Penn Medicine
Center for Evidence-based Practice

CENTER for EVIDENCE-BASED PRACTICE

Use of artificial intelligence tools in evidence reports

January 2024

As the annual volume of published clinical studies in medicine, nursing, and related subjects continues to grow, evidence-based practice centers have sought to automate the processes of finding and analyzing evidence, to make the best use of limited resources. Interest in the use of computer algorithms and machine learning in systematic reviews predates the advent of artificial intelligence tools like ChatGPT. Applying these technologies to evidence-based practice is also an area of considerable research interest, and CEP has lent its support to various research studies at Penn and at other institutions.

Machine learning and literature searching

The challenge in applying artificial intelligence to the science of systematic review and evidence synthesis is that any such tool ultimately reflects the bias of its developers. In the case of machine learning systems, of which Google is a prime example, their programming creates a bias towards conventional wisdom, preferentially showing search results that previous users have clicked on. While this improves the specificity of search results, it worsens sensitivity and is a potential source of bias, as less-favored results sink lower and lower in algorithmic ranking of perceived relevance. The bias reinforces itself, as users are more likely to click on results shown at the top of the page, and other references are ignored.

Furthermore, the algorithms may not be so reliable in identifying and ranking references where differences in terminology, interventions, or outcomes reported, or even spelling errors make one reference different from the others. And because the algorithms are complicated, proprietary, and constantly being updated, search results are not reproducible, foreclosing one of the most fundamental elements of the scientific method: reproducing previous studies to verify their findings.

Finally, literature searching algorithms are more likely to give lower ranking to smaller studies and studies demonstrating little or no effect of the intervention in question. This has the same effect on search results as publication bias at the stages of writing and peer-reviewing systematic reviews.

For purposes of gaining a basic understanding of a topic, finding seed papers, and finding terms to be included in a structured search, Google, the "Find Similar" feature of Ovid MEDLINE, search term mining, and other algorithm-guided products are valuable tools. These may be used in search development and do not need to be disclosed, but they may not be part of the finished search strategy. With the approval of the Director, they may be used *a priori* as a supplement to but not a replacement for a structured search using index terms and keywords in a comprehensive bibliographic database. However, these algorithm-guided tools are not a substitute for the knowledge and experience of a CEP analyst, as the analyst, with the benefit of experience and the information provided by clinical partners, will likely know of terms that would be overlooked by a computer program.

Reference ranking algorithms such as the Embase "Sort by Relevance" feature rely on simple algorithms and may not have the same machine learning components as Google, but lack reproducibility and will miss or downplay references with misspellings or using alternate descriptions of an intervention. They too should be avoided in final systematic review searches but have a role in search development. For example, it is appropriate to use a ranking algorithm when testing a NOT term being considered for inclusion in a search. By prioritizing the most potentially relevant hits, the algorithm makes it easier to find references that would be missed by the resulting search, ultimately improving the sensitivity of the search.

Deduplication, reference screening, and article retrieval

CEP routinely uses automated reference deduplication tools built in to reference management programs like Covidence and RefWorks. They are based on documented algorithms that compare database fields like author, title, and pagination to identify sets of references that are likely to be duplicates. These tools may be used in preparation of CEP reports, but an analyst must review the machine-generated list of potential duplicates and mark any references that were mistakenly listed as duplicates. Computer-assisted deduplication and article retrieval are routine CEP methods and do not need to be specifically disclosed in the methods section of a report because they do not pose a risk of overlooking relevant evidence.

The algorithms used for ranking search results can also be used as an aid to screen results. While the algorithms will not be as effective as an experienced analyst, they can serve as a supplement to manual screening of search results, calling out possibly relevant references for re-review much as a pattern-recognition algorithm can act as a "second opinion" in the detection of possible cancers in an x-ray image. Since CEP standard practice is for each reference in a set of search results to be screened by a single research analyst, and we frequently do not perform duplicate screening, this kind of secondary screening will increase the sensitivity of screening and reduce the risk of missing relevant evidence, as long as the ranking algorithm does not completely replace the analyst's screening.

Validated search filters such as the Cochrane RCT filter are being added to reference management programs. If the filters are being used to exclude studies, it is preferable to apply them at the search stage rather than the screening stage. They may be used as an aid to screening in the same way that search ranking algorithms may be used in screening.

Some software packages also include features to automatically retrieve full text open-access articles from digital object identifiers incorporated in database records, extract a list of references from a published article, or to automatically search for articles citing or cited by a specified paper. These tools simply automate repetitive tasks and do not introduce a risk of missing relevant evidence or misinterpreting evidence. They may be used routinely and do not need to be disclosed.

Machine translation of foreign-language articles

The reliability and efficiency of machine translation tools such as Google Translate has progressed to the point where they be used in routine work abstracting data from published articles. When using evidence from foreign-language articles, the CEP analyst must sufficiently translate the article to be able to understand the population, methods, and results, and adequately assess threats to the validity of the results and conclusions.

Quality appraisal and data abstraction

Another area where artificial intelligence tools have been used in systematic review is the appraisal of study quality and the abstraction of data. For example, computer algorithms can detect text in a published article that may describe how a study was performed, and potentially judge whether a study meets specified quality criteria such as randomization or blinding. The problem with this approach is that study authors can intentionally or unintentionally present their methods and results in a way that makes the study look more reliable than it truly is. Judgement is required to assess quality appraisal because the presentation of study methods and results is much less standardized in content and structure. The possibilities for such bias are too numerous for a computer program to be able to recognize all of them. The proponents of "living" systematic reviews and guidelines have called for increased standardization of research outcomes and reporting, and the storage of study data in central repositories, but until such processes are universally followed, artificial intelligence will not be able to replace a human research analyst.

Likewise, study authors may frequently emphasize results that favor their hypothesis and conclusions, and downplay or avoid reporting results that contradict them. This is why CEP avoids including evidence from studies published only in abstract form: there is not room in an abstract to report all methods and results, so authors can select the most favorable results to report. The mechanisms by which authors can skew their presentation of results are endless, and no AI algorithm has yet been able to detect all of these forms of bias and abstract complete and unbiased results from a published article. For this reason, it is unlikely that AI will ever be able to replace a trained research analyst in abstracting study results.

CEP Policy

Primary searches, including those in MEDLINE, Embase, CINAHL, JBI databases, and the Cochrane Library must use reproducible search terms. All hits from these databases should be screened by one or more of the analysts working on the report. Use of artificial intelligence tools to highlight terms that may bear on a screening decision or as means of obtaining a second screening of each reference is acceptable. Use of reference ranking algorithms may be necessary in screening results from other databases, particularly those that lack structured multi-line search capabilities. If such algorithms are used, search results should be footnoted accordingly, and the number of references actually screened should be reported along with the total number of references in the set.

Artificial intelligence tools may be used to find and highlight terms that may bear on study quality, but in all cases, the actual quality appraisal must be carried out by a CEP analyst. Artificial intelligence tools should not be used in data abstraction.

Any use of artificial intelligence tools or reference ranking algorithms in final searches, reference screening, or quality appraisal must be disclosed in the report. Use of AI or algorithms in the development stages of a search does not need to be disclosed as long as the finished search is documented and fully reproducible. Disclosures should include the tool or algorithm used, with date or version number if available, the purpose of use, and if the tool was used for ranking search results, how many references were reviewed by a CEP analyst and how many were ruled out by the algorithm.

Initial version: September 2023
Updated: February 2024