AHRQ Agency for Healthcare Research and Quality

Evidence-based Practice Center Methods Report Protocol

Project Title: ***Tools to (Semi)Automate Evidence Synthesis***

I. Background and Purpose of the Review

Background

Despite the large amount of research into methods to automate or semi-automate labor- and time-intensive steps in evidence synthesis methodology, the uptake of automation in the conduct of evidence synthesis products (ESPs) has been low.[1] Evidence-based Practice Centers (EPCs) are just beginning to use these tools, but the EPCs that do use them use different tools, for different tasks, and use them in different ways. Nevertheless, there is a great deal of interest among EPCs in the potential for tools that use artificial intelligence (AI) or machine learning (ML) to automate or semi-automate steps in the evidence synthesis process, thereby reducing the time and expense without compromising rigor.

As the EPC Program enters this field, it is crucial to understand what research has been done and what tools already exist to avoid redundant effort and to help EPCs determine which tools may improve their processes. In addition, this is a very active field, so a single review will become outdated very fast. Thus, a living review is warranted.

Purpose of the Review

This living review will summarize the evidence on the use of ML and AI tools in the conduct of any specific aspects of ESPs commonly produced by EPCs (e.g., abstract screening, data extraction, summary writing). The intended audience includes evidence synthesis practitioners, tool developers, and evidence synthesis methods developers. For the purpose of this review, we define a tool as a system, method, instrument, model, application, software, or package that uses AI or ML to semi-automate or fully automate any specific aspect(s) of ESP production. Semi-automation refers to automation of some of the work while relying on humans to complete certain tasks (e.g., abstract screening tools that update models iteratively but still require human screening). Full automation refers to the ability of the computer to do all of the work (e.g., abstract screening that is pretrained and then labels all new citations without further human involvement).

II. Key Questions and Eligibility Criteria

Key Question (KQs)

KQ 1: What tools that use artificial intelligence (AI) or machine learning (ML) to semi- or fully automate any aspect of evidence synthesis product (ESP) production have had their performance evaluated?

> **KQ 1a**: How do these tools perform?

Study Eligibility Criteria

We have amended the standard Population, Intervention, Comparators, Outcomes, Timing, Setting ("PICOTS") framework to better fit this topic. Specific eligibility criteria include:

Domain

- Any aspect of ESP production, including
    - Question development/topic scoping
    - Development of literature searches
    - Title/abstract screening
    - Full-text screening
    - Data extraction
    - Data summarization
    - Data synthesis
    - Risk of bias/quality assessment
        - Based on study design (e.g., randomized, observational, diagnostic test accuracy)
    - Certainty/strength of evidence/strength of evidence (e.g., per GRADE)
    - Writing of reports, plain language summaries, manuscripts, etc.
    - Other
- <u>Limit</u> to the biomedical domain (i.e., topic areas that fall within the general remit of the EPC Program)
- <u>Exclude</u> tools that are not designed to automate any part of a standard systematic review process (e.g., tools to create concept maps, tools to monitor the literature)

Tools

- Tools are systems, methods, instruments, models, applications, software, and packages that use AI or ML to semi-automate or fully automate any aspect of ESP production.
- <u>Include</u> any publicly available tool that can be accessed via the Internet (for download or direct use online) and used without having to contact the publication/study author.
    - There is ***<u>no</u>*** requirement that the tool be free. The tool may have a charge and may be accessible only through a paywall or through a username/password limited site, etc.
- <u>Include</u> free-standing tools or those embedded within (or used via) coding languages (or other software (e.g., R packages, open-source tools, the abstract screening module of SRDRplus)
    - <u>Exclude</u> Evaluations of models that have not been incorporated into a tool.
- Must be intended for ESP production.

Comparators

- Human doing the described ESP task.
- Any other tool that meets the Tools inclusion criteria.
- <u>Include</u> studies of any tool evaluated across multiple datasets (i.e., real or simulated search results for various topics/projects), with or without a comparator.

Outcomes

- Quantitative performance, specifically the utility of using versus not using the tool in a ESP. For example:
    - Literature screening:
        - The number of finally included papers correctly identified (e.g., sensitivity/recall/yield)
        - The reduction in screening burden (e.g., precision, specificity, work saved over sampling, gain, f-score, number needed to read),
        - Time
    - Data extraction and risk of bias assessment:

        - The accuracy of extraction tools in comparison to human data extraction (e.g., frequency of an exact match between the human-annotated and tool output or annotation, ability of the tool to identify sentences carrying relevant information on a particular element),
        - Time
    - Strength/certainty of evidence:
        - Accuracy of citation, annotation, etc.
        - Accuracy compared to human assessment
    - Data summarization and other text generation tasks:
        - Quality in comparison with human-written summaries (e.g., readability, comprehensiveness, accuracy, relevancy, etc.),
        - Time
- Exclude: qualitative (narrative only) assessments (e.g., "We tried Tool X, and the tool was successfully incorporated into our workflow.")
- Exclude: Usability analyses

Study Designs

- Primary studies
- Systematic reviews, to be used as a source of citations
- Exclude: narrative (non-quantitative) summaries, usability studies
- Exclude: Analyses that include only a single review, example article, literature search, screened project, topic, etc.
- Exclude: Conference abstracts and preprints older than 2 years (from January 2022 for the initial review)

Timing

- Unlimited
    - Reports before 2021 will be identified through the reference lists of existing SRs
    - Reports from 1/1/2021 onward will be identified through a combination of database literature searches and screening the reference lists of evidence synthesis products
    - Updates will take place every 6 months and will overlap the last search timeframe by 1 year (update searches will be deduplicated with existing corpus prior to updated screening)

III. Methods

The review will follow AHRQ EPC Program methodology, as described in its Methods Guide,[2] with specific processes described in the EPC guidance for rapid reviews.[3]

Literature Identification

To keep the scope of this project feasible given resource limitations, we propose to use the studies in Khalil 2022[4] and Jimenez 2023[5] as the source of studies up to 2021. We will search PubMed, Embase , and the ACM Digital Library from January 1, 2021 to present, with search updates every 6 months. We will also search the Cochrane Library for methodology reviews. We will hand search two journals that are not indexed in either bibliographic database: Journal of the European Association for Health Information and Libraries (JEAHIL) and Cochrane Evidence Synthesis and Methods. To identify conference papers and preprints, primarily in the computer science literature, we plan to search BioarXiv and medRxiv through Embase and ACL Anthology. Search terms will include terms for ESPs (systematic review, scoping review, evidence map, etc.) and evidence synthesis product methodology (search, screening, data extraction, etc.); terms for AI and ML, including terms for specific known tools; and terms for evaluation

(evaluation, workload, sensitivity, etc.) The studies will be limited to those available in English. Date limits will be implemented as described in the eligibility criteria. The search strategies for all databases will be developed by two experienced evidence synthesis librarians, with one doing primary development and the second reviewing based on the Peer Review of Electronic Search Strategies (PRESS) checklist.[6] We will revise the searches for updates if gaps are identified. The current full Medline and Embase search strategies are in Appendix A.

We will take advantage of the abstract screening ML capacities of SRDR+ (https://srdrplus.ahrq.gov/) to limit resources spent on abstract screening. We will train the ML algorithm as follows: (1) We will review the reference lists from known existing evidence synthesis products and clinical practice guidelines to identify potentially relevant studies (approximately 140 citations), which will be entered into SRDR+ and screened by all team members, with resolution of all conflicts in conference. (2) Subsequently, other known citations and the 500 "most relevant" citations per PubMed will be added to the corpus, and citation screening will proceed in duplicate. Upon completion of this set of citations, all citations found by the full literature searches will be added to the already-screened citations in SRDR+, and abstract screening will continue in duplicate, with conflicts adjudicated in conference or by a third screener. (3) As screening progresses, the pretrained SRDR+ machine learning algorithm will continue to adapt and will sort the list of unscreened abstracts such that the most potentially relevant articles are presented first. This process will make screening more efficient and will enable us to capture the preponderance of relevant articles relatively early in the abstract screening process. (4) We will stop double screening when we have rejected at least 200 consecutive citations. (5) Each 6 months, new citations will be added to the corpus and we will resume screening until we again reject at least 200 consecutive citations.

Potentially relevant citations will be retrieved in full text and screened during extraction by a single screener with verification by a second methodologist.

Data Extraction and Data Management
Data from eligible studies will be extracted into the SRDR+ database. The unit of analysis will be the intersection of tool and task, so if a tool offers (semi-)automation for multiple tasks, it will be assessed separately for each task.

For each study, we will extract

- Basic information about the tool and the automation method
    - Tool URL
    - What's the technology
        - Model types (e.g., SVM, deep learning, LLM), as reported in the studies
        - Version and extension number; For AI, specific name and version (e.g., ChatGPT 4, Claude 2)
        - Developer
        - Environment (e.g., web-based, Python, R)
- The ESP task(s) automated
- The number and clinical topic(s) of ESPs tested
- Paywall (y/n/levels/free options)
- Reported results for the utility of using versus not using the tool in a ESP.
    - Date tool assessed
    - Inputs (i.e., the instructions given to the AI, how model trained)
    - Outputs (e.g., description of the generated content, ranking, labels)
    - Reported results

Data Synthesis and Presentation
We will iteratively summarize the evidence into an interactive evidence map. Full extraction data for each study will be presented in a series of evidence tables. We will not attempt to summarize data statistically across studies, assess the quality of included studies, or assess the strength of evidence for any tool or set

of tools. A white paper will include a narrative review of all studies found in the initial iteration of the project. Upon publication of the white paper and with each subsequent update, we will make the SRDR+ database available for public viewing and downloading.

IV. References

1. 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 Jun 18;8(1):143. doi: 10.1186/s13643-019-1062-0. PMID: 31215463.
2. AHRQ. AHRQ Methods for Effective Health Care. Rockville (MD): Agency for Healthcare Research and Quality (US); 2017. https://effectivehealthcare.ahrq.gov/products/collections/cer-methods-guide. Accessed on August 15 2023.
3. Hartling L, Guise JM, Kato E, et al. AHRQ Comparative Effectiveness Reviews. EPC Methods: An Exploration of Methods and Context for the Production of Rapid Reviews. Rockville (MD): Agency for Healthcare Research and Quality (US); 2015.
4. Khalil H, Ameen D, Zarnegar A. Tools to support the automation of systematic reviews: a scoping review. J Clin Epidemiol. 2022 Apr;144:22-42. doi: 10.1016/j.jclinepi.2021.12.005. PMID: 34896236.
5. Cierco Jimenez R, Lee T, Rosillo N, et al. Machine learning computational tools to assist the performance of systematic reviews: A mapping review. BMC Med Res Methodol. 2022 Dec 16;22(1):322. doi: 10.1186/s12874-022-01805-4. PMID: 36522637.
6. McGowan J, Sampson M, Salzwedel DM, et al. PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement. J Clin Epidemiol. 2016 Jul;75:40-6. doi: 10.1016/j.jclinepi.2016.01.021. PMID: 27005575.

V. Definition of Terms and Abbreviations

AI	Artificial intelligence
AHRQ	Agency for Healthcare Research and Quality
EPC	Evidence-based Practice Center
GQ	Guiding Question
MeSH	Medical Subject Heading
ML	Machine learning
ESP	Evidence synthesis product
SRDR+	Systematic Review Data Repository-Plus

VI. Summary of Protocol Amendments

This protocol has been reviewed by two members of the EPC Program with a particular interest in the use of ML and AI in systematic reviews. It will be posted publicly on the AHRQ Website. If we need to amend this protocol, we will give the date of each amendment, describe each change and give the rationale in this section.

VII. EPC Team Disclosures

EPC core team members must disclose any financial conflicts of interest greater than $1,000 and any other relevant business or professional conflicts of interest. Related financial conflicts of interest that cumulatively total greater than $1,000 will usually disqualify EPC core team investigators from participation in the review.

VIII. Role of the Funder

This work was funded by the Agency for Healthcare Research and Quality through the following contracts: 75Q80120D00001, 75Q80120D00004, 75Q80120D00006, and 75Q80122C00002. The findings and conclusions in this document are those of the authors, who are responsible for its contents;

the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

IX. Registration

This protocol will be posted on the AHRQ website.