not linear function of workload, and includes true negatives in its calculation. This puts it in the same class as the existing measures specificity and positive likelihood ratio.

Two investigators in conference assigned each set of measures to a profile. Of the 32 potential profile combinations across both scenarios, 17 had at least one measure set (14 for scenario 1 and 4 for scenario 2), and eight had two measure sets (6 for scenario 1 and 1 for scenario 2).

We presented the measures, calculations, and attributes to three experts (two systematic reviewers and a librarian) in a series of research meetings. This was not done as a part of a formal qualitative study but allowed us to get feedback on the measure sets and attributes from local experts before incorporating them into the discrete choice experiment. After discussion with the experts, the feedback was that the chosen attributes were appropriate and sufficient to describe the measures. The measure sets and their attribute profiles are in Table 2.2.2.

Table 2.2.2. Measure Sets

Attribute profile	Measure Set 1	Measure Set 2
00001	Scenario 1: N relevant records found (TP), specificity, N total records	Scenario 1: N relevant records found (TP), Positive likelihood ratio, N total records
00010	Scenario 1: N relevant records found (TP), N irrelevant records found (FP), N total records	none

Attribute profile	Measure Set 1	Measure Set 2
00011	Scenario 2: TP, absolute screening reduction, N	none
00101	Scenario 2: N relevant records found (TP), combined Sensitivity of the process (yield) and Burden at a target yield of 95%, N	none
01000	Scenario 1: N relevant records found (TP), precision/positive predictive value (PPV), N total records	none
01001	Scenario 1: N relevant records found (TP), Negative Predictive Value (NPV), N total records	Scenario 1: N relevant records found (TP), Correlation coefficient, N total records
01010	Scenario 1: N relevant records found (TP), Number Needed to Read (NNR), N total records	none
01100	Scenario 1: N relevant records found (TP), F1 score (sensitivity and precision are equally important), N total records	Scenario 1: N relevant records found (TP), F20 score (sensitivity is 20 times as important than precision), N total records
01101	Scenario 1: N relevant records found (TP), Accuracy/Error, N total records	Scenario 2: N relevant records found (TP), Work saved over sampling (WSS) at a sensitivity of 95%, N
10001	Scenario 1: Sensitivity/recall, specificity, N total records	Scenario 1: Sensitivity/recall, Positive likelihood ratio, N total records
10010	Scenario 1: Sensitivity/recall, N irrelevant records found (FP), N total records	none
10011	Scenario 2: Sensitivity of the process (yield), burden/gain, N	Scenario 2: Sensitivity/recall, absolute screening reduction, N
11000	Scenario 1: Sensitivity/recall, precision/positive predictive value (PPV), N total records	none
11001	Scenario 1: Sensitivity/recall, Negative Predictive Value (NPV), N total records	Scenario 1: Sensitivity/recall, Correlation coefficient, N total records
11010	Scenario 1: Sensitivity/recall, Number Needed to Read (NNR), N total records	none
11100	Scenario 1: Sensitivity/recall, F1 score (sensitivity and precision are equally important), N total records	Scenario 1: Sensitivity/recall, F20 score (sensitivity is 20 times as important than precision), N total records
11101	Scenario 1: Sensitivity/recall, Accuracy/Error, N total records	none

Notes: Scenario 1 in purple and scenario 2 in green. TP = true positive; FN = false negative; FP = false positive; TN = true negative; N = number.

2.2.3 Designing the survey

Ideally, we would compare every measure set in a scenario to every other, a full factorial design. For Scenario 2, there were only five measure sets and they corresponded to four attribute profiles. So, we were able to create a full factorial design, including all possible comparisons in six questions. However, for Scenario 1, a full factorial design would have required 91 questions to fully compare all 14 attribute combinations represented by the measures. We considered this unfeasible. Therefore, we chose to use a D-optimal design,[39] in which a subset of comparisons is selected based on an empirical evaluation of which attributes and interactions have the most significant effects. This design strategy allowed us to prespecify the number of questions and then automatically select the pairs of measure sets for each

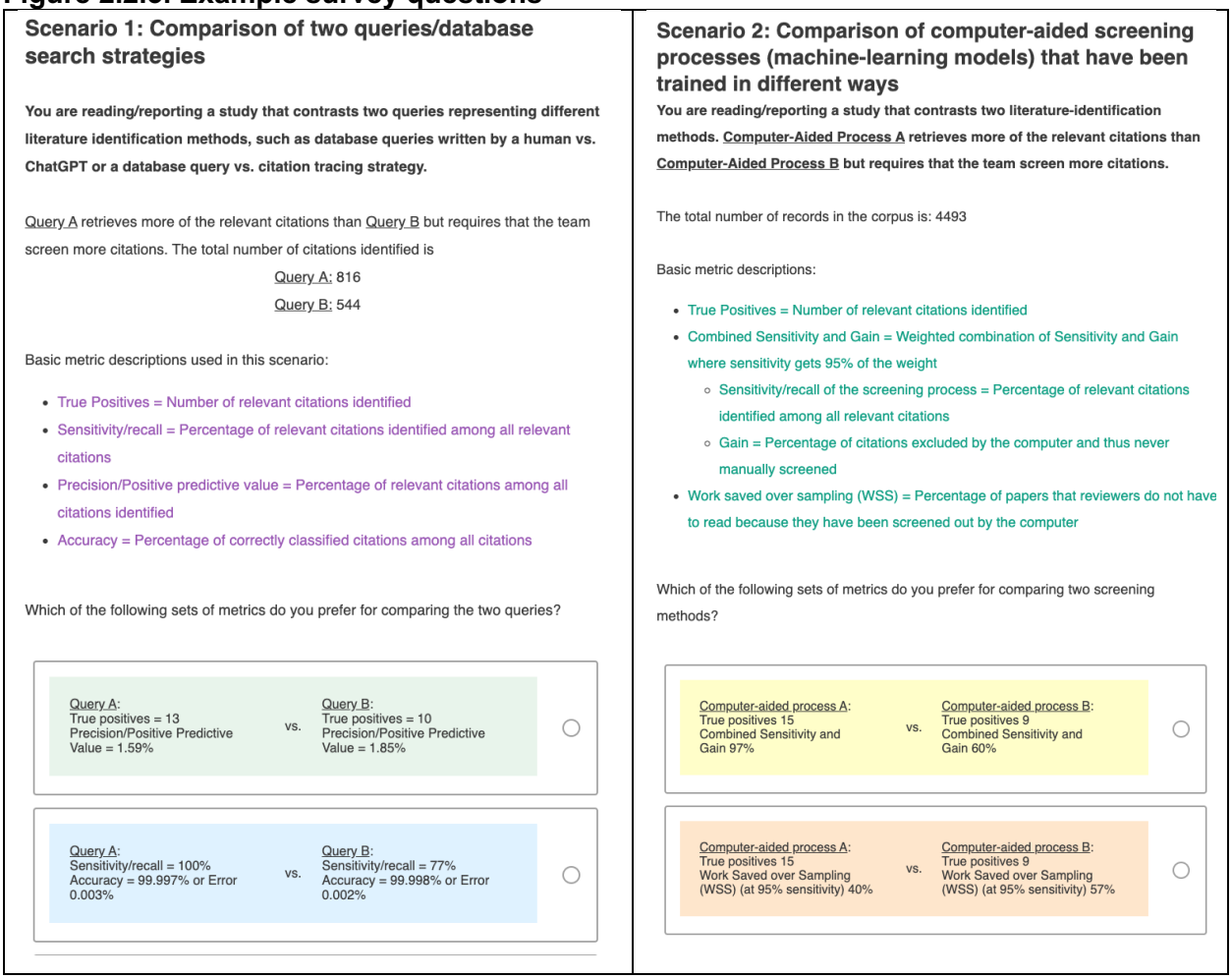
question. Empirically, the target size to reduce the prediction error should be greater than two to three times the number of attributes, which in this case would suggest a minimum of 10 to 15 questions.[40] We settled on 19 questions, as this is well above the minimum, and internal testing showed that a survey with 25 questions (the 19 for scenario 1 plus the six for scenario 2) could be completed in about 10 to 20 minutes.

The D-optimal design used to select the 19 set pairs from the full set of 91 initially created a series of submatrices containing 19 the 91 possible set pairs. It then evaluated the information matrix (variance/covariance matrix) of each submatrix and chose the one with the largest determinant.[41] The goal of this design is to create a subset that best covers the full range of possible attribute combinations.. Although, by definition, there is a level of information loss with this method.

The D-optimal design yielded a set of 19 ordered set pairs for Scenario 1, and the full factorial design yielded 6 unordered questions for Scenario 2. We implemented this design through a survey that included 25 questions, in which the user was asked to choose between two measures represented by different profiles (Figure 2.2.3). The survey also included a question about the user's professional role as it pertains to their experience with these measures: librarian, systematic review methodologist, computer scientist/software engineer, or statistician).

Because every set has a non-negative chance of being chosen over every other set, we did not include any "honeypot" questions that would by definition have a "correct" answer. Thus, we assume that every choice represents that of a thoughtful user.

Figure 2.2.3. Example survey questions



2.2.4 Circulating the survey

We sent the survey to known authors of studies of tools, as well as to known librarians and systematic review authors. In addition to direct emails, the survey was circulated to list-servs for the Cochrane Information Retrieval Group, the Cochrane authors, the Campbell librarians' group, the Agency for Healthcare Research and Quality Evidence-based Practice Centers (EPC) librarians' group, the EPC directors, project leads, and program managers groups, and the Society for Research Synthesis Methodology between December 15, 2023, and February 9, 2024. The chance to win a $100 gift card or charity donation was offered to all survey completers. Finally, a link to the survey was shared via the platform X (formerly Twitter). A reminder email was sent on January 30, 2024, a week before the survey closed. The email text is in the appendix.

2.2.5 Analysis

The analysis for this project was exploratory, rather than hypothesis driven. Thus, we explored multiple subsets of the data and looked for patterns of choice behavior. We evaluated the data in two separate datasets: (1) data in which the respondent chose between measures for at least one question (referred to as "any response"); and (2) data in which the respondents chose a measure for every question (referred to as "complete responders"). We did not use the full data for any analysis, because the data in which a respondent did not select any measure for any question is not useful to understanding the choices between measures. We did not conduct any imputation. Finally, we stratified the data by respondent role (librarian, systematic review methodologist, computer scientist/software engineer, or statistician) to see if the results differed.

Our analysis is based on McFadden's random utility theory.[38] In this case, we posit that the respondent will map each measure set in a given pair implicitly to the five attributes and will choose the option that they consider the best alternative. The relative importance of the five attributes will guide the choice, allowing the analysis to identify the most important attributes across respondents. For example, when given the pair:

Set A: True positives + precision + total number identified
versus
Set B: Sensitivity + accuracy + total number identified

They may choose set B because attribute 1 (that the relevant records correctly identified be given as a ratio) is most important, and they weigh the differences in other attributes as less important. Conversely, a respondent who chooses set A, might consider the fact that it does not have attribute 5 (changes slowly under class imbalance) the most important factor. Based on the responses given by each participant, we can estimate the probability of them choosing another option with similar or differing attributes.[38]

To estimate these preference weights, we used a conditional, multinomial logit model. The conditional aspect refers to the fact that for every measure set chosen by a given participant, there is a second measure set not chosen, due to the experimental design. Thus, the sets in each pair are not independent and must be analyzed together. The model estimates the odds that a given attribute will be preferred based on the choices made across all respondents and all 20 pairs. These odds ratios make up the utility function, which defines how respondents choose measures.[38] However, there is also a random component, based on unobserved factors of the choice (for example, a user's familiarity with a given measure),[40] which is captured in the confidence intervals of each odds ratio.

We initially fit a model with only the main effects of the five attributes. This gave us the marginal utility of each attribute. To explore interactions among attributes, we specified a model with all interactions for which there was at least one measure in common. To select the interactions from this model that are most likely to guide preferences, we used Least Absolute Shrinkage and Selection Operation (LASSO) regression and cross validation as described by Reid and Tibshirani.[42] This analysis allowed us to simplify the model and better identify combinations of attributes with the strongest likelihood of being chosen. The LASSO analysis provided a series of possible models based on different regularization penalties for large coefficients. The regularization penalty keeps the model from putting too much weight on any attribute or combination of attributes that might be particular to the data. In the cross-validation step, we tested the LASSO-generated regularization penalties to minimize the error on predictions made from model parameters on sequentially held out sets of 10% of the data to establish which model performed best. We used the interactions that survived in the best performing model plus all of the main attributes in the final conditional logistic regression model.

We ran separate LASSO and cross-validation analyses, as well as final regression models, in the following datasets: complete responders, any response, any response limited to librarians, any response limited to systematic review methodologists. The number of respondents was too small to create datasets for any response limited to computer scientists/software engineers (N participants = 8) or statisticians (N participants = 6).

To establish which specific measures respondents were most likely to prefer based on the relative importance of the attributes, we cross multiplied the attribute profiles of each measure set with the weight (as an odds ratio) of the likelihood that each attribute or combination of attributes was preferred. This allowed us to use the weights identified for each attribute and to predict which measure sets would be more likely to be chosen by respondents, despite the fact that we were unable to ask them to compare every set as we would have had we done a full factorial design. We could also use these weights to predict the desirability of a new set of measures, such as, for example, the Diagnostic Odds Ratio discussed above.

2.3 Results

2.3.1 Response

The survey was circulated for nearly 2 months (December 13, 2023, to February 9, 2024). In that time 156 respondents recorded at least one choice for at least one scenario. Table 2.3.1 shows the breakdown of the respondents' self-identified roles (a respondent could choose more than one role) and Figure 3 shows the number of respondents, overall and by role, who answered each question.

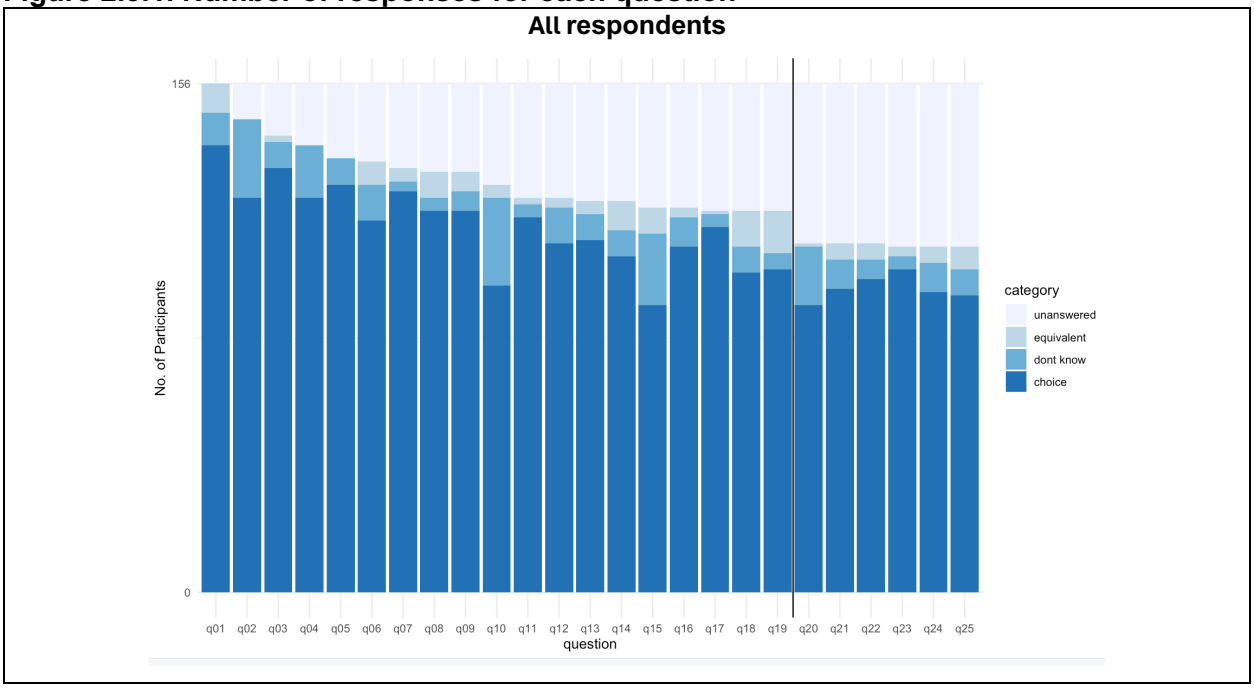
Table 2.3.1. Roles of the participants included in the analysis

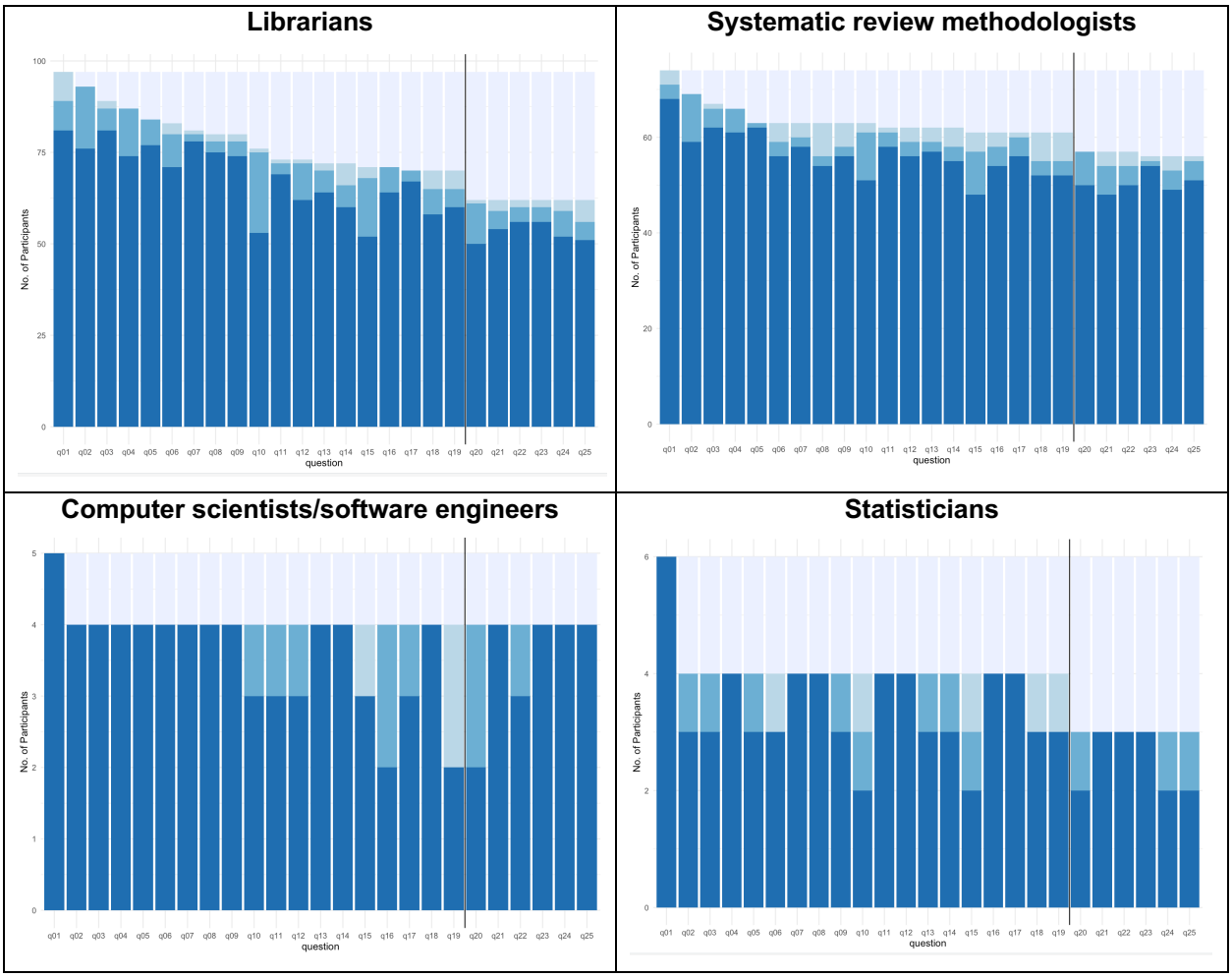
	Scenario 1		Scenario 2
Role	**Number, complete responders***	**Number, any response†**	**Number, any response**
Total	51	151	102
Librarian	38	114	74
Systematic review methodologist	42	95	73
Computer scientist/software engineer	0	8	2
Statistician	1	6	3

* complete responders chose a measure for every question, †any response chose between measures for at least one question. No one in scenario 2 made a choice for every question.

Figure 2.3.1 shows that the response rate declined across the survey by question. No respondent made a choice for every question across both scenarios. However, every question had a choice by a majority of respondents. We did not collect demographic data, so it was not possible to explore non-response rates in detail, but it appears that systematic review methodologists were somewhat more likely to answer every question. Librarians were less likely to answer the questions for scenario 2, which is not surprising, given that they often do not participate in subsequent steps of the systematic review after the literature search is complete. The numbers of computer scientist/software engineers and statisticians was so low that it is difficult to discern any patterns in their response rates, except to say that the following results are likely most applicable to tool consumers (librarians and systematic review methodologists) than to tool producers. The rest of the results are presented separately by scenario.

Figure 2.3.1. Number of responses for each question





Legend: Each blue line represents the response to a question. The darker blue lines represent the number of respondents who made a choice between measures for that question the medium blue lines represent that the respondents selected either "I don't know enough about the measure to make a choice" or "I consider the measures equivalent". The light blue lines indicate nonresponse to the question. The questions to the left of the black line are for scenario 1, the questions to the right for scenario 2.

2.3.2 Scenario 1. Comparison of two queries/database search strategies with all citations manually screened.

Scenario 1 includes 11 sets of measures, representing 14 attribute profiles. The results of a main effects conditional logistic regression model with no interaction terms are presented in Table 2.3.2-1. The results are the same in direction and similar in magnitude across everyone and for everyone, librarians, and systematic review methodologists. This analysis indicates that respondents prefer sensitivity over true positives and prefer measures that increase linearly with burden (i.e., "Did I find all relevant records, and how hard did I have to work to do so?"). There was also a more modest (though statistically significant) preference for measures calculated using true negatives. Respondents did not seem to care either way if the measure was affected by prevalence or was designed with a valuation between sensitivity and workload.

Table 2.3.2-1. Results of the main effects conditional logistic regression model

	Predicted odds of being chosen, OR (95% CI)			
Attribute	**Any response**	**Complete responders**	**Any response, librarians**	**Any response, systematic review methodologists**
1 (number vs. ratio of relevant records correctly identified)	**1.93 (1.74-2.13)**	**1.92 (1.66-2.22)**	**1.94 (1.70-2.22)**	**2.14 (1.86-2.46)**
2 (dependent on prevalence)	1.02 (0.89-1.16)	0.98 (0.81-1.20)	0.89 (0.75-1.06)	1.14 (0.94-1.38)
3 (explicitly weights sensitivity and workload)	0.88 (0.77-1.00)	0.86 (0.71-1.04)	0.96 (0.81-1.14)	0.83 (0.69-1.01)
4 (increases linearly with workload)	**2.09 (1.79-2.44)**	**1.88 (1.50-2.37)**	**2.30 (1.88-2.81)**	**1.82 (1.46-2.26)**
5 (changes slowly under class imbalance)	**1.28 (1.12-1.46)**	**1.26 (1.04-1.53)**	**1.05 (0.88-1.25)**	**1.44 (1.19-1.74)**
Likelihood ratio test	**334.6***	**137.9***	**234.2***	**183.8***
N observations	4264	1938	2586	2148
N people	151	51	93	71
R^2	0.151	0.137	0.173	0.164
AIC	2631	1215	1568	1315

* The likelihood ratio tests were all highly statistically significant with $p \leq 2\ e^{-16}$; Statistically significant results are in bold; AIC = Akaike information criterion; R^2 = R-Squared

We then explored whether a person's preference for one attribute affects their preference for another. For example, if measures that incorporate true negatives are preferred, do respondents prefer them to also be dependent on prevalence (e.g., accuracy) or not (e.g., specificity). The interactions we explored and the measures they share are presented in Table 2.3.2-2.

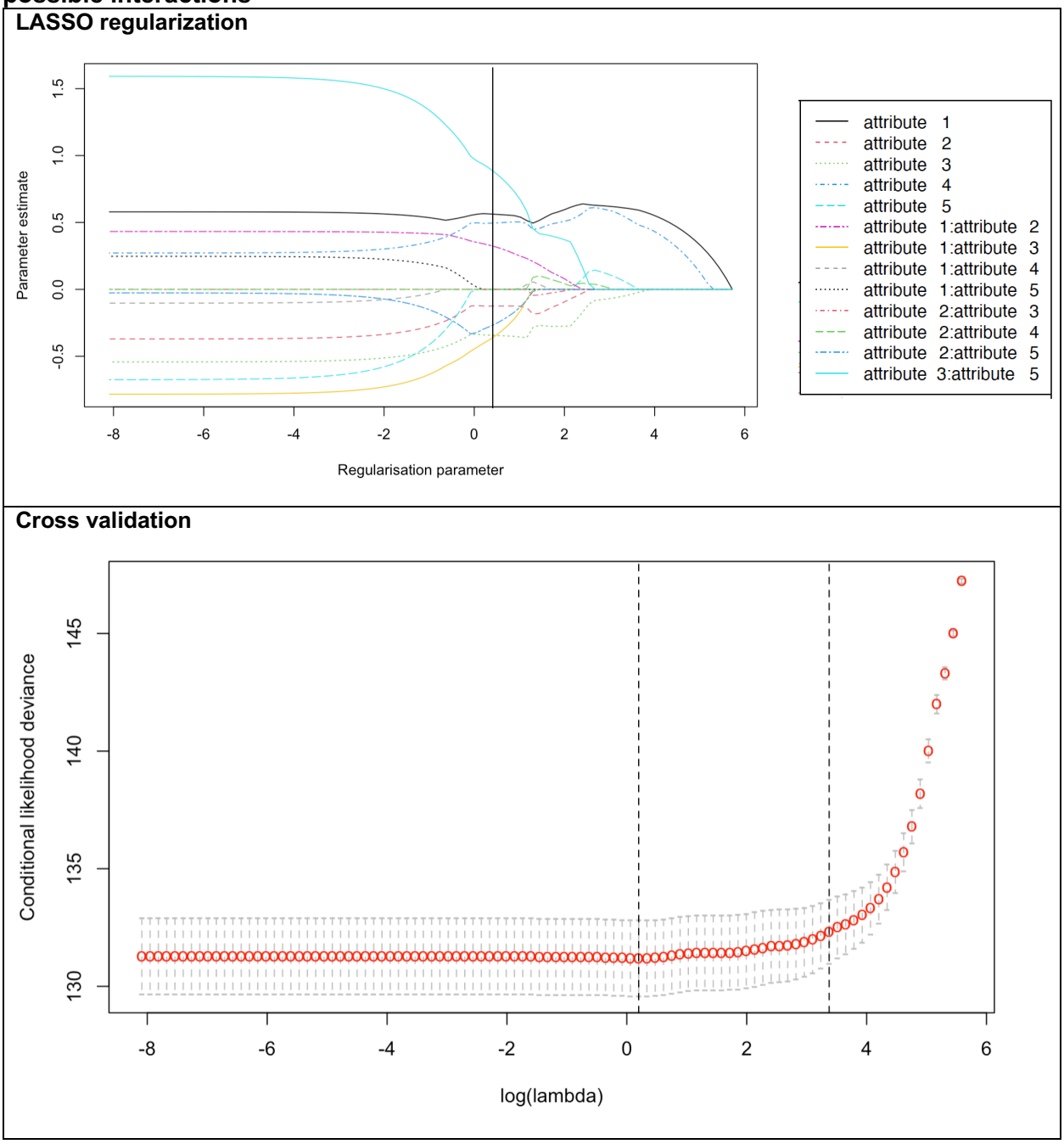
Table 2.3.2-2. Interactions explored and associated measures

First Attribute	Second Attribute	Common Measures
1 (number vs. ratio of relevant records correctly identified)	All others	All
2 (dependent on prevalence)	*3 (**e**xplicitly weights sensitivity and workload)*	F1/F20 score accuracy
	*4 (**i**ncreases linearly with workload)*	number needed to read
	*5 (**s**low changing measures)*	negative predictive value correlation coefficient accuracy
3 (explicitly weights sensitivity and workload)	*4 (**i**ncreases linearly with workload)*	no measures
	*5 (**c**hanges slowly under class imbalance)*	accuracy
4 (increases linearly with workload)	*5 (**c**hanges slowly under class imbalance)*	no measures

For variable selection and regularization, we ran the LASSO and cross-validation on all attributes and all interactions for which there were measures in common for each dataset. Because the results for the any response, complete responders, librarians, and systematic review methodologists groups were similar and yielded the same set of interactions for the final model, only the figures for the any response group are shown. Figures and results for the complete responders, librarians, and systematic review methodologists are shown in Appendix Figure 1 and Appendix Table 1. Figure 2.3.2 shows that as the regularization

parameter (penalty for model complexity and possible overfitting) increases from left to right, the model remains stable with all attributes and interactions until just before 0, at which point their estimates begin to converge to no effect (left panel) and the model performance on cross validation begins to decline (right panel).

Figure 2.3.2. Results of LASSO regression and cross validation on the model with all possible interactions



Legend: The top plot shows lines for the log odds ratio of each attribute and interaction from no regularization penalty for model complexity at left to maximum penalty at right. The horizontal black line represents the "best

model" based on regularization parameter values as established through cross validation (λ=0.20). Several attributes can be seen to converge to 0 before the best model. The bottom plot shows the cross-validation results, again with the full model at left, and the model with one variable at right. In this plot the left vertical dashed line shows the regularization parameter value (λ=0.20) for the best model, and the right dashed line shows the regularization parameter value (λ=3.65) for the most parsimonious model within 1 standard error of the best model.

The best regularization parameter based on the cross-validation analysis was 0.20 and yielded the estimates presented in Table 2.3.2-3. The most parsimonious model retained only three attributes (1: sensitivity measures, 3: combination measures, and 4: workload reduction measures). For our final regression, we chose to retain all attributes and the interactions that were not eliminated in the best model.

Table 2.3.2-3. Results of LASSO and cross-validation analysis

Variables	Best model, OR	Most parsimonious model within 1 SE of best model, OR
1 (number vs. ratio of relevant records correctly identified)	1.76	1.81
2 (dependent on prevalence)	0.88	1
3 (explicitly weights sensitivity and workload)	0.71	0.98
4 (increases linearly with workload)	1.64	1.62
5 (changes slowly under class imbalance)	1	1
Interaction of 1 and 2	1.41	1
Interaction of 1 and 3	0.67	1
Interaction of 1 and 4	1	1
Interaction of 1 and 5	1	1
Interaction of 2 and 3	1	1
Interaction of 2 and 4	1	1
Interaction of 2 and 5	0.74	1
Interaction of 3 and 5	2.55	1

Best model = the model with the least amount of deviation on cross-validation.

The final model is presented in Table 2.3.2-4. Again, the results are similar for complete responders and any response, and generally agree in direction for subgroups based on role. The only main attribute that remains statistically significant in this model is the preference for sensitivity. In terms of interactions, the interaction of 3 and 5 is large and statistically significant for all subsets, and the interaction of 1 and 2 is close to or just statistically significant for most groups.

Table 2.3.2-4. Results of the final conditional logistic regression model across datasets

	Predicted odds of being chosen, OR (95% CI)			
Attribute	**Any response**	**Complete responders**	**Any response, librarians**	**Any response, systematic review methodologists**
1 (number vs. ratio of relevant records correctly identified)	**1.80 (1.31-2.47)**	**1.72 (1.08-2.74)**	1.48 (0.97-2.24)	**2.50 (1.58-3.94)**
2 (dependent on prevalence)	0.68 (0.35-1.33)	0.77 (0.29-2.04)	0.42 (0.17-1.03)	0.81 (0.32-2.06)

	Predicted odds of being chosen, OR (95% CI)			
Attribute	**Any response**	**Complete responders**	**Any response, librarians**	**Any response, systematic review methodologists**
3 (explicitly weights sensitivity and workload)	0.57 (0.31-1.05)	0.55 (0.23-1.30)	0.53 (0.23-1.19)	0.53 (0.23-1.23)
4 (increases linearly with workload)	1.24 (0.64-2.39)	1.27 (0.49-3.27)	1.05 (0.43-2.54)	1.04 (0.41-2.60)
5 (changes slowly under class imbalance)	0.59 (0.17-2.05)	0.77 (0.13-4.49)	0.33 (0.06-1.71)	0.58 (0.10-3.21)
Interaction of 1 and 2	**1.57 (1.01-2.46)**	1.36 (0.70-2.61)	**2.35 (1.31-4.22)**	1.25 (0.67-2.34)
Interaction of 1 and 3	**0.54 (0.31-0.91)**	0.81 (0.37-1.78)	**0.44 (0.22-0.88)**	0.47 (0.22-1.00)
Interaction of 2 and 5	0.99 (0.34-2.87)	0.88 (0.20-3.90)	1.27 (0.30-5.31)	1.03 (0.24-4.39)
Interaction of 3 and 5	**4.38 (1.67-11.48)**	3.16 (0.78-12.73)	**6.62 (1.83-23.98)**	**5.39 (1.38-21.02)**
Likelihood ratio test	**351.7***	**144.8***	**252***	**193.3***
N observations	4264	1938	2586	2148
N people	151	51	93	71
R^2	0.158	0.144	0.186	0.172
AIC	2622	1216	1558	1314

* The likelihood ratio tests were all highly statistically significant with $p \leq 2\ e^{-16}$; Statistically significant results are in bold; AIC = Akaike information criterion; R^2 = R-Squared

We also ran a model that incorporated librarians as a factor. We were interested in the interaction of respondent role and attributes (i.e., how different professionals would value attributes differently from each other). With relatively few data for computer scientists/software engineers and statisticians, we focused on whether librarians differed from others, primarily from systematic review methodologists, which formed the other large group. The full results of this analysis are presented in Table 2.3.2-5. In general, the main effects and attribute interactions are similar to those shown in the other models, but it explicitly shows that librarians dislike measures that are dependent on prevalence (attribute 2) and measures that change slowly under class imbalance (attribute 5) compared with other roles.

Table 2.3.2-5. Results of the final conditional logistic regression incorporating librarians vs. other roles

	Predicted odds of being chosen, OR (95% CI)	
Attribute	**Any response**	**Complete responders**
1 (number vs. ratio of relevant records correctly identified)	**1.84 (1.31-2.59)**	**1.75 (1.07 - 2.87)**
2 (dependent on prevalence)	0.84 (0.41-1.69)	0.92 (0.33 - 2.54)
3 (explicitly weights sensitivity and workload)	0.51 (0.27-0.95)	0.52 (0.21 - 1.26)
4 (increases linearly with workload)	1.09 (0.54-2.19)	0.96 (0.36 - 2.61)
5 (changes slowly under class imbalance)	0.82 (0.23-2.94)	1 (0.16 - 6.12)

Attribute	Predicted odds of being chosen, OR (95% CI)	
	Any response	**Complete responders**
Interaction of 1 and 2	1.53 (0.98-2.41)	1.31 (0.68 - 2.55)
Interaction of 1 and 3	**0.55 (0.32-0.95)**	0.82 (0.37 - 1.81)
Interaction of 2 and 5	0.95 (0.32-2.79)	0.88 (0.19 - 3.98)
Interaction of 3 and 5	**4.35 (1.63-11.6)**	3.12 (0.74 - 13.07)
Interaction of 1 and librarian role	0.99 (0.80-1.21)	1.03 (0.77 - 1.39)
Interaction of 2 and librarian role	**0.74 (0.56-0.97)**	0.73 (0.49 - 1.10)
Interaction of 3 and librarian role	1.24 (0.94-1.63)	1.14 (0.77 - 1.71)
Interaction of 4 and librarian role	1.27 (0.92-1.75)	**1.71 (1.07 - 2.73)**
Interaction of 5 and librarian role	**0.62 (0.47-0.82)**	**0.63 (0.42 - 0.94)**
Likelihood ratio test	**385.5***	**174.2***
N observations	4264	1938
N people	151	51
R^2	0.173	0.172
AIC	2598	1197

***** The likelihood ratio tests were all highly statistically significant with $p \leq 2\ e^{-16}$; Statistically significant results are in bold; AIC = Akaike information criterion; R^2 = R-Squared

Tables 2.3.2-6a and 2.3.2-6b show a comparison of the performance across the models tested in the two datasets (any response and complete responders). In both datasets, the more complex models have better performance on all parameters than the simpler models. Although no model explains a large proportion of the variability in responses by question, it is clear that the measure attributes explain a significant amount of the odds that one measure is preferred over another. These results also suggest that the role of the respondent (librarian vs. systematic review methodologist or other roles) accounts for some of the variability not explained by the measure attributes alone, but, given the low R^2 values, there are clearly other factors not accounted for in the models.

Table 2.3.2-6a. Model performance in any response data (4264 total observations)

	Main effects model	**Final model with interactions**	**Final model with role**
R^2	0.151	0.158	0.173
AIC	2631	2622	2598
LR Test compared with intercept-only model	333.6 ($p \leq 2$ e-16)	351.7 ($p \leq 2$ e-16)	385.5 ($p \leq 2$ e-16)
LR Test compared with Main effects model	-1310.5	-1301.9 ($p \leq 1.8$ e-3)	-1285.1 ($p = 7.3$ e-8)

AIC = Akaike information criterion; R^2 = R-Squared

Table 2.3.2-6b. Model performance in complete responders data (1938 total observations)

	Main effects model	Final model with interactions	Final model with role
R^2	0.137	0.144	0.172
AIC	1938	1938	1197
LR Test compared with intercept-only model	137.9 ($p \leq 2$ e-16)	144.8 ($p \leq 2$ e-16)	174.2 ($p \leq 2$ e-16)
LR Test compared with Main effects model	-602.71	-599.24 ($p = 0.1386$)	-584.57 ($p=3.54$ e-5)

AIC = Akaike information criterion; R^2 = R-Squared

Based on the regression coefficients of the final model with interactions in the any response dataset, we calculated the odds that each set of metrics would be chosen by a participant. Table 2.3.2-7 shows the results of that analysis, ranked from the most to least likely to be chosen by anyone. These odds ratios are compared with the odds of choosing sensitivity/specificity, as this is a common measure that is well known to everyone.

This analysis indicates that sensitivity and number needed to read (NNR) are the most meaningful measures. There is also a strong preference for sensitivity/recall and number of irrelevant records found, as well as sensitivity and precision/positive predictive value (which is the inverse of NNR). There seems to be a general dislike of the F-score measures and true positives.

It is clear that the preference for sensitivity drove much of respondent's choices. Within sensitivity, the top burden measures were NNR, FP, precision, and accuracy. Within true positives, they were FP, accuracy, NNR, and precision.

Table 2.3.2-7. Predicted likelihood of selection for each measure set as compared with sensitivity/specificity

	Predicted odds of being chosen, OR (95% CI)		
Measure set	**Overall**	**Librarians**	**Systematic Review Methodologists**
Sensitivity/recall, *Number Needed to Read (NNR),* *N total records*	**2.23 (1.64-3.02)**	**3.20 (2.14-4.77)**	**1.82 (1.17-2.84)**
Sensitivity/recall, *N irrelevant records found (FP),* *N total records*	**2.08 (1.13-3.84)**	**3.23 (1.41-7.36)**	1.80 (0.77-4.22)
Sensitivity/recall, *precision/positive predictive value (PPV),* *N total records*	1.80 (0.84-3.87)	**3.05 (1.10-8.43)**	1.76 (0.60-5.15)
Sensitivity/recall, *Accuracy/Error,* *N total records*	1.43 (0.70-2.90)	1.92 (0.75-4.93)	1.40 (0.52-3.77)
N relevant records found (TP), *N irrelevant records found (FP),* *N total records*	1.16 (0.60-2.23)	2.19 (0.91-5.27)	0.72 (0.29-1.80)

	Predicted odds of being chosen, OR (95% CI)		
Measure set	**Overall**	**Librarians**	**Systematic Review Methodologists**
Sensitivity/recall, Negative Predictive Value (NPV), N total records	1.06 (0.58-1.04)	1.26 (0.56-2.80)	1.04 (0.45-2.39)
Sensitivity/recall, Correlation coefficient, N total records	1.06 (0.58-1.04)	1.26 (0.56-2.80)	1.04 (0.45-2.39)
Sensitivity/recall, specificity, N total records	1.00 (ref)	1.00 (ref)	1.00 (ref)
Sensitivity/recall, Positive likelihood ratio, N total records	1.00 (ref)	1.00 (ref)	1.00 (ref)
N relevant records found (TP), Accuracy/Error, N total records	0.94 (0.48-1.85)	1.26 (0.51-3.12)	0.96 (0.37-2.48)
N relevant records found (TP), Number Needed to Read (NNR), N total records	**0.79 (0.63-0.99)**	0.92 (0.69-1.23)	**0.58 (0.42-0.81)**
N relevant records found (TP), precision/positive predictive value (PPV), N total records	0.64 (0.34-1.20)	0.88 (0.37-2.06)	0.56 (0.23-1.37)
N relevant records found (TP), specificity, N total records	**0.56 (0.40-0.76)**	0.68 (0.45-1.03)	**0.40 (0.25-0.63)**
N relevant records found (TP), Positive likelihood ratio, N total records	**0.56 (0.40-0.76)**	0.68 (0.45-1.03)	**0.40 (0.25-0.63)**
Sensitivity/recall, F1 score (sensitivity and precision are equally important), N total records	**0.55 (0.38-0.81)**	0.70 (0.43-1.15)	**0.44 (0.26-0.77)**
Sensitivity/recall, F20 score (sensitivity is 20 times as important than precision), N total records	**0.55 (0.38-0.81)**	0.70 (0.43-1.15)	**0.44 (0.26-0.77)**
N relevant records found (TP), Negative Predictive Value (NPV), N total records	**0.37 (0.21-0.66)**	**0.36 (0.17-0.77)**	**0.33 (0.15-0.72)**
N relevant records found (TP), Correlation coefficient, N total records	**0.37 (0.21-0.66)**	**0.36 (0.17-0.77)**	**0.33 (0.15-0.72)**
N relevant records found (TP), F1 score (sensitivity and precision are equally important), N total records	**0.36 (0.28-0.47)**	**0.46 (0.33-0.64)**	**0.30 (0.21-0.44)**
N relevant records found (TP), F20 score (sensitivity is 20 times as important than precision), N total records	**0.36 (0.28-0.47)**	**0.46 (0.33-0.64)**	**0.30 (0.21-0.44)**

Within the combinations (attributes that included two measure options), specificity was preferred to positive likelihood ratio (88% to 12%), negative predictive value was preferred to correlation coefficient (81% to 19%), and the F20 score (sensitivity is 20 times as important as precision) was preferred to the F1 score (sensitivity and precision are equally weighted) (66% to 34%).

2.3.3 Scenario 2. Comparison of computer-aided screening processes (machine-learning models) that have been trained in different ways, only some records manually screened

Scenario 2 includes six sets of measures (five measure combinations), representing four attribute profiles. Table 2.3.3-1 shows the number of times a measure with each attribute was chosen across all six questions, overall and by role. The librarians and systematic review methodologists had similar preferences as shown in Scenario 1. The numbers for the computer scientists/software engineers and statisticians are much smaller, so it is difficult to draw conclusions, but they appear to prefer the same metrics as those in the other roles.

Table 2.3.3-1. Rates of choice for each attribute across all six questions

Attribute	Overall, n/N (%)	Librarians, n/N (%)	Systematic review methodologists, n/N (%)	Computer scientists/ software engineers, n/N (%)	Statisticians, n/N (%)
1 (number vs. ratio of relevant records correctly identified)	220/288 (76%)	128/166 (77%)	116/152 (76%)	9/11 (82%)	8/9 (89%)
2 (dependent on prevalence)	108/278 (39%)	57/157 (36%)	67/155 (43%)	3/10 (30%)	2/7 (29%)
3 (explicitly weights sensitivity and workload)	198/551 (36%)	107/313 (34%)	115/302 (38%)	9/20 (45%)	7/14 (50%)
4 (increases linearly with workload)	361/567 (64%)	212/325 (65%)	187/302 (62%)	12/22 (55%)	8/16 (50%)

n = number of times the attribute was selected, N the total number of times the attribute was assessed

The results of a main effects conditional logistic regression model with no interaction terms are in Table 2.3.3-2. We could only include three attributes in this model because all measure scenarios include TNs in their calculations (attribute 5), and attributes 3 and 4 included non-overlapping sets, which lead to collinearity. Thus, the odds ratio for attributes 4 would be the inverse of that for attributes 3.

Table 2.3.3-2. Results the conditional logistic regression model with no interactions

	Predicted odds of being chosen, OR (95% CI)		
Attribute	**Overall**	**Librarians**	**Systematic review methodologists**
1 (number vs. ratio of relevant records correctly identified)	**2.40 (1.75-3.30)**	**2.29 (1.50-3.49)**	**2.68 (1.73-4.15)**

	Predicted odds of being chosen, OR (95% CI)		
Attribute	**Overall**	**Librarians**	**Systematic review methodologists**
2 (dependent on prevalence)	1.24 (0.91-1.68)	1.17 (0.77-1.76)	1.45 (0.96-2.19)
3 (explicitly weights sensitivity and workload)	0.56 (0.41-0.76)	0.50 (0.33-0.75)	0.61 (0.40-0.93)
Likelihood ratio test	98.84*	64.13†	50.1‡
N observations	1118	638	604
N people	102	58	55
R^2	0.169	0.191	0.159
AIC	682.0944	384.1001	374.5588

The likelihood ratio tests were all highly statistically significant with *$p \leq 2\ e^{-16}$; †$p \leq 7.7\ e^{-14}$; ‡ $p \leq 7.6\ e^{-12}$

The main effects conditional logistic regression model indicates that respondents value sensitivity over true positives, they do not much care if the measure is affected by prevalence, and they do not like combined measures, but they do like models to increase linearly with workload. Exploratory interaction analyses indicated that there are no interactions among attributes.

Table 2.3.3-3 shows the predicted likelihood of each measure being selected based on the conditional logistic regression results. The comparison for all measures is true positives and absolute screening reduction, which has an odds ratio of 1. As in scenario 1, the preferred measures separate sensitivity from workload and present workload in a way that increases linearly. From this analysis, it seems that they want to know how many of the relevant records are found, ideally as a ratio (sensitivity) and how many abstracts do not have to be screened, either as a ratio (gain) or as a whole number (absolute screening reduction). This pattern holds for librarians and systematic review methodologists, with the exception that systematic review methodologists are more neutral about work saved over sampling. Respondents seemed to slightly prefer absolute screening reduction (N measures = 127, 58%) to burden/gain (93, 42%). Combined measures are considered less useful.

Table 2.3.3-3. Predicted likelihood of selection for each measure set as compared with true positives/absolute screening reduction

	Predicted odds of being chosen, OR (95% CI)		
Measure set	**Overall**	**Librarians**	**Systematic Review Methodologists**
Sensitivity of the screening process, Burden/Gain	**2.40 (1.75-3.30)**	**2.29 (1.50-3.49)**	**2.68 (1.73-4.15)**
Sensitivity of the screening process, Absolute screening reduction	**2.40 (1.75-3.30)**	**2.29 (1.50-3.49)**	**2.68 (1.73-4.15)**
True positives, Absolute screening reduction	1.00 (ref)	1.00 (ref)	1.00 (ref)
WSS at 95% sensitivity	**0.69 (0.51-0.94)**	**0.58 (0.39-0.87)**	0.89 (0.59-1.33)
Combined Yield and Burden (at 0.95 yield)	**0.56 (0.41-0.76)**	**0.50 (0.33-0.75)**	**0.61 (0.40-0.93)**

2.4 Discussion

This is the first study we know of that uses decision science methods to systematically establish which measures for literature identification tools or methods are most meaningful to stakeholders, including medical librarians, systematic review methodologists, tool developers, and statisticians. The goal of this analysis was to not only quantify existing measures, and which should be preferred, but also to establish which attributes of these measures make them preferred or not. Based on this analysis, it appears that systematic review methodologists and librarians would like to see studies report the measures used to establish whether the tool identifies all relevant records as a ratio, and they like measures that clearly indicate how much work is being saved. For example, precision/positive predictive value is a commonly reported measure, but it tends to be so small that it can be difficult to interpret; it is easier to conceptualize having to read 30 versus 60 records to find one that is relevant, than to interpret a difference in precision of 0.033 compared with 0.016.

This project has several limitations. First, the response from computer scientists and software engineers, who are often the developers of the tools and the ones reporting their performance, was disappointing, with only eight respondents total. Thus, this analysis is heavily defined by tool users. The fact that the results did not differ greatly based on all respondent or user subgroup suggests that at least for the few who responded, their preferences are similar to those of the users. Second, the attributes were determined by two of us and evaluated in discussions with three other experienced systematic review methodologists and librarians, but they were not formally elicited or tested on a larger group. Thus, it is possible that one or more of the attributes is mis-specified, that they are not sufficiently independent of each other, and that they do not sufficiently cover the decision-making space. Third, for scenario 1, we used a D-optimal design to reduce the number of comparisons for evaluation from 91 to 20. There is a risk that in doing so that we may miss important interactions among multiple attributes. Finally, we did not do an imputation analysis, nor did we analyze which respondents did not choose a metric for a given question. However, the results were similar for those who answered all questions and those who answered any questions, suggesting that this may not have had a major impact on findings.

2.5 Conclusions

The findings of this study suggest that, at a minimum, those evaluating methods or tools for literature identification in systematic reviews should report sensitivity as a percentage of reference standard records and separately a measure of workload that can be easily interpreted (such as number needed to read, absolute screening reduction, or gain). Given the preference for sensitivity as a percentage, it is likely that there would be a similar preference for absolute screening reduction as a percent of the entire corpus.

In the next chapter, we use these measures (sensitivity and number needed to read) to evaluate a novel tool that leverages language models to generate search queries from review titles and questions.

CHAPTER 3: *LITERATURE SEARCH SANDBOX*: A LANGUAGE MODEL THAT GENERATES SEARCH QUERIES

3.1 Introduction

Medical librarians serve an important role in developing unbiased, comprehensive search queries that appropriately balance the requirement to identify all relevant studies with time, resource, and budget constraints.[14] However, not every review team has access to an experienced librarian. And even for trained librarians, manual development of search queries is time-consuming.[15] Therefore, there is a need for tools that assist in the design and development of high-quality search queries. Training and evaluating models capable of powering such tools requires examples of natural language inputs (e.g., titles of systematic reviews) and structured search queries (e.g., Boolean queries one might issue to PubMed). We hypothesize that a model fine-tuned on actual search queries will perform better than proprietary large language models (LLMs) like ChatGPT used in zero-shot application (i.e., prompted, but not fine-tuned).

To train tools to generate systematic review search queries, we need data that incorporates textual information about the review, such as the title and questions it aims to answer, as well as real search queries. The existing evaluation collections that we know of are small, including only 25 to 94 reviews, and were not deemed usable for training a language model to generate queries. Specifically, the existing collections that we know of include the CLEF Technology Assisted Review collection,[43 44] two collections from Scells et al.,[45 46] and a collection of systematic review updates.[47] All but one of these datasets[46] comprise searches in Ovid format only, which limits their utility for this project. Ovid is a proprietary interface and little research is done directly through it.[48] Additionally, Ovid queries are composed of a series of numbered lines that are combined using Boolean operators, making them difficult for humans to read. Finally, Ovid-style queries cannot be used through the open-access (and widely used) PubMed interface (or in other databases) without extensive manual modification, and there is no standardized approach to manually modify an Ovid query into a PubMed query. The dataset we describe in this manuscript is considerably larger than existing datasets—comprising over 10,000 reviews—and it includes search strings that we manually translated into PubMed format.

Existing tools to assist in the identification of literature for systematic reviews either require a large amount of human input (for example, tools that use a set of previously identified studies to suggest terms

that must then be incorporated into the search query by the search designer)[36 49] or are complementary only to human search query generation (they suggest potentially relevant citations based on an unstructured query and/or a "labeled" set of identified studies). We were therefore concerned that the existing tools would not generate reproducible methods, and that it would be difficult to identify the source of performance issues.[32] In this work, we capitalize on recent advances in Natural Language Processing (NLP) by fine-tuning a Language Model (LM) to create systematic review search queries that are formatted like the search queries reviewers are accustomed to working with. These output queries can be easily evaluated and edited, and can be reported using a standard reporting format, allowing for peer review and improving reproducibility.[50 51]

3.2 Methods

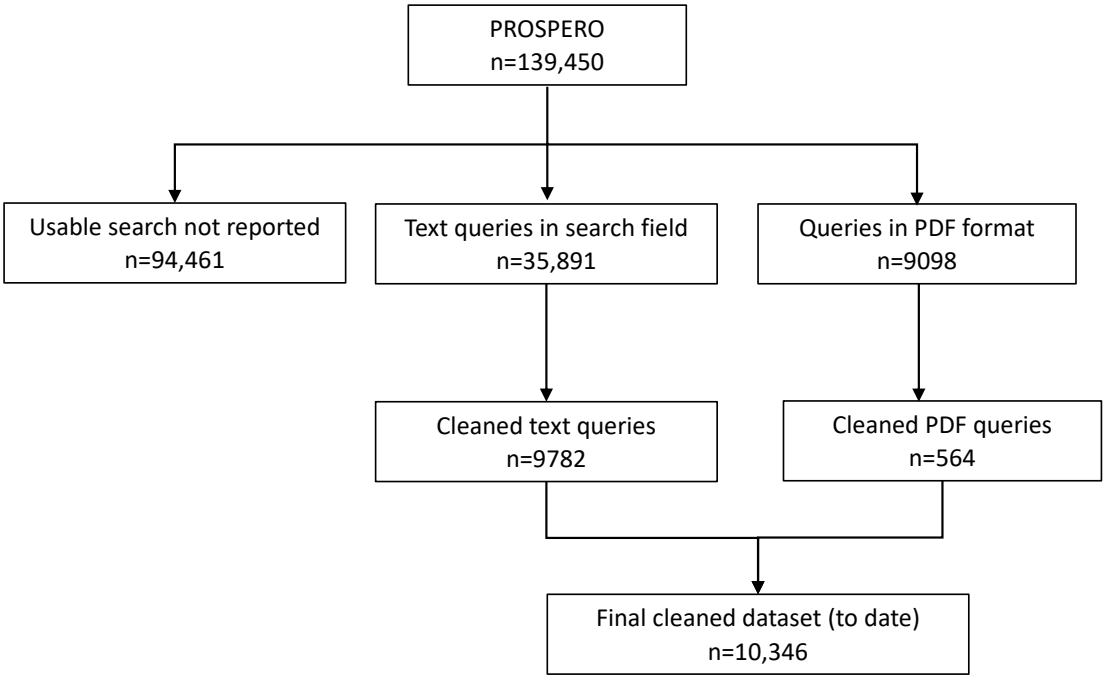
We first assembled two datasets of systematic review topics and search queries: one for training, representing 10,346 reviews, and one for evaluation, representing 57 reviews. From the training set, we separated out 99 reviews to use as a validation dataset, with which to choose the best model parameters. We then trained models on the (reduced) training dataset and, after settling on all hyperparameter settings using the validation dataset, evaluated them using the evaluation dataset.

3.2.1 Source and curation of the training dataset

PROSPERO is an international prospective register of systematic review protocols maintained by the Centre for Reviews and Dissemination at the University of York and funded by the UK National Institute for Health Research.[52 53] Full registration in PROSPERO requires the review team to fill out a minimum of 22 fields, including the review title, the review question(s), and the text of (or a link to) the review search query or queries. PROSPERO registration is highly encouraged by numerous, and required by some, journals that publish systematic reviews, therefore the registry has become both large and representative of published systematic reviews.

The PROSPERO team shared the data for 139,450 records, from inception (February 2011) to December 24, 2021. The fields shared were: RecordID, CRDAccessionNumber, RecordType (animal or clinical), Review title, Review question, Searches, Query, Condition, Participants, Intervention, and Keywords. These constitute a subset of the minimum required fields for registration. The flow of the cleaning process to date is shown in Figure 3.2.1. Data cleaning will continue for the foreseeable future and the dataset will be updated as more records are cleaned.

Figure 3.2.1. Flow diagram for PROSPERO queries to date



Much of the compilation and annotation of this dataset was done manually. Initially, we stripped the export of all HTML formatting, but otherwise made no changes to any fields other than the query. In the query field, we made the following changes:

First, we removed all methodological filters. This included study-type filters, date limits, and so on. This was done because methodological filters appear often, are easy to reproduce, and, because they are so common, might distract the LM from the more critical subject-specific terms.

Second, we subdivided the records into (a) those with links in the “Query” column to queries in PDF format, and (b) those without. For those with PDF links, all search cleaning was done manually by accessing the PDF, copying the query, translating it to PubMed syntax if necessary, fixing any obvious errors, and pasting the cleaned strategy into a new field entitled “pubmed”. Translating the syntax included such tasks as adding PubMed field tags instead of the Ovid backslash and “exp” (explode), replacing proximity operators with “AND”, and converting it from a numbered structure into a single search string that can be pasted into the PubMed search box. For records with data entered directly in the “Searches” field, we wrote an R[54] script to capture the search strategies nested within larger chunks of text, based on patterns of characters, such as opening and closing parentheses, and the words "AND", "OR", and "NOT" in proximity to parentheses. We then manually evaluated the output and fixed any remaining syntax. The cleaned strategies were pasted into the “pubmed” field.

Third, we saved two separate versions of the dataset: in one, the field tags were standardized (e.g., [Mesh], [Mesh terms], and [mh] were standardized to [Mesh]; [tiab] and [title/abstract] were standardized to [tiab]). In the second version, we stripped all field tags, quotation marks, backslashes, and commas, because these tags could be misapplied by the model, leading to faulty syntax and poor query performance. PubMed's interface automatically matches untagged words to their phrase index and to MeSH terms; thus, it is unlikely that this change would reduce the sensitivity of model-generated queries. However, it does have an impact on precision, increasing the number of records that need to be read (NNR) to find one relevant record.

Finally, we created a subset of 99 records to use as a validation set for model development and hyperparameter experiments. For this subset, we needed a reasonable number of relatively high-quality queries. Thus, we included a random selection of 10 queries from PDFs (which tended to be high-quality queries) and 89 of the best queries from those entered directly in the text field.

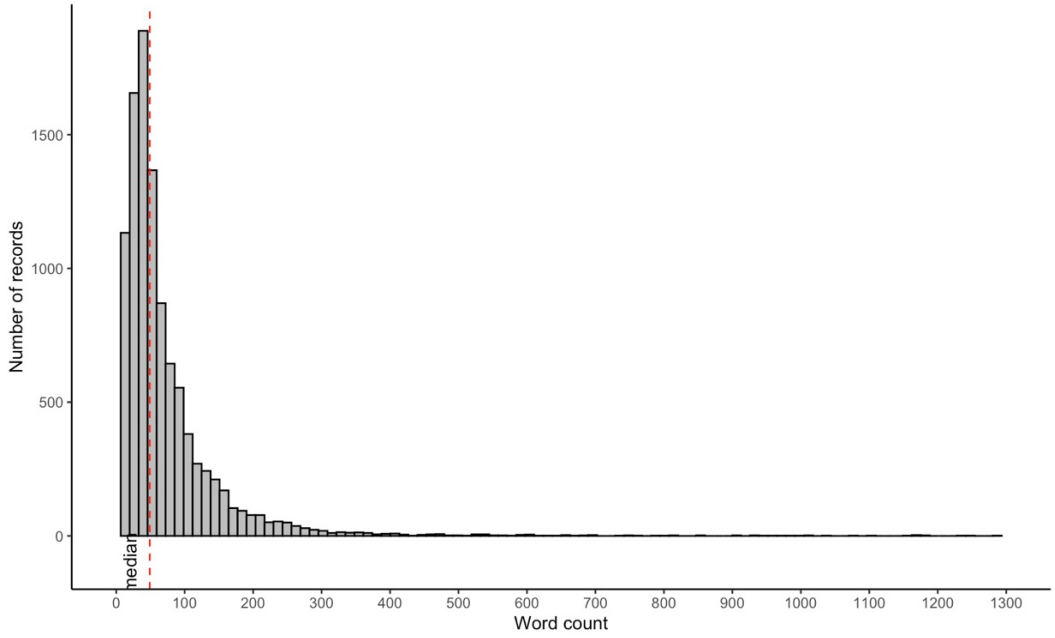
The current versions of the final datasets are in https://github.com/jayded/pubmed-query-generation

3.2.2. Analysis of the training dataset

The dataset currently includes 10,346 cleaned records with the following columns: RecordID, CRDRecordNumber, Review title, Review question, pubmed, Condition, Participants, Intervention, and Keywords. The following description pertains to the dataset stripped of tags; however, the patterns are similar for the dataset with tags.

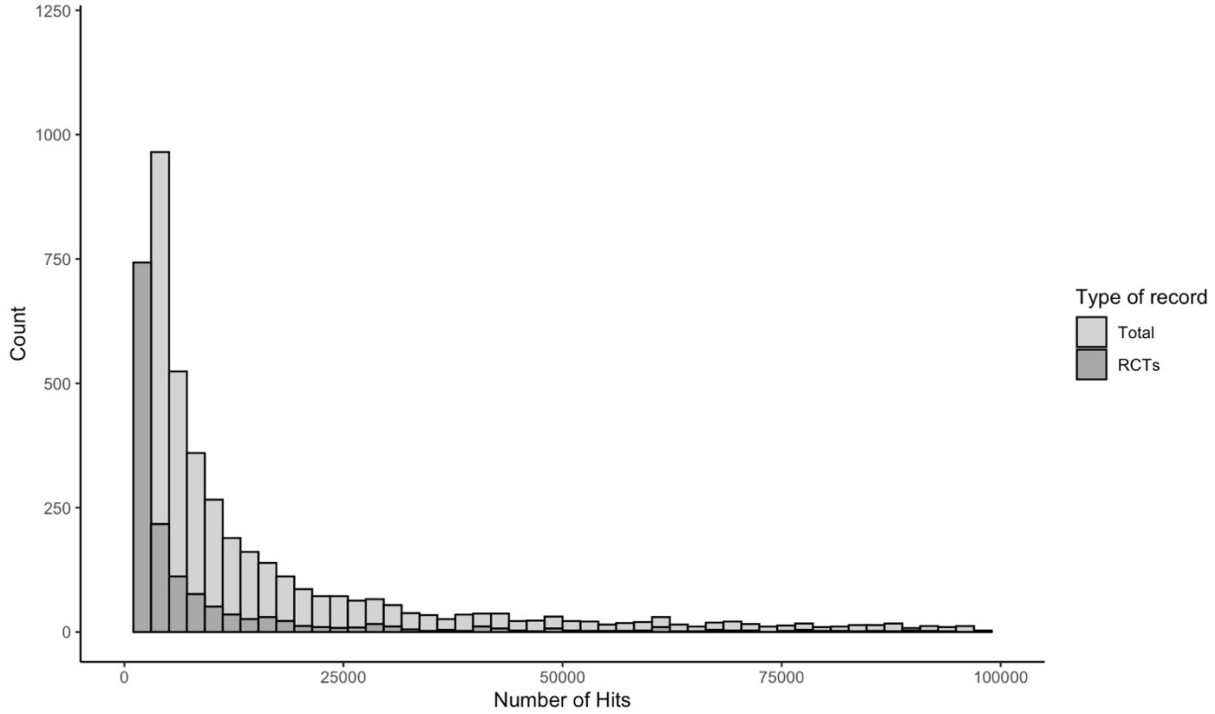
Queries are highly skewed with respect to the number of words they contain (Figure 3.2.2-1). Most searches include few terms, indicating that the training data is comprised primarily of broad searches. Search queries with few, generic terms will retrieve more irrelevant citations than longer queries with specific terms and more concepts connected with AND or NOT operators. The median number of terms is 49 (IQR 30-86; range 1-2,755). Almost all (95%) have fewer than 200 terms, and fewer than 1% have more than 400 terms.

Figure 3.2.2-1. Number of words per query



The total number of records identified by each query ranges from 0 to 23,180,504 (Figure 3.2.2-2); 597 queries return no records, which suggests they contain a fatal error; almost 200 queries return fewer than 100 records, indicating that they are unlikely to be adequate for a systematic review. Conversely, 653 queries return more than 100,000 records, also indicating that they are unlikely to be practically usable for a systematic review. These data are also highly skewed: 75% of searches return fewer than 7,000 citations. The median is 1,237 (IQR 208-86,682). To get a better sense of how much the removal of filters is increasing the size of the returned citation set, we crossed all searches with the PubMed randomized controlled trials filter. Many systematic reviews are limited to RCTs, so we thought this would be a good proxy for more realistic screening burden. The median number of RCTs retrieved is 57 (IQR 7 to 368, range 0 to 609,069), or about 5% of the median total number of records retrieved.

Figure 3.2.2-2. Number of hits per query



RCTs = randomized controlled trials

3.2.3 Source and curation of the evaluation dataset

The evaluation dataset comprised two separate subsets: the first contained 38 of the 40 citations in Wang's seed citation dataset[46] (which we will refer to as the "Wang dataset"). We removed two records from this set that retrieved either only one or no PubMed identifiers (PMID). To this, we added a second set of 19 reviews known to us (the "Adam dataset"), 18 of which were designed for systematic reviews conducted in the Agency for Healthcare Research and Quality's (AHRQ) Evidence-based Practice Center (EPC) program[55-71] or for the Department of Veterans Affairs (VA).[72] These searches were all peer reviewed by another systematic review librarian using the Peer Review Electronic Search Strategies (PRESS) checklist.[73] The final record in this set was for a Cochrane review that we were aware of; it is unclear whether this was peer reviewed.[74] We did not systematically search for additional reviews.

The fields include the review title and either the key questions for the EPC, VA, or Cochrane searches (in the Adam dataset) or the search description (in the Wang dataset), the reported PubMed search query for each review as developed by the review team (which we will refer to as the "human-generated query"), and the PMIDs of the final citations included in the review. For comparison with model-generated queries, we removed the tags, quotes, and commas from the data. While this means that the tool does not take over the entire task, it allows the tool to more effectively do the more challenging aspect of search query development (i.e., identify search terms and combine them correctly using Boolean operators),

leaving the matching of terms to database specific headings and attaching pretested filters (such as the Cochrane RCT filter) to manual post-processing (which could also be automated).

Very rarely would a query include only key words with no limitations by field, quotation marks, and filters, as our cleaned datasets have, so these numbers do not accurately represent real queries. In addition, we found that the sensitivity of the reported PubMed search queries (in the datasets) are not 100% for all searches. A study by Bramer et al. of 58 systematic reviews estimated that less than a quarter of the Medline queries (23%) identify all relevant citations, and only 40% identify 95% of the relevant citations.[75] Other relevant citations (even those with PMIDs) are often identified through queries of other databases (such as Embase and the Cochrane Register of Clinical Trials), by scanning the citation lists of other relevant systematic reviews, and by asking domain experts and peer reviewers. Thus, PubMed queries and searches are not expected to find all included studies.

3.2.4 Analysis of the evaluation dataset

The 57 reviews in the evaluation set (38 in the Wang dataset, 19 in the Adam dataset) include a wide variety of topics and range considerably in terms of the underlying review complexity. As shown in Table 3.2.4, the number of studies that were included in the systematic reviews ranges from 4 to 612, with a median of 30. The median sensitivity and precision of the human-generated evaluation search queries is 100%, with a median NNR of 580. The total number of hits and NNR are higher than they were before we stripped the data. The median NNR with field tags and punctuation in the original data was 125 (IQR 57-456) compared with 580 (IQR 161-,466) in the stripped data; and the median number of hits in the original data was 4,058 (IQR 811-21,416) compared with 22,812 (IQR 1,999-76,642) in the stripped data. The searches in the Adam dataset are more sensitive (median sensitivity 100%) than those in the Wang dataset (median sensitivity 91%). However, they are also less precise (median NNR 590 versus 526).

Table 3.2.4. Performance of search queries in the evaluation dataset (stripped of tags)

	Overall, median (IQR) [range]	Wang, median (IQR) [range]	Adam, median (IQR) [range]
Included studies, N	30 (11-74) [4-612]	12 (8-34) [4-137]	75 (56-93) [17-612]
PubMed hits, N	22,812 (1,999-76,642) [180-1,084,387]	12,078 (1,229-60,436) [180-271,423]	45,242 (21,799-78,759) [3,362 -1,084,387]
Sensitivity, %	100 (88-100) [34.2-100]	91 (84-100) [34.2-100]	100 (100-100) [85.2-100]
NNR	580 (161-3,466) [13-30,439]	526 (136-3,553) [13-30,439]	590 (282-836) [63-15,491]

SD = standard deviation, IQR = interquartile range, N = number, NNR = number needed to read

3.2.5 Details of model training

We initially trained models on both the tagged and stripped datasets, but the dataset stripped of tags performed significantly better in preliminary experiments based on manual evaluation of a subset of the produced strategies in comparison with the reference strategies, so we proceeded using only this version of the training data. We explored fine-tuning multiple models, including variants of Flan T5[76] and BioBART,[77] before settling on a Mistral-Instruct 7 billion parameter chat/instruct LM.[78] All fine-tuning experiments were conducted using the Huggingface library (v4.37.1),[20] an open-source library that provides a vast array of pre-trained models primarily focused on natural language processing. Instruction-tuned LMs have been trained to follow natural language instructions or templates.

We used the following instruction to the LM, which we added to all training samples (i.e., each review): "*Translate the following into a Boolean search query to find relevant studies in PubMed. Do not add any explanation. Do not repeat terms. Prefer shorter queries*." This instruction was followed by either the review's title or its title and objectives/key questions. We did not systematically evaluate whether different instruction prompts would have affected the output. Because we were fine-tuning the model, it is unlikely that the particular instruction used affected performance much; however, confirming this would have been computationally expensive.

Due to model size and resource limitations, we fine-tuned this model using Parameter Efficient Fine-Tuning,[79] specifically with Low Rank Adaptation (LoRA) methods.[80] Low Rank Adaptation inserts a small pair of matrices A and B, sized r by k (input dimensionality) and *d* by r (output dimensionality) and adds BA*x* to each linear layer output (where *x* is the *k*-dimensional input to the linear layer). Rather than updating all parameters in a model, LoRA trains only these smaller matrices, allowing for faster fine-tuning than updating all model parameters (potentially at some performance cost).

During fine-tuning, we used the adafactor optimizer[81] and ran and evaluated a grid search across learning rates (1e-4, 1e-5, 1e-6), and LoRA parameters (rank factor r: 16, 32, 64, 128, 256; scaling factor alpha: 16, r). For each hyperparameter combination we trained for 10 epochs with a batch size of 1 (owing to computational constraints), using floating point 16 precision. Although we monitored the training and validation losses, we used PubMed query results from the validation set to select the best model (for Review Titles); we chose the model with the highest recall between the validation set reference query and the model-generated query (0.54). For the title and key questions model, we used the same hyperparameters as the best title model.

When generating queries (i.e., for "decoding"), we used the following hyperparameters (passed to the Huggingface generate method[20]): maximum new tokens 2,048, minimum generation length 5, and no repeat ngram size 5 (banning repeats of the exact 5 tokens; this parameter was not ablated). Details of the parameters chosen are listed in Table 3.2.5.

Initial attempts to fine-tune Mistral-Instruct-7b tended to yield repetition in outputs. Prior to launching a grid search over hyperparameters, we performed an analysis on the training data for review title length and target query (stripped) lengths. Hypothesizing that the variation in inputs and outputs expected from the model would provide mixed signals for learning, we analyzed input and output lengths. We marked each piece of the training data as belonging to the middle 50% (the middle quartiles) and retained only data in both middle quartiles, leaving 4,324 instances for training. The code for the models is in https://github.com/jayded/pubmed-query-generation.

Table 3.2.5. Parameters used in fine-tuning the models

Parameter	Definition	Model specifications
Base model	Pretrained models build in "world knowledge," putting in weights for tokens	Mistral • 7 billion parameters • 32 layers
Number of trainable parameters	Number of variables or weights that the model adjusts during the learning process	63,488 = LoRA rank * model embedding size (size of a single LoRA matrix = 496) * 2 (one projects down, and one projects back up)
Optimizer	Uses the history to adjust the model's momentum and direction	Adafactor
Regularizer	Adds noise to labels to avoid overfitting the model	none
Learning rate	Determines speed of momentum in optimizer	0.00001
Decay	Measure of how the model performance declines when presented with unseen data	default
Batch size	Defines the number of random samples used for training	1
N epochs	How many times the training dataset is run through the model during training	6 (the 6th epoch, from 0-9, was the "best" by the NCBI recall measure)
LoRA alpha	Scaling factor for the weights in the LoRA matrices	16
LoRA rank	Determines the size of the LoRA matrices	64
LoRA dropout	Reduces the risk of overfitting	0.1
Decoding		Greedy = choose the most likely token at each prediction step
Maximum new tokens	Will make the model stop at this point, whether query is finished or not	2048 tokens
Minimum generation length	Every query has to have a minimum number of tokens	5 tokens
No repeat ngram size	Number of exact tokens that cannot be repeated	5

Val loss	Penalty for bad predictions in the validation data	1.21

LoRA = Low-rank adaptation; token = word part

3.2.6 Evaluation of the models

The final models were tested on the evaluation data described in section 3.2.3.We used the title and key questions or description for each review in the evaluation data to generate a query using each model described in section 3.2.5 (one trained only on the title, the second on the title plus the key questions). The resulting queries were run in PubMed and compared with the reference PMIDs (those included in the corresponding systematic review's final report) to calculate sensitivity (or recall; the number of included citations correctly identified) and NNR (the number of citations that must be read to identify one relevant citation).

Further analyses explored whether model performance differed by evaluation review source (Wang or Adam). The search description in the Wang dataset comprises only a few words, while the key questions in the Adam dataset comprise a full description of the questions the review seeks to answer. We also evaluated whether any queries "failed", defined as retrieving none of the review's final included PMIDs. Exceedingly high retrieval could also be considered failing. However, in this case, the fact that we removed all methodological filters, field tags, and quotes for all queries led to artificially high NNR, so that was not considered as a failure of the query.

A secondary analysis looked at the percentage of the relevant citations identified in the first 500 records based on PubMed's Best Match algorithm[82] for the reviews that had perfect sensitivity for all queries. This choice was made for ease of analysis, as queries with perfect sensitivity across all queries would, by definition, retrieve the same citations. PubMed's Best Match algorithm uses the term weighting function BM25 to initially weight records based on the terms in the user query, then re-ranks the top ranked citations using LambdaMART. These results were compared with the NNR to see if perhaps the low precision searches could be salvaged (i.e., whether the high sensitivity came from simply casting a very large net, or if, preferably, the search had relevant terms that would allow PubMed's algorithms to identify the most likely to be relevant studies accurately and rank them higher).

Finally, the MeSH terms, publication types, and years of publication of the citations missed by each strategy were retrieved and compared with those of the citations found to see if there were any patterns in the subjects, publication types, or years in the records missed by the search as compared with those correctly identified.

3.3 Results

3.3.1 Performance of the two trained models

The performance of the queries generated by the two models (trained on the title only and trained on title plus key questions) was similar. The title only median sensitivity was 85% (IQR 40% to 100%); title plus key question median sensitivity 86% (IQR 51% to 100%). The title only median NNR was 1,206 (IQR 205 to 5,810). The title plus key question median NNR was 908 (IQR 171 to 3,906). Table 3.3.1 shows the summarized results for each model on the evaluation set, both in total and for each of the two evaluation datasets.

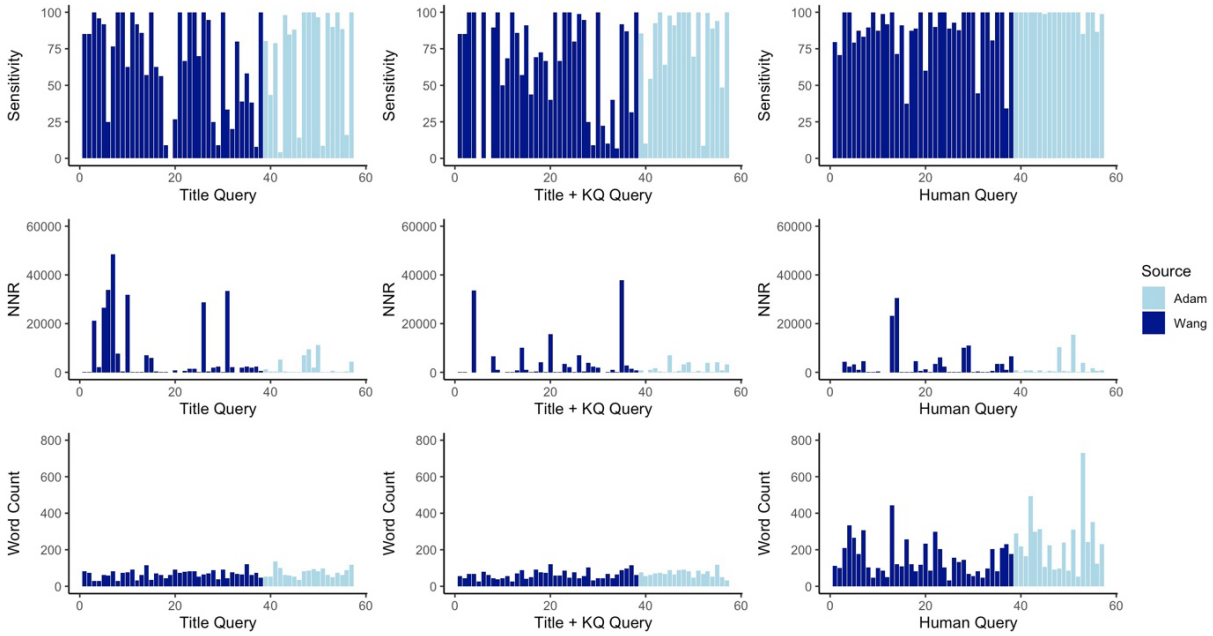
Table 3.3.1. Summarized results for each model on the evaluation set, Median (IQR)

Dataset	Human query*	Mistral-Instruct 7b trained on title and key questions query	Mistral-Instruct 7b trained on title only query
Overall			
Sensitivity, %	100 (88-100)	86 (51-100)	85 (40-100)
NNR, N	580 (161-3,466)	908 (171-3906)	1206 (205-5810)
Word count, N	144 (96-240)	66 (49-83)	71 (55-83)
Failed queries, N (%)	0 (0%)	2 (3%)	1 (2%)
Wang data			
Sensitivity, %	91.29 (84.29-100)	69.23 (40.94-99.74)	58.33 (38.34-100)
NNR, N	526 (135-3553)	1477 (155-4689)	1653 (115-7031)
Word count, N	120 (84-208)	65 (44-74)	67 (52-76)
Failed queries, N (%)	0 (0%)	2 (3%)	1 (2%)
Adam data			
Sensitivity, %	100 (100-100)	91.83 (68.09-100)	88.16 (69.93-99.75)
NNR, N	590 (282-836)	661 (188-3226)	331 (227-4588)
Word count, N	230 (116-304)	70 (57-84)	70 (56-90)
Failed queries, N (%)	0 (0%)	0 (0%)	0 (0%)

All measures median (Interquartile Range), except where explicitly noted; NNR = number needed to read, N = number or count
Failed queries = retrieved none of the review's included PubMed identifiers (PMIDs), Word count = number of words in the search queries.
* Human queries were the PubMed queries reported in the evaluation dataset that had been developed by the original review team. Note these numbers are the same as the overall numbers in Table 1.

Performance varied widely across the evaluation set reviews. For three reviews, at least one model identified none of the relevant citations; while for 20 reviews at least one model identified all of the relevant citations. Figure 3.3.1 shows the sensitivity, NNR, and word count for the queries generated by each model and the human query for each review.

Figure 3.3.1. Performance of models on each review



KQ = key question

No query failed to run in PubMed or retrieved no records, indicating that none had a fatal error. A few from each group retrieved excessively large numbers of records; even the gold standard queries had some with an extremely high NNR. This is likely because the queries were stripped of characters that would reduce the number of records retrieved, such as quotation marks and field tags. For two queries, the title and key question model query identified none of the PMIDs included in the systematic reviews (0% sensitivity), and the title only model query identified none of the included PMIDs for one topic.

For reviews that had 100% sensitivity across all queries (N=10), we assessed whether there was an association between the NNR and the percent of the total relevant citations in the first 500 records based on PubMed's Best Match algorithm. Across the 10 projects, the first 500 records included a median of 27% of the relevant citations for human queries (IQR 7% to 77%), 18% of the relevant citations for the title and key question model-generated queries (IQR 7% to 34%), and 12% of the relevant citations for the title only model-generated queries (IQR 2% to 60%). In general, the smaller the NNR, the larger the number captured in the first 500, though in one case the first 500 contained all of the relevant citations for all three queries, despite a wide range of NNRs.

3.3.2 Results of replication of Wang analysis

Wang et al.[33] investigated whether the large, general LLM ChatGPT-3 could be used to formulate Boolean search queries, tested against the CLEF TAR dataset and the same seed citation dataset used to evaluate the models in this paper. They found that this method had extremely poor sensitivity (between 4% and

13%) but good NNR of between 10 and 53 (calculated from their reported precision). However, it is worth noting that they retained the methods filters, which may have harmed sensitivity and improved precision/NNR. Critically, they reported that performance was highly correlated with the types of information given in prompts, with prompts that integrated examples of correctly worded queries greatly improving performance.[33] We re-ran the Wang analysis using GPT-4(-0314) to see if the performance had improved.

We used four of Wang's initial prompts:

Prompt 1: "For a systematic review titled [review_title], can you generate a systematic review?"

Prompt 2: "You are an information specialist who develops Boolean queries for systematic reviews. You have extensive experience developing highly effective queries for searching the medical literature. Your specialty is developing queries that retrieve as few irrelevant documents as possible and retrieve all relevant documents for your information need. Now you have your information need to conduct research on [review_title]. Please construct a highly effective systematic review Boolean query that can best serve your information need."

Prompt 3: "Imagine you are an expert systematic review information specialist; now you are given a systematic review research topic, with the topic title [review_title]. Your task is to generate a highly effective systematic review Boolean query to search on PubMed (refer to the professionally made ones); the query needs to be as inclusive as possible so that it can retrieve all the relevant studies that can be included in the research topic; on the other hand, the query needs to retrieve fewer irrelevant studies so that researchers can spend less time judging the retrieved documents."

Prompt 4: "You are an information specialist who develops Boolean queries for systematic reviews. You have extensive experience developing highly effective queries for searching the medical literature. Your specialty is developing queries that retrieve as few irrelevant documents as possible and retrieve all relevant documents for your information need. You are able to take an information need such as [example_review_title] and generate valid PubMed queries such as: [example_review_query]. Now you have your information need to conduct research on [review_title], please generate a highly effective systematic review Boolean query for the information need."

We also tested two new prompts to see if perhaps we could improve on their experience.

New Prompt 1: "You are a helpful information specialist who develops Boolean queries for aiding systematic reviewers. These queries attempt to search PubMed to find all studies relevant to a systematic

review topic. Your queries are targeted at finding studies, not reviews. Please draft me a short and reasonably comprehensive query searching for studies related to my work [review_title].

New Prompt 2: "You are a helpful information specialist who drafts Boolean queries for systematic reviewers. These queries attempt to search PubMed to find all studies relevant to a systematic review topic (other reviews are not relevant). Please draft me a short and reasonably comprehensive query searching for studies related to my work [review_title]".

The results are presented in Table 3.3.2. Interestingly, the Wang's prompts, when run in GPT4, had worse sensitivity than the GPT-3 results reported in the paper. Some prompt engineering produced prompts in GPT4 with similar sensitivity to those reported in the paper, but nothing close to sufficient for practical use.

Table 3.3.2. Summarized results for each model on the evaluation set

Dataset	Sensitivity, %	NNR, N
Wang GPT-4 (recreated for this analysis)		
Prompt 1	0 (0-5.55)	62 (27-202)
Prompt 2	0 (0-9.11)	26 (1-201)
Prompt 3	0 (0-2.55)	21 (3-163)
Prompt 4	0 (0-2.0)	87 (27-367.5)
Updated work		
New prompt 1 GPT-4	12.5 (0-45.4)	62 (23-151)
New prompt 2 GPT-4	5.9 (0-27.6)	29 (12-101)
Mistral-Instruct-7b trained on title and key questions	86 (51-100)	908 (171-3906)
Mistral-Instruct-7b trained on title only	85 (40-100)	1206 (205-5810)

Results are median (worst performance-best performance), NNR = number needed to read, N = number or count.

All queries for the regenerated prompts and the new prompts are in https://github.com/jayded/pubmed-query-generation

3.4 Discussion

The datasets we have compiled should be useful for exploring and testing automation methods for query generation, and perhaps other aspects of search design.

The model-generated queries yielded moderate sensitivity, but at present fall below the sensitivities required for use in systematic review searching, where the goal is to include *all* relevant studies. One of

the librarians I interviewed for the qualitative evaluation described them as a "sandbox", which seems appropriate. They are good enough to use in query development but not for stand-alone use.

Interestingly, the model-generated queries performed better on the Adam dataset (which was largely derived from systematic reviews for the AHRQ EPC program) than on the Wang dataset (which was mostly based on Cochrane reviews). In general EPC topics tend to be more complex than those in Cochrane reviews. However, it may be that the Wang topics were more challenging, given that the sensitivity of the human generated queries for this dataset was also lower. The precision of the human generated queries was also worse for the Wang data, though the precision of the human generated queries in the Wang data was slightly better.

The NNR data were less encouraging. The model-generated queries were simple in comparison with the human queries, including fewer terms and simple syntax (e.g., no nested Booleans; no "NOT" operators), and more complex queries are often more precise. However, machine learning has been shown to drastically reduce the number of abstracts that must be screened by humans. Abstract screening tools that iteratively rank records by their likelihood of relevance based on previously tagged human labels have been shown to reduce the screening workload by as much as 50% with minimal loss of relevant records.[26] Further research has shown the utility of training algorithms based on the 500 records considered most likely to be relevant by the PubMed algorithm.[83] These methods could be combined with automatic query generation to reduce the impact of the higher numbers of records identified.

3.4.1 Limitations

The search strategies in this dataset are likely to be a good representation of final publishable search strategies. But they are imperfect, as they were often the initial search for the review published in the protocol, rather than the final peer-reviewed query. The model might have performed better had it been trained on final queries, which are likely to have been of better quality. In addition, they may not always perfectly represent the original query posted in PROSPERO, as many were manually or automatically altered in the cleaning process. These searches should not in any way be used to form conclusions about each systematic review's query or its quality. Finally, we cleaned the queries with pasted search terms first, likely enriching the training dataset with shorter and more generic queries.

Models trained on the Population and Intervention/Exposure fields might perform better. We chose to train and test the model on only the review titles field and the objectives or key questions field for two reasons. First, not everyone is able to break a topic or set of questions down into its component parts, and a model that can do this will help with this process. Second, and related, the more the model requires the

user to break down the topic and propose Population and Intervention/Exposure terms, the less it actually helps in the process. Additionally, there is room for improvement based on query complexity, and possibly by decomposing queries into simpler parts and then expanding them.

We fine-tuned and evaluated only a single, representative open-source LM (Mistral-Instruct-7b). Our aim was to establish feasibility of generating search queries automatically via LMs, rather than to establish which particular LM performs "best" at this task. Larger models would likely yield better results, but we do not have the computational resources to fine-tune them at this time. Proprietary models, like ChatGPT, may also be used to aid search design, but these cannot be fine-tuned. Moreover, we are interested in designing and sharing usable, open-source LMs that can be used as search aids.

3.4.2 Results in context of related research

There is increasing interest in using machine learning and artificial intelligence to design or refine search queries. Several groups have experimented with using commercially available large language models, like ChatGPT and Claude.ai to write search queries, with little success. As shown in the results above, Wang et al. achieved only between 4% and 13% sensitivity over a series of prompts in GPT3, and when we re-ran their prompts in GPT4, we achieved a median sensitivity of 0. Minor prompt engineering with GPT4 got us to 6% to 13% sensitivity, but that is still much too low to be of any use in evidence synthesis. In a recent editorial, based on informal experimentation with GPT 3 and 4, Qureshi et al.[84] noted that ChatGPT produced plausible-looking queries, but on close examination the generated queries contained fabricated MeSH terms, and overly limited the results with a very narrow publication-type filter. The query they tested retrieved no citations when run in PubMed. Similar issues have been described on social media and among librarian groups.[84]

3.5 Conclusions

We have developed a relatively large dataset of natural language descriptions of reviews and corresponding structured Boolean searches, which can be used to train and evaluate LMs that map the former onto the latter. We demonstrated that an open-source and modestly sized LM (Mistral-Instruct-7b) can perform this task reasonably well.

Future research should focus on improving the dataset with more high-quality queries and assessing whether other fields, such as the population and intervention fields, would produce better performance. Nevertheless, it appears that pretraining a large language model on relevant strings improves its performance greatly over generic large language models. These models cannot replace careful and thoughtful query design, Nevertheless, they can be used to summarize a topic using Boolean logic,

providing suggestions for key words and the framework for the query. In conjunction with PubMed's Best Match algorithm, they may also be a good starting place for searchers either scoping a topic or looking for a non-comprehensive list of articles.

Toward this end, we integrated the Mistral models trained on the untagged data into an online application, using Anvil.[85] I then asked several experienced librarians from the AHRQ EPC program, the Brown University Biomedical library, and a hospital to test the app and give me feedback on the results. This qualitative study is the subject of Specific Aim 3.

CHAPTER 4. *"CALL IT A SANDBOX"*: A QUALITATIVE EXPLORATION WITH PRACTICING MEDICAL LIBRARIANS REGARDING THE USEFULNESS OF A LANGUAGE MODEL THAT DESIGNS SEARCH QUERIES FOR SYSTEMATIC REVIEWS

4.1 Introduction

Systematic review searches must be thorough, objective, and reproducible.[9] However, the number of journal articles and other reports of medical research (e.g., clinical trial records, conference abstracts, preprint articles) is increasing exponentially.[86-88] In addition to being time-consuming and expensive, the current process may introduce a potentially significant bias; namely, that searches are artificially limited to return only the number of citations considered feasible for the team to screen. Medical librarians serve an important role in developing unbiased, comprehensive search queries that appropriately balance the requirement to identify all relevant studies within the constraints of the team's time and budget.[14] However, not every review team has access to an experienced librarian, and, even for trained librarians, the development of search queries is time-consuming.[15] Thus, there is a need for tools that assist in the design and development of high-quality search queries. A recent review of the tools in The Systematic Reviews Toolbox,[28] a searchable repository of tools that can be used during various stages of systematic review production, returned 102 tools related to the automation of searching, of which 21 were directly applicable to search query design.[29]

Machine learning models—language models, in particular—may assist with the crafting of high-quality search queries for systematic reviews. Language models are defined as models, trained on large amounts of text to be able to predict which words generally follow specific sequences of other words. For specific aim 2, we fine-tuned two Mistral Instruct 7b language models on more than 10,000 real systematic review search queries from the PROSPERO database, from which we had stripped field tags (e.g. [MeSH], [tiab]), quotation marks, and other punctuation.[53] One model was fine-tuned with the title as input and the search query as output; the other was fine-tuned on the title and key questions as input and the search query as output. In a quantitative analysis, the models had a median sensitivity of approximately 85% and required that the team screen approximately 1,000 abstracts for every included citation. Full details of the dataset, model training processes, and the full quantitative analysis are in Chapter 3.

The uptake of automation in systematic review literature identification is low. According to O'Connor et al., there are two primary reasons for this.[37] First, few tools find all relevant articles, though many may be good enough to be used as an adjunct to traditional methods. Second, existing tools are not easily integrated into current workflows.[37] When developing a tool to semi-automate any aspect of literature identification, it is important to understand how it will be used. A group of systematic review librarians, including me, recently evaluated the usability of query-generation related tools in The Systematic Reviews Toolbox[28] across the steps of search query generation. We found many tools that were helpful in generating search terms, matching keywords to subject terms (such as MeSH terms for MEDLINE or Emtree terms for Embase), or evaluating searches by graphically presenting the relationships among terms and concepts.[29] For that evaluation, we created tables that summarized the usability of each tool based on objective and subjective criteria (usefulness, ease of use, time to learn, access, user support, and import/export options) to help searchers quickly identify and appraise the usability of existing tools. However, in the process of developing these tables, we found that, generally, existing tools do not work well for complex topics, often requiring that the user break the search down into its individual components before using the tool, and that no tool was able to assist users in creating or testing their search logic.[29]

In developing a new tool for search query design, we believed that getting input from expert users, in this case medical librarians experienced in developing searches for systematic reviews, would inform implementation efforts, as well as future development. Thus, I conducted a series of semi-structured interviews with practicing medical librarians to assess how they would use the tool and how they would like to see it evolve. To see how librarians might implement the tool in their own workflows with searches they knew well, as well as to allow for in depth discussions, I chose to individually interview each librarian, rather than to convene a focus group or send a survey.

4.2 Methods

I used a qualitative design, involving virtual semi-structured interviews with medical librarians who tested a tool designed to generate search queries based on a review's title and/or key questions (hereafter referred to as the 'query tool') during the interview. The findings of this study are reported in accordance with the Standards for Reporting Qualitative Research (SRQR) Checklist.[89] Because this information was sought in the librarians' professional capacity, rather than their personal feelings and opinions, this study was deemed by the Brown University Office of Research Integrity Institutional Review Board (IRB) to not meet the federal definition of human subjects research. Nevertheless, efforts were taken to inform the

participants of the nature of the research, offer them the opportunity to quit at any time, and to ensure their anonymity and privacy.

4.2.1 Sample identification and Recruitment strategy

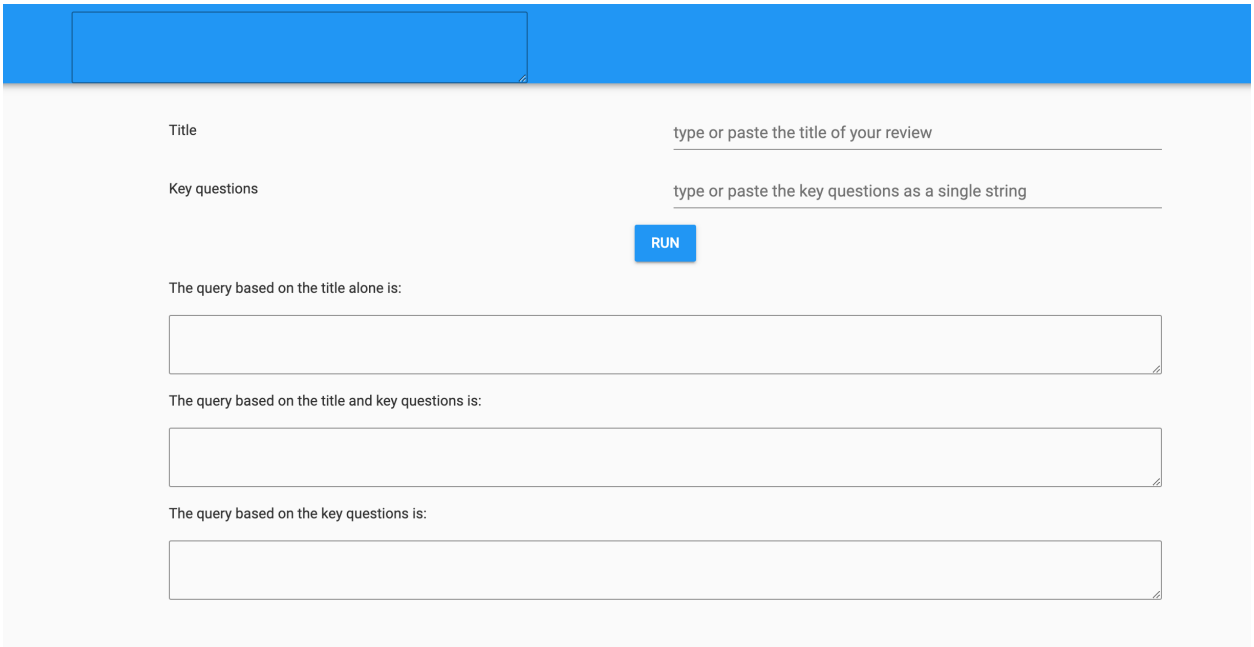
Eligible participants were medical librarians with experience searching in PubMed. I used purposive sampling based on the participant's work setting. This allowed me to understand a diversity of perspectives and to explore the applicability of the query tool in different medical library settings. Specifically, I sampled 1) systematic review librarians affiliated with Agency for Healthcare Research and Quality (AHRQ)-funded Evidence-based Practice Centers (EPCs) because their jobs fundamentally involve working with evidence synthesis teams to design comprehensive search queries; 2) university-affiliated academic librarians who support medical and public health students because they consult and teach literature identification for evidence synthesis; and 3) hospital-affiliated librarians because they work with front-line clinicians and policymakers who are implementing evidence-based healthcare. I aimed to recruit four EPC librarians, three academic librarians, and a hospital librarian, with a total recruitment goal of eight participants.

Participants were invited via email. Direct emails were sent to U.S. academic and hospital librarians known to the author, as well as to the EPC librarian listserv. A brief description of the project, eligibility criteria, and contact information were included in the email. The email sent to recruited librarians is in the Appendix. Librarians who were willing to participate replied to me via email, and we arranged for a 45-minute Zoom® call.

4.2.2 Interview

First, we created a web-based interface for the query tool.[85] The interface is simple with space for the user to enter the title and key questions for their review and outputs for the generated queries. Because there are two separate models, different queries are generated and appear in the boxes labeled "The query based on the title alone is:" and "The query based on the title and key questions is:" A third option for developing a query based on the key questions (alone) was presented in the interface but was not used. See Figure 4.2.2 for a screenshot of the interface.

Figure 4.2.2. Web interface for users to run queries in the fine-tuned Mistral Instruct 7b model



Librarians were not sent the query tool before the interviews. They were asked to bring a completed search for which they had the topic information, the search they had run, and any citations they used to evaluate the quality of their own query. During the interview, they were asked to share their screens and pilot the query tool on their topic. The tool generated two queries, one for the model trained on the title and one for the model trained on the title plus key questions. The key question only box was left blank because we did not train a model on key questions only. The librarian was then asked to evaluate one or both generated queries in the PubMed interface and compare it with their manually developed query. As they were doing so, I encouraged each participant to reflect on their experiences and asked a series of questions in a semi-structured interview format, guided by an interview guide that contained a series of questions, adapted from the 2015 Peer Review of Electronic Search Strategies (PRESS) Checklist Reviewer Form[73] for search strategy peer review.

The question topics included: 1) How well does the query reflect the translation of concepts to search terms? 2) Were Boolean operators used correctly? 3) Were relevant free-text terms appropriately included (and rejected)? and 4) Was the query structured appropriately? To address the barriers to uptake mentioned by O'Connor et al.,[37] a lack of trust in the performance of tools, and difficulty incorporating them into current workflows, the interview guide included questions about how the librarian might incorporate the query tool into their workflow, whether the query tool would be helpful or harmful for non-expert searchers when either beginning to think about a search or preparing to meet with a medical

librarian, and how the query tool could be improved to make it more useful. The full interview guide is presented in the Appendix.

4.2.3 Data collection

On average, the interviews lasted for 45-minutes and were conducted via Zoom® with cameras on and the participant sharing their screen. After an initial introduction to myself and the study, I informed the participant that all identifying information would be redacted and got their consent to participate and record the interview. I then began the recording. At the end of the interview, I invited the participant to share any other thoughts. Immediately after the end of the interview, I wrote notes about salient points made during the interview and themes that I noticed. These notes were not formalized but were used to inform the analysis of the transcripts. Interviews were conducted between February 6 and 16, 2024. After the interview, participants were invited to continue experimenting with the query tool and share any additional thoughts with me via email.

4.2.4 Analysis

I used a modified version of the framework analysis described by Gale et al.,[90] based on the framework provided by the PRESS checklist,[73] which is an adaptation of thematic analysis. Thematic analysis focuses on the development of themes, which involves "the systematic search for patterns to generate full descriptions capable of shedding light on the phenomenon under investigation."[90]

To ensure participant anonymity, each librarian was assigned a participant number. I transcribed all interviews verbatim using the Zoom® automatic transcription tool. I then reviewed each transcript against the original recording to ensure the accuracy of the transcripts. I also removed all identifying information from the transcript and any subsequent email communication. My notes and all email communication after the interview was anonymized and appended to the end of the edited transcript. Finally, I uploaded the final anonymized transcripts to the qualitative analysis software NVivo (version 1.7.1), with access limited to myself.

In NVivo, I coded each cleaned transcript by going through it and associating the blocks of text to a set of descriptive or conceptual labels, referred to as codes. Coding classified the data and allowed for the collocation of the pieces of text across multiple interviews that addressed the same concepts. Prespecified codes (deductive codes) were taken from the interview guide. New topics that emerged from the interviews were coded separately (inductive codes). The final themes are elucidated in Table 4.2.4, along with the underlying codes and whether those codes were deductive or inductive. The text in each code, aided by the notes I had taken immediately after the interview, were then charted into a framework

matrix, in which the data for each category was summarized into a short description of findings. In this way, the PRESS checklist concepts were reduced from four to three, and they were reordered. Most saliently, query structure emerged as the most commonly discussed concept, and it inextricably incorporated both concept translation and the correct use of Boolean operators. In addition, whether the model-generated queries led to appropriate MeSH term mapping within PubMed was discussed often and so was included as a PRESS concept that I had not initially thought would be relevant. The final summarized categories, with minimal interpretation, became the themes presented in the results, with quotes to illustrate as appropriate.

Finally, I checked the transcripts again to ensure that the themes I had developed made sense in the context of the full interviews. I also shared the findings with the participants to ensure that they were comfortable with their quotes and to elicit any comments on the findings. Four responded with suggestions, which were incorporated into the results. None of these suggestions changed the interpretation, but they better elucidated some of the nuances.

Table 4.2.4. Themes as derived from the coded text

Theme	Related codes (*code type*)	Definition
Theme 1: Sensitivity	**1.** Overall evaluation: **1a.** sensitivity (*deductive*)	Is the tool-developed query finding the relevant studies?
Theme 2: Size of the returned citation set	**1.** Overall evaluation: **1b.** burden (*deductive*)	Is the tool-developed query finding too many or too few citations?
Theme 3: PRESS checklist concepts **3a.** Query Structure **3b.** MeSH Term Matching **3c.** Key Words	**2.** PRESS evaluation: **2a.** Translation and structure (*deductive*) **2b.** Boolean and proximity (*deductive*) **2c.** Text words you would add (*deductive*) **2d.** Text words you would remove (*deductive*) **2e.** MeSH (*inductive*)	How does the query tool perform in relation to the quality measures included in the most common search query peer review framework?
Theme 4: Usefulness to searchers	**3.** Usefulness (*deductive*) **3a.** for non-expert searchers (*deductive*) **3b.** for librarians (*deductive*)	What are the ways in which the query tool is useful to librarians and how do they think it might be useful (or dangerous) for non-expert searchers?
Theme 5: Future directions **5a:** Better training searches **5b:** Split by PICO element **5c:** Interactivity	**4.** Future directions (*deductive*) **4a.** Improve training data (*inductive*) **4b.** Train on different data (*inductive*) **4c.** Post-processing (*inductive*)	What are some suggestions on how to improve the query tool to make it more useful?

Abbreviations: MeSH=medical subject headings, PICO = Population, Intervention, Comparator, Outcome, PRESS= Peer Review of Electronic Search Strategies

4.3 Results

I recruited eight experienced medical librarians, including four AHRQ EPC librarians, three academic librarians, and a hospital librarian. All agreed to participate. Across respondents, the median number of years working as a librarian was 19 (range 5-25) and the median number of searches done per year was 35

(range 20-250). Each librarian tested between one and three topics. The search topics varied widely, including the following:

- Reproductive health
- Violence
- Cancer
- Mental health
- Pain management
- Acupuncture
- Child development
- Irritable bowel syndrome
- Cognitive functioning
- Medical cannabis

Because the sample of each librarian type was small (one to four librarians), I could not attribute any specific themes to a specific librarian type. Thus, the themes below generally reflect opinions of the full group.

4.3.1 Theme 1. Sensitivity

The librarians considered overall sensitivity to be *"pretty darn good" [participant 1].* Several librarians agreed that the tool-developed queries might be useful in identifying relevant studies (as seed citations) that could be used in search query development. In some cases, the librarians were able to compare the tool-generated searches against a set of known relevant citations (seed citations). In looking at the seed citations that the tool-based query missed, one librarian commented:

> *"I can't say … that I found a single real article in [the missed citations] that focused on glioblastoma [the search topic]. So, I think that's pretty darn good."[Participant 1]*

Another agreed, noting:

> *"Executing a search in the tool using just the title retrieved 25 of 26 citations, and the one not retrieved is a narrative review that would potentially only serve as background material."[Participant 4]*

A third librarian, in looking at the first page of the results, commented that several were likely to be relevant, although they did not formally check that this search retrieved all the articles they found to be relevant.

> *"I am seeing what possibly could be relevant abstracts. For example, I know that this one here is actually one that I sent off to the requester, because it hits most of the hot spots."[Participant 8]*

This seemed to give the librarians a sense that the query tool could be useful to them, but they did note that the strategies generated by the query tool retrieved too may citations to be immediately useful.

4.3.2 Theme 2. Size of the returned citation set

The librarians were unanimous that the size of the retrieved corpus with the tool-generated queries was too large: *"it's also retrieving way, way, way, way too many citations." [Participant 7]* Across interviews, the numbers of citations found ranged from 2 to 10 times larger than the librarians' original search queries.

However, as they examined the queries in detail, the librarians noted that it would be easy to bring the numbers down by adding quotation marks or adjacency tags; methods, date, and other related validated filters; and/or by removing words and acronyms that increase the number of irrelevant citations retrieved:

> *"Pain by itself, of course, will blow [the search] up and may or may not bring in articles that are going to address [pain] management. … Librarians, we look at this, and we know immediately how to tighten it up right, how to make it run a little more efficiently [and] get rid of noise."[Participant 6]*

4.3.3 Theme 3. PRESS Checklist

The following sections describe some of the ways in which the tool-generated queries performed well and not so well, in the domains set out in the PRESS Checklist.[73]

4.3.3a Subtheme 3a. Search Structure

Because the interface links to two independently trained models, it returns two similar but distinct queries. Although not required to, all the librarians played with both searches, comparing them to each other and to their original search queries for the topic. For several topics, the query tool was able to generate queries with the concepts appropriately defined, as one librarian summarized:

> *"It does a half-decent job with…mapping the question to a basic search, at least identifying what are the different chunks"[Participant 7]*

Nevertheless, the query tool struggled with complex topics or implicit concepts, often missing conceptual chunks that could be used to decrease the number of search results returned. For example, in a search about history of trauma and reproductive health screening, the tool-generated query missed the implicit connection between history of trauma and screening behavior: the person's attitude toward health services. This attitude mediates the relationship between trauma and willingness to be screened and was an implicit aspect of the review topic. The librarian included a series of terms for this concept, which was combined with the other concepts using AND, which served to reduce the number of identified citations.:

> *"It's not stated in the question. So that's more of something that it's kind of like an invisible third. There's a history of trauma and there's reproductive health screening, but the missing connector between those is the attitude towards the health services."[Participant 7]*

A few librarians suggested that the title and key question searches could be used together by mixing and matching concepts, taking some of the terms and even conceptual pieces from each and putting them together into a new query: "So *one of the takeaways from this is, you've gotta play with both searches."[Participant 5]*

4.3.3b Subtheme 3b. Medical Subject Heading (MeSH) Term Matching

Because the models were trained on data stripped of field tags, the queries rely on PubMed's interface to match the keywords provided with the appropriate MeSH terms. This may be a flawed strategy. With few exceptions, when the queries were copied into the PubMed interface and the "search details" were examined, the librarians agreed that the key words were correctly mapped to relevant MeSH headings, but they were also mapped to many irrelevant headings. In one example, on mental health screening for children, the tool-generated query's terms for screening prompted PubMed to match it to a MeSH term for cancer screening: "Early Detection of Cancer"[Mesh].

> *"So, it's kind of overmatching to MeSH terms … there's kind of a known issue in PubMed with the specificity of the terms applied by [PubMed's machine learning models for MeSH term matching]."[Participant 2]*

4.3.3c Subtheme 3c. Keywords

Most of the librarians identified keywords that were missing from the tool-generated queries but included in their queries for the topic. In most cases, the queries captured the obvious terms, but missed ones that were perhaps not as immediate. As one librarian noted,

> *"[I had] to sit down and map out all of the various ways in which one can be traumatized. So, there's many more [terms] that I had grabbed… It's missing a lot of those, like domestic abuse and intimate partner violence, sexual crime, that sort of thing."[Participant 7]*

In some cases, the tool-generated search query simply returned permutations of the same term:

> *"And then, here, "acupuncturation". Is that actually a word? Okay. acupuncturists or acupuncturers or acupuncture or acupunctures. I mean, there's a lot of different variations of this word I would not have included. So, I would have excluded these terms."[Participant 8]*

However, for some topics, the tool-generated queries suggested, if not novel terms, terms that could be helpful in the initial design of the search:

> *"I really like that. It found all the different … style of guns. When I first started developing my own search strategy, it sounds silly, but it didn't immediately come to me to think of pistols or revolvers or rifles. [Those terms were] added after looking through other types of searches and other things related to it. … So that's pretty helpful."[Participant 5]*

This suggests that there may be some benefit to searchers in using the query tool. To explore this further, I asked explicitly about how useful the query tool would be to librarians and to other, less experienced, searchers.

4.3.4 Theme 4. Usefulness to searchers

The librarians suggested that the query tool could be useful to trained searchers, but there was a strong concern that inexperienced searchers may not critically appraise the output prior to executing the search. Overall, it seemed that the consensus was the query tool was useful as a "sandbox" or a place to discover terms and create queries to explore in PubMed and other database interfaces:

> *"This is a sandbox right. like this is not a search generator. … AI tools in general are doing is summariz[ing] topics. … This is where, you know, go to play around. [But users need to have] the mindset that [they] need to either talk to a librarian or apply a little critical thinking, too."[Participant 6]*

Librarians described the query tool's usefulness, particularly at the beginning of the search process in identifying relevant terms and possible search structure:

> *"I probably wouldn't just put this in and run it, but putting it in and [thinking] 'oh, I didn't even think about reproductive coercion; that's not a term I used in my search.' … I think it's got a lot of promise at the basic level of just helping you identify terms."[Participant 3]*

Another librarian suggested that due to the tool-generated queries' relatively good sensitivity and in use with features like PubMed's Best Match, these queries could be useful in scoping a topic:

> *"So, we do a lot of scoping on a daily basis. And then also, we have a team of seven librarians. I can see how something like this would help to standardize our process, so that we're all kind of at the same starting point and using the same tool to at least do scoping. I could see that this would actually speed up our scoping quite a bit, … [and] potentially streamline PubMed searches – simplify them in a way. We often over-complicate search strings in an effort to approach a topic from all angles." [Participant 4]*

Nevertheless, one librarian noted that it would not be worthwhile to add it to their workflow, because the added burden of scanning through the identified citations would not be worth the time savings of not having to design a search.

> *"I would not want to work with the searches that it spit out, because they're too broad and irrelevant. I don't have enough time in the day to run through a ton of irrelevant abstracts."[Participant 8]*

Moreover, most librarians were less comfortable with the idea of inexperienced searchers using the query tool. Their concerns were largely centered around the possibility that a novice searcher might take the query tool output as the final product rather than as a starting point:

> *"I think it could probably go both ways, to at least get them started on search term ideas but also potentially … instill in them too much false confidence. Because PubMed will do automatic term, mapping, etc., they [might think] that this is all that they would need, and they don't really need to look it over too much. Would they realize that this is not the end, like they need to have stuff done to it, to play around with it?"[Participant 5]*

Another librarian compared this behavior to that being seen with the increased use of generic large language models, such as ChatGPT, and the potential for the promulgation of substandard research.

> *"When people use [Chat]GPT right? Then they're like, 'Oh, this is perfect. I'll just copy paste this in.' It doesn't matter if there's hallucinations; doesn't matter if stuff has been made up; doesn't matter if things are incorrect. And then they're basing their review on whatever came back. … I've seen this happen a lot with manuscripts where peer reviewers for the journals don't have a clue about how to build a good search strategy. So, they're looking at this and being like, 'I guess, it's the best search strategy. And then it's getting published. … It's kind of a slippery slope you know. I feel like that's the crux of the whole issue: if we build it, will they come? But also, will they then have too much confidence in the tool itself?"[Participant 6]*

The consensus was summarized well by Participant 3, "*At its current state, it's probably best used by people who know its limitations."* Another librarian also summarized this usefulness to experienced searchers particularly well.

> *"I guess there's two different kinds of people, someone who can look at this, and only use it exactly as presented, which isn't bad. It looks like it covers [this] topic pretty well. … But then there are other people who can look at it and spend a little bit more time, and say, 'let me take from this what's useful,' and I think it would be really helpful in that instance."[Participant 1]*

The academic librarians, in particular, expressed that they are overburdened with reference requests for reviews, and a query tool like this might be useful in getting a reviewer started or a student prepared for a reference interview with a trained librarian.

> *"I think this could be good pre-work for folks getting started on a project, especially if they try the search, and maybe they put it in PubMed and say, 'Hey. This is the search that I started with. But I don't understand why I'm getting … all of these trauma results?' We could take a look and modify it from there. I think it's better than starting from*

> *nothing. … I think this would be great for students learning like how searches work."[Participant 7]*

As it appeared that the librarians recognized the potential benefits of integrating tool-generated searches into their workflows, I inquired about how they would like to see this type of tool evolve.

4.3.5 Theme 5. Future directions to improve utility

The feedback on how to improve the query tool's usefulness can be summarized into these three general suggestions: (a) training the query tool on better queries, (b) training the query tool on PICO elements and having the interface split the results by search concept, and (c) using post-processing to better emulate a librarian's role in a reference interview.

4.3.5a Subtheme 5a. Better Training Searches

All the librarians agreed that the queries currently produced by the tool were too broad and generic. A few suggested training the query tool on validated hedges. A validated hedge or filter is a query that has been explicitly designed to capture a concept, such as a publication type (e.g., randomized controlled trials) or clinical topic (e.g., Covid-19).

> *"[You could use] a search hedge, search filter … but I'm also thinking about your organizations that produce clinical practice guidelines. … That would be a good way to get some really high-quality searches on specific topics."[Participant 3]*

A few librarians proposed making the interface bidirectional, so that librarians could submit their final searches to be used to refine the models, or tapping into existing search repository archives as a source of higher quality searches than those published in PROSPERO, but again there was the concern of *"who's peer reviewing the searches that are going to train the model?"[Participant 6]*

4.3.5b Subtheme 5b. Split by PICO (Population, Intervention, Comparator, Outcome) Element

Several librarians suggested that the query tool might perform better if trained on PICO-based keywords, rather than titles and entire key questions. They also suggested that the interface be modular, allowing the user to "*paste in your population, paste in your intervention … because it seems to do better with some concepts than others."[Participant 3]* This would potentially lead to a more modular output, where "*Maybe if it fails to accurately map the question into an overall strategy, you could still kind of work with the pieces*" *[Participant 7]*. One librarian suggested that it would be helpful to have the query tool break up the query by concept, as well as including the full query.

> *"Break it into sort of two or three basic things … [because] it's not doing some of the more advanced techniques like nesting. ... Let's say you did break it out. And you had 4*

boxes show up based on your PICO. Then would there be another section where it put them all together so that someone could copy and paste it into PubMed."[Participant 6]

4.3.5c Subtheme 5c. Interactivity

Finally, several librarians suggested that the query tool could benefit from increased interactivity, which could be achieved in two ways: first, the tool could incorporate basic post-processing features (e.g., automatically suggesting MeSH terms based on keywords) or enabling the automatic addition of validated method filters (e.g., the Cochrane RCT filter).

"I can see [it] being useful to have an off the shelf … "click here to limit to RCT filter.""[Participant 3]

The second form would be to replicate a reference interview, where the query tool could elicit explicit input from the user to better develop the query, probably at least in part by better leveraging the PICO.

"If there is a way to have it [have] almost a brief interactive reference interview of sorts ... and then is able to pull the key concepts out of that."[Participant 6]

Another way this could work would be by eliciting a set of relevant articles:

"So maybe keywords, or … PMIDs, seed-article type things. So, to have it read the seed article, pull relevant terms, and then put them [in the query]. … That's the mental process that I do: I take the title or the question, pull out some of the key things, and then look at seed articles or other reviews to fine tune, adding stuff or removing. So, I think that could be a good option." [Participant 5]

4.4 Discussion

4.4.1 Summary

These semi-structured interviews with eight librarians about a query tool that would generate search queries based on titles and key questions leads to three takeaways.

First, the queries developed are sensitive enough to be a good starting point in the search development process. For the most part, they correctly identified the relevant concepts and produced initial keywords. The comparably large number of citations they return is problematic but may be ameliorated by re-adding quotation marks, standard vocabulary (MeSH terms), field terms, and filters. A few of the librarians tested this and did find that the numbers were reduced, but we did not assess it formally. Additionally, ranking algorithms, such as PubMed's Best Match algorithm,[82] may be useful in prioritizing relevant results.

Second, the tool-generated queries are not an end point in and of themselves. They lack both the necessary sensitivity and precision to be used without scrutiny. Every librarian stated that it would be very

important to educate users, particularly non-librarian users, about this limitation. It is, however, possible that it could be used as a teaching tool or a place for non-expert searchers to start.

Finally, the librarians proposed interesting suggestions for future development of this technology. These included training on better data, allowing for modular searches by concept, and adding interactivity to better mimic the reference interview, as well as algorithmic post-processing to assist in the process of re-applying quotation marks, standard vocabulary (MeSH terms), field terms, and filters.

4.4.2 Limitations

This project was exploratory and therefore relied on librarians with whom I was familiar. This means that the results may not be representative of librarians in different settings or countries or with different foci. Because of the small sample of librarians, I did not explore the ways that different settings (research, academic, and hospital) affected the librarian's reaction to the query tool, but I think this would be of interest. Nevertheless, the librarians were experts in the field with many years of experience, which made them uniquely able to provide thoughtful and informed input that is helpful in determining next steps in the effort to leverage AI in the identification of studies for evidence synthesis. Secondarily, the fact that I did the interviews and analyses alone is likely to have led to both flaws in practice and bias. Although, I consulted with experienced qualitative researchers and incorporated their feedback in the design, execution, and reporting of this study, it would have been better to work with a team with whom I could debrief about the coding and discuss such issues as (1) the decisions regarding the matrix and (2) the translation of codes to matrix elements to themes.

4.5 Conclusions

All the interviewed librarians indicated that they are very busy, with large numbers of requests for search query design help, either from review teams (EPC librarians), students (academic librarians), or clinicians (hospital librarian). All librarians also indicated interest and openness to technology that could improve the search query design process. Although none felt that this query tool was ready to be incorporated into their current workflows, they had useful suggestions for future development based on their expertise in designing search queries and teaching others to do so. Future research should involve incorporating some of the suggestions to improve the tool's usability and performance and continuing to "check in" with librarians and other users as the technology evolves to ensure that tool development remains appropriate for current practice. This has been done very successfully in the integration of natural language processing in abstract screening, in that review teams are beginning to incorporate machine learning to reduce abstract screening workload.

In addition to continuing to incorporate expert medical librarians in evaluating the usefulness of query-design tools, future work should reach out to others in the field, including systematic review methodologists and students, with whom librarians work, to include them in the development process and establish how the tools might be most useful to them. Finally, conversations with users can guide implementation of tools, such as how best to incorporate post-processing steps, and whether a tool could be integrated with a database to enable simultaneous query refinement and abstract screening.

These areas of research are changing and expanding rapidly. It will be critical to continue to incorporate users in the development of tools to ensure that research effort is not wasted and that the tools are best suited to use by their intended audience.

REFERENCES

1. About Cochrane Reviews [Available from: https://www-cochranelibrary-com.revproxy.brown.edu/about/about-cochrane-reviews accessed October 9, 2021.
2. Johnson BT, Hennessy EA. Systematic reviews and meta-analyses in the health sciences: Best practice methods for research syntheses. *Soc Sci Med* 2019;233:237-51. doi: 10.1016/j.socscimed.2019.05.035 [published Online First: 20190528]
3. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Bmj* 2021;372:n71. doi: 10.1136/bmj.n71 [published Online First: 20210329]
4. Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *Bmj* 2017;358:j4008. doi: 10.1136/bmj.j4008 [published Online First: 20170921]
5. Gough D, Oliver S, Thomas J. An introduction to systematic reviews: Sage 2017.
6. Patrini M, Arienti C, Lazzarini SG, et al. Cochrane Rehabilitation collaborates with the World Health Organization to establish a Package of Rehabilitation Interventions based on the best available evidence in stroke. 2020
7. Joorabchi A, Doherty C, Dawson J. 'WP2Cochrane', a tool linking Wikipedia to the Cochrane Library: Results of a bibliometric analysis evaluating article quality and importance. *Health Informatics Journal* 2020;26(3):1881-97.
8. Cumpston MS, McKenzie JE, Welch VA, et al. Strengthening systematic reviews in public health: guidance in the Cochrane Handbook for Systematic Reviews of Interventions. *Journal of Public Health* 2022;44(4):e588-e92.
9. Lefebvre C GJ, Briscoe S, Littlewood A, Marshall C, Metzendorf M-I, Noel-Storr A, Rader T, Shokraneh F, Thomas J, Wieland LS. Chapter 4: Searching for and selecting studie. In: Higgins JPT TJ, Chandler J, Cumpston M, Li T, Page MJ, Welch VA ed. Cochrane Handbook for Systematic Reviews of Interventions The Cochrane Collaboration 2019.
10. Allen IE, Olkin I. Estimating time to conduct a meta-analysis from number of citations retrieved. *Jama* 1999;282(7):634-5. doi: 10.1001/jama.282.7.634 [published Online First: 1999/10/12]
11. 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;4:5. doi: 10.1186/2046-4053-4-5 [published Online First: 2015/01/16]
12. Thomas J. Diffusion of innovation in systematic review methodology: why is study selection not yet assisted by automation. *OA Evidence-Based Medicine* 2013;1(2):1-6.
13. Thomas J, McNaught J, Ananiadou S. Applications of text mining within systematic reviews. *Res Synth Methods* 2011;2(1):1-14. doi: 10.1002/jrsm.27 [published Online First: 2011/03/01]
14. Rethlefsen ML, Farrell AM, Osterhaus Trzasko LC, et al. Librarian co-authors correlated with higher quality reported search strategies in general internal medicine systematic reviews. *J Clin Epidemiol* 2015;68(6):617-26. doi: 10.1016/j.jclinepi.2014.11.025 [published Online First: 2015/03/15]
15. Bullers K, Howard AM, Hanson A, et al. It takes longer than you think: librarian time spent on systematic review tasks. *J Med Libr Assoc* 2018;106(2):198-207. doi: 10.5195/jmla.2018.323 [published Online First: 2018/04/11]
16. Healy HS, Regan M, Deberg J. Examining the Reach and Impact of a Systematic Review Service. *Med Ref Serv Q* 2020;39(2):125-38. doi: 10.1080/02763869.2020.1726150

17. 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 Synth Methods* 2017;8(3):355-65. doi: 10.1002/jrsm.1250 [published Online First: 2017/07/01]
18. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need, Part of Advances in Neural Information Processing Systems, 30, 2017.
19. Devlin J, Chang M-W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:181004805* 2018
20. Wolf T, Debut L, Sanh V, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:191003771* 2019
21. Active learning for biomedical citation screening. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining; 2010. ACM.
22. Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium; 2012. ACM.
23. Carey N, Harte M, Mc Cullagh L. A text-mining tool generated title-abstract screening workload savings: performance evaluation versus single-human screening. *J Clin Epidemiol* 2022 doi: 10.1016/j.jclinepi.2022.05.017 [published Online First: 20220530]
24. Harrison H, Griffin SJ, Kuhn I, et al. Software tools to support title and abstract screening for systematic reviews in healthcare: an evaluation. *BMC Med Res Methodol* 2020;20(1):7. doi: 10.1186/s12874-020-0897-3 [published Online First: 2020/01/15]
25. Callaghan MW, Müller-Hansen F. Statistical stopping criteria for automated screening in systematic reviews. *Syst Rev* 2020;9(1):273. doi: 10.1186/s13643-020-01521-4 [published Online First: 2020/11/30]
26. Large scale empirical evaluation of machine learning for semi-automating citation screening in systematic reviews. 41st Society for Medical Decision Making Annual Meeting; 2019; Portland, OR.
27. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;36(4):1234-40.
28. Johnson EE, O'Keefe H, Sutton A, et al. The Systematic Review Toolbox: keeping up to date with tools to support evidence synthesis. *Syst Rev* 2022;11(1):258. doi: 10.1186/s13643-022-02122-z [published Online First: 20221201]
29. Paynter R AG, Hedden-Gross A, Twose C, Voisin C. Systematic Review Search Strategy Development Tools: A Practical Guide for Expert Searchers. Rockville, MD: Agency for Healthcare Research and Quality, 2024.
30. Bui DD, Jonnalagadda S, Del Fiol G. Automatically finding relevant citations for clinical guideline development. *J Biomed Inform* 2015;57:436-45. doi: 10.1016/j.jbi.2015.09.003 [published Online First: 2015/09/13]
31. 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. ACM.
32. Adam GP, Pappas D, Papageorgiou H, et al. A Novel Tool that Allows Interactive Screening of PubMed Citations Showed Promise for the Semi-Automation of Identification of Biomedical Literature. *J Clin Epidemiol* 2022 doi: 10.1016/j.jclinepi.2022.06.007 [published Online First: 20220620]
33. Wang S, Scells H, Koopman B, et al. Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search? *arXiv preprint arXiv:230203495* 2023
34. Ananiadou S, McNaught J. Text mining for biology and biomedicine: Citeseer 2006.
35. Michelson M, Reuter K. The significant cost of systematic reviews and meta-analyses: A call for greater involvement of machine learning to assess the promise of clinical trials. *Contemp Clin Trials Commun* 2019;16:100443-43. doi: 10.1016/j.conctc.2019.100443
36. Adam GP, Wallace BC, Trikalinos TA. Semi-automated Tools for Systematic Searches. *Methods Mol Biol* 2022;2345:17-40. doi: 10.1007/978-1-0716-1566-9_2 [published Online First: 2021/09/23]

37. O'Connor AM, Tsafnat G, Thomas J, et al. A question of trust: can we build an evidence base to gain trust in systematic review automation technologies? *Syst Rev* 2019;8(1):143. doi: 10.1186/s13643-019-1062-0 [published Online First: 2019/06/20]
38. Manski CF. Daniel McFadden and the econometric analysis of discrete choice. *The Scandinavian Journal of Economics* 2001;103(2):217-29.
39. Traets F, Sanchez DG, Vandebroek M. Generating optimal designs for discrete choice experiments in R: the idefix package. *Journal of Statistical Software* 2020;96:1-41.
40. Rao VR, Rao VR. Choice based conjoint studies: design and analysis. *Applied Conjoint Analysis* 2014:127-83.
41. de Aguiar PF, Bourguignon B, Khots MS, et al. D-optimal designs. *Chemometrics and Intelligent Laboratory Systems* 1995;30(2):199-210. doi: https://doi.org/10.1016/0169-7439(94)00076-X
42. Reid S, Tibshirani R. Regularization paths for conditional logistic regression: the clogitL1 package. *Journal of statistical software* 2014;58(12)
43. CLEF 2017 technologically assisted reviews in empirical medicine overview. CEUR Workshop Proceedings; 2017.
44. CLEF 2018 technologically assisted reviews in empirical medicine overview. CEUR workshop proceedings; 2018.
45. A test collection for evaluating retrieval of studies for inclusion in systematic reviews. Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2017.
46. Wang S, Scells H, Clark J, et al. From Little Things Big Things Grow: A Collection with Seed Studies for Medical Systematic Review Literature Search. *arXiv preprint arXiv:220403096* 2022
47. A Dataset of Systematic Review Updates. Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval; 2019. ACM.
48. Ovid Medline Database Guide Waltham, MA: Wolters Kluwer Health; 2024 [Available from: https://ospguides.ovid.com/OSPguides/medline.htm accessed April 12, 2024.
49. Paynter RA, Featherstone R, Stoeger E, et al. A prospective comparison of evidence synthesis search strategies developed with and without text-mining tools. *J Clin Epidemiol* 2021 doi: 10.1016/j.jclinepi.2021.03.013 [published Online First: 2021/03/24]
50. Zhang A LZ, Li M, Smola AJ. Dive into Deep Learning2020.
51. Lewis M, Liu Y, Goyal N, et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:191013461* 2019
52. Booth A, Clarke M, Dooley G, et al. The nuts and bolts of PROSPERO: an international prospective register of systematic reviews. *Systematic reviews* 2012;1(1):1-9.
53. Schiavo JH. PROSPERO: an international register of systematic review protocols. *Medical reference services quarterly* 2019;38(2):171-80.
54. R Core Team. R: A language and environment for statistical computing. [program]. Vienna, Austria: R Foundation for Statistical Computing, 2017.
55. Adam GP, Di M, Cu-Uvin S, et al. AHRQ Comparative Effectiveness Technical Briefs. Strategies for Improving the Lives of Women Aged 40 and Above Living With HIV/AIDS. Rockville (MD): Agency for Healthcare Research and Quality (US) 2016.
56. Balk E, Adam GP, Kimmel H, et al. AHRQ Comparative Effectiveness Reviews. Nonsurgical Treatments for Urinary Incontinence in Women: A Systematic Review Update. Rockville (MD): Agency for Healthcare Research and Quality (US) 2018.
57. Balk EM, Adam GP, Cao W, et al. Long-term effects on clinical event, mental health, and related outcomes of CPAP for obstructive sleep apnea: a systematic review. *J Clin Sleep Med* 2024 doi: 10.5664/jcsm.11030 [published Online First: 20240201]
58. Balk EM, Adam GP, Cao W, et al. AHRQ Comparative Effectiveness Reviews. Management of Colonic Diverticulitis. Rockville (MD): Agency for Healthcare Research and Quality (US) 2020.

59. Balk EM, Adams GP, Langberg V, et al. Omega-3 Fatty Acids and Cardiovascular Disease: An Updated Systematic Review. *Evid Rep Technol Assess (Full Rep)* 2016(223):1-1252. doi: 10.23970/ahrqepcerta223
60. Balk EM, Ellis AG, Di M, et al. AHRQ Comparative Effectiveness Reviews. Venous Thromboembolism Prophylaxis in Major Orthopedic Surgery: Systematic Review Update. Rockville (MD): Agency for Healthcare Research and Quality (US) 2017.
61. Balk EM, Gazula A, Markozannes G, et al. AHRQ Comparative Effectiveness Reviews. Lower Limb Prostheses: Measurement Instruments, Comparison of Component Effects by Subgroups, and Long-Term Outcomes. Rockville (MD): Agency for Healthcare Research and Quality (US) 2018.
62. Balk EM, Konnyu KJ, Cao W, et al. AHRQ Comparative Effectiveness Reviews. Schedule of Visits and Televisits for Routine Antenatal Care: A Systematic Review. Rockville (MD): Agency for Healthcare Research and Quality (US) 2022.
63. Drucker A, Adam GP, Langberg V, et al. AHRQ Comparative Effectiveness Reviews. Treatments for Basal Cell and Squamous Cell Carcinoma of the Skin. Rockville (MD): Agency for Healthcare Research and Quality (US) 2017.
64. Konnyu KJ, Thoma LM, Bhuma MR, et al. AHRQ Comparative Effectiveness Reviews. Prehabilitation and Rehabilitation for Major Joint Replacement. Rockville (MD): Agency for Healthcare Research and Quality (US) 2021.
65. Panagiotou OA, Markozannes G, Kowalski R, et al. AHRQ Technology Assessments. Short- and Long-Term Outcomes after Bariatric Surgery in the Medicare Population. Rockville (MD): Agency for Healthcare Research and Quality (US) 2018.
66. Saldanha IJ, Adam GP, Kanaan G, et al. AHRQ Comparative Effectiveness Reviews. Postpartum Care up to 1 Year After Pregnancy: A Systematic Review and Meta-Analysis. Rockville (MD): Agency for Healthcare Research and Quality (US) 2023.
67. Saldanha IJ, Cao W, Bhuma MR, et al. Management of primary headaches during pregnancy, postpartum, and breastfeeding: A systematic review. *Headache* 2021;61(1):11-43. doi: 10.1111/head.14041 [published Online First: 20210112]
68. Steele D, Adam GP, Di M, et al. AHRQ Comparative Effectiveness Reviews. Tympanostomy Tubes in Children With Otitis Media. Rockville (MD): Agency for Healthcare Research and Quality (US) 2017.
69. Steele DW, Adam GP, Saldanha IJ, et al. Postpartum Home Blood Pressure Monitoring: A Systematic Review. *Obstet Gynecol* 2023;142(2):285-95. doi: 10.1097/aog.0000000000005270 [published Online First: 20230613]
70. Steele DW, Becker SJ, Danko KJ, et al. AHRQ Comparative Effectiveness Reviews. Interventions for Substance Use Disorders in Adolescents: A Systematic Review. Rockville (MD): Agency for Healthcare Research and Quality (US) 2020.
71. Guirguis-Blake JM, Evans CV, Perdue LA, et al. U.S. Preventive Services Task Force Evidence Syntheses, formerly Systematic Evidence Reviews. Aspirin Use to Prevent Cardiovascular Disease and Colorectal Cancer: An Evidence Update for the US Preventive Services Task Force. Rockville (MD): Agency for Healthcare Research and Quality (US) 2022.
72. Jutkowitz E, Hsiao J, Celedon M, et al. VA Evidence-based Synthesis Program Reports. Accelerated Diagnostic Protocols Using High-sensitivity Troponin Assays to "Rule In" or "Rule Out" Myocardial Infarction in the Emergency Department: A Systematic Review. Washington (DC): Department of Veterans Affairs (US) 2023.
73. 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 [published Online First: 2016/03/24]
74. Woods B, Aguirre E, Spector AE, et al. Cognitive stimulation to improve cognitive functioning in people with dementia. *The Cochrane database of systematic reviews* 2012(2):Cd005562. doi: 10.1002/14651858.CD005562.pub2 [published Online First: 20120215]

75. Bramer WM, de Jonge GB, Rethlefsen ML, et al. A systematic approach to searching: an efficient and complete method to develop literature searches. *J Med Libr Assoc* 2018;106(4):531-41. doi: 10.5195/jmla.2018.283 [published Online First: 2018/10/03]
76. Chung HW, Hou L, Longpre S, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:221011416* 2022
77. Yuan H, Yuan Z, Gan R, et al. BioBART: Pretraining and Evaluation of A Biomedical Generative Language Model. *arXiv preprint arXiv:220403905* 2022
78. Jiang AQ, Sablayrolles A, Mensch A, et al. Mistral 7B. *arXiv preprint arXiv:231006825* 2023
79. Parameter-Efficient Fine-Tuning Library [program]: HuggingFace.
80. Hu EJ, Shen Y, Wallis P, et al. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:210609685* 2021
81. Adafactor: Adaptive learning rates with sublinear memory cost. International Conference on Machine Learning; 2018. PMLR.
82. Fiorini N, Canese K, Starchenko G, et al. Best Match: New relevance search for PubMed. *PLoS Biol* 2018;16(8):e2005343. doi: 10.1371/journal.pbio.2005343 [published Online First: 20180828]
83. 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.1082 [published Online First: 2013/09/01]
84. Qureshi R, Shaughnessy D, Gill KA, et al. Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? *Systematic Reviews* 2023;12(1):72.
85. Anvil [program]. Cambridge, England: The Tuesday Project Ltd., 2024.
86. Elliott JH, Mavergames C, Becker L, et al. The efficient production of high quality evidence reviews is important for the public good. *Bmj* 2013;346:f846. doi: 10.1136/bmj.f846 [published Online First: 2013/02/15]
87. Tsafnat G, Dunn A, Glasziou P, et al. The automation of systematic reviews. *Bmj* 2013;346:f139. doi: 10.1136/bmj.f139 [published Online First: 2013/01/12]
88. Wallace BC, Dahabreh IJ, Schmid CH, et al. Modernizing the systematic review process to inform comparative effectiveness: tools and methods. *J Comp Eff Res* 2013;2(3):273-82. doi: 10.2217/cer.13.17 [published Online First: 2013/11/19]
89. O'Brien BC, Harris IB, Beckman TJ, et al. Standards for Reporting Qualitative Research: A Synthesis of Recommendations. *Academic Medicine* 2014;89(9):1245-51. doi: 10.1097/acm.0000000000000388
90. Gale NK, Heath G, Cameron E, et al. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. *BMC Medical Research Methodology* 2013;13(1):117. doi: 10.1186/1471-2288-13-117

APPENDIXES SPECIFIC AIM 1

Email sent to survey recruits

As a part of my dissertation research, I am seeking your input as a user or developer of systematic review literature identification tools and methods. We hope to define appropriate sets of metrics for the performance of methods of literature identification for systematic reviews (searching and screening).

https://brown.co1.qualtrics.com/jfe/form/SV_9ypzoMjwqt4QqOO

The survey should take about **15 minutes**, and, upon completion, you can choose to submit your email address for a **drawing of a $100 gift card** or to have $100 donated to the charity of your choice. Your email will not be linked to your responses, and all responses will be summarized in aggregate.

Background:

Based on recent systematic reviews of semi-automated methods for literature identification, we identified over 40 metrics that were used in at least one paper to assess the ability of a tool/method to identify relevant studies and exclude irrelevant studies.

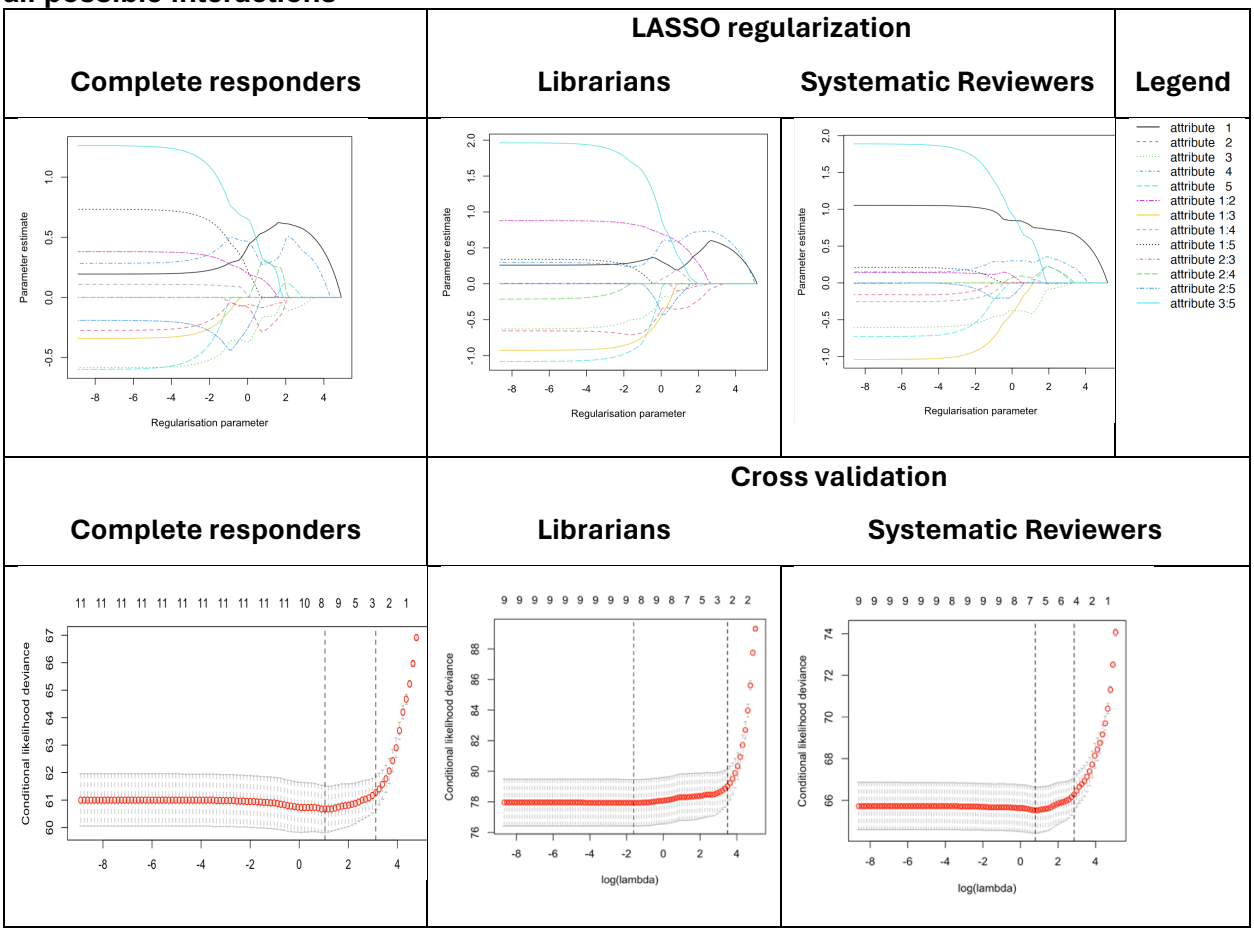
The goal of this project is to establish which metrics are most useful to both tool/method users and developers, and which would be most intuitive to a general audience.

This survey includes 18 of the 40 identified metrics, whose calculations are based on 2 by 2 tables. You will be asked to indicate your preference, as a professional in the field, for different metrics to evaluate the effectiveness of literature identification (search design or screening prioritization) methods. There are no "right" or "wrong" answers; you should simply indicate which metrics you prefer.

https://brown.co1.qualtrics.com/jfe/form/SV_9ypzoMjwqt4QqOO

Thank you very much in advance for your input. If you have any questions, feel free to contact me at gaelen_adam@brown.edu.

Appendix Figure 1. Results of LASSO regression and cross validation on the model with all possible interactions



Legend: Left plot shows lines for the log odds ratio of each attribute and interaction from no regularization penalty for model complexity at left to maximum penalty at right. Several attributes can be seen to converge to 0 before the best model. The right plot shows the cross-validation results, again with the full model at left and the model with 1 variable at right. In this plot the left vertical dashed line shows the regularization parameter value (λ=0.20) for the best model, and the right dashed line shows the regularization parameter value (λ=3.65) for the most parsimonious model within 1 standard error of the best model.

Appendix Table 1. Results of LASSO and cross-validation analysis

	Complete responders		Librarians		Systematic reviewers	
Attribute	**Best model, OR**	**Most parsimonious model within 1 SE of best model, OR**	**Best CV model, OR**	**Most parsimonious model within 1 SE of best model, OR**	**Best CV model, OR**	**Most parsimonious model within 1 SE of best model, OR**
1 (number vs. ratio of relevant records correctly identified)	1.75	1.73	1.38	1.68	2.36	1.99
2 (dependent on prevalence)	0.77	1	0.5	1	1	1
3 (explicitly weights	0.85	0.96	0.62	1	0.66	0.97

	Complete responders		Librarians		Systematic reviewers	
Attribute	**Best model, OR**	**Most parsimonious model within 1 SE of best model, OR**	**Best CV model, OR**	**Most parsimonious model within 1 SE of best model, OR**	**Best CV model, OR**	**Most parsimonious model within 1 SE of best model, OR**
sensitivity and workload)						
4 (increases linearly with workload)	1.33	1.43	1.29	1.87	1.34	1.29
5 (changes slowly under class imbalance)	1	1	0.43	1	0.95	1.07
Interaction of 1 and 2	1.12	1	2.23	1	1.15	1
Interaction of 1 and 3	1	1	0.47	1	0.60	1
Interaction of 1 and 4	1	1	1	1	1	1
Interaction of 1 and 5	1	1	1.16	1	1	1
Interaction of 2 and 3	0.93	1	1	1	1	1
Interaction of 2 and 4	1.33	1	1	1	1.02	1.09
Interaction of 2 and 5	1	1	1	1	0.81	1
Interaction of 3 and 5	1.35	1	4.89	1	2.95	1

CV = cross validation; OR = odds ratio; SE = standard error

APPENDIXES SPECIFIC AIM 3

Interview guide

Questions

1. Which of the three searches did you choose to evaluate?
    a. Title only
    b. KQ only
    c. Title + KQ
2. When you ran the search in PubMed
    a. Did it trigger an error? Y/N If Y, which
    b. How many citations were returned
    c. When compared with the seed citations,
        i. What proportion of the seed citations did it capture?
        ii. Were there any surprises/comments (e.g., did it do better/worse than expected)?
        iii. If you added appropriate search filters, would the number be reasonable for your team to screen?
3. Looking more closely at the search terms
    a. Translation:
        i. Are all concepts covered? Y/N If N, what concepts are missing?
    b. Boolean and Proximity operators:
        i. Could you easily match the terms with correct MeSH terms (direct them to look at advanced search)?
        ii. Would you change any of the simple Booleans to proximity operators would you turn any into phrases?
    c. Text words:
        i. What text words would you add?
        ii. What would you remove?
    d. Structure:
        i. Would you change the structure? Y/N If Y, in what way
4. Based on your draft/final search for this topic for comparison
    a. Would you change your search in any way based on this output?
5. Do you think this tool would be helpful or harmful for non-expert searchers in either beginning to think about a search or preparing to meet with a medical librarian?
    a. Would you use it as a starting point (why/why not)?
6. How can we improve this tool to make it more useful to you?

Email sent to survey recruits

For the final piece of my dissertation research, I am looking for recruits to spend an hour on Zoom with me sometime in the next 6 weeks, evaluating an AI model trained to create search strategies based on titles and key questions.

This would be in your professional capacity (i.e., I will not be collecting any personal information), and I would request that you have a search for which you have a final (or nearly final) PubMed search strategy and a set of seed citations that you would test the search strategy against.

If you are interested, please email me at gaelen_adam@brown.edu.