Automated tools for evidence synthesis: Briefing note

October 2023

Author: Richard Ward

The Strategy Unit's Evidence and Knowledge team use rapid and pragmatic methods to complete reviews within weeks, to prompt thinking and inform decision-making. An example of such a pragmatically focused review may be seen here (Aldridge, 2021), and a wider blogpost outlines why pragmatism is such an important part of our approach (Turner, 2023).

Maintaining this pragmatic approach can be challenging. As the information we seek to assimilate and distil continues to grow in size and complexity, with "*an excess of 4000 health research articles being published daily*" (Khalil et al., 2022). So, we have asked ourselves: how can we create greater efficiencies in time and how our efforts are spent, while maintaining quality and rigour in our work?

Technology-based tools may provide practical solutions. Advances in artificial intelligence (AI), computing and linguistics have seen the development of tools which can process and analyse large volumes of textual data, rapidly and 'automatically'. AI-based tools are being increasingly applied to expedite stages of the evidence synthesis (ES) cycle (Cierco Jimenez et al., 2022; Marshall & Wallace, 2019). However, existing research on this topic focuses more on systematic reviews in an academic context. And we've seen limited research, guidance, or formal evaluation to inform AI-based tool adoption in a more practice-oriented context.

Figure 1. Core concepts underpinning automated tools for evidence synthesis (IBM, 2023a, 2023b; Marshall & Wallace, 2019).

Artificial Intelligence (AI) broadly refers to a field of research whose goal is to replicate human intelligence and perception for the purpose of completing specific complex tasks. AI also describes a set of algorithmic technologies which can been applied or 'trained' to achieve the automated or semi-automated completion of complex repetitive tasks.

Machine learning (ML) refers to a technology or application of AI whereby computers are iteratively trained to perform specific tasks using statistical modelling of large amounts of data. The programme tries to automatically learn from experiences over time by making inferences about the data, without the need for fixed rule-based programming for each command or task.

Natural Language Processing (NLP) refers to the application of computational methods and technologies to simulate human ability to understand and manipulate spoken and written language. This may be applied to automatically process and analyse unstructured human language texts, such as journal articles.

Text mining (TM) describes an application of AI which uses NLP (amongst other technologies) to process and transform unstructured text within a set of documents into structured text which may be processed and classified by ML algorithms. TM describes the use of this process for the purpose of automatic information extraction and classification, identifying textual patterns and trends within the data.

In this briefing note we share the beginnings of our journey, as a public-sector organisation, exploring innovative approaches to ES using AI-based tools. We outline some of the core concepts which underpin these tools (see above - figure 1), while sharing some reflections on recent primary and secondary research, including:

- Examples of open-source tools which might be applicable to the way we work; and
- Reflections on how well benefits and drawbacks might translate to more rapid or pragmatic contexts and what this might mean for us as an organisation.

So, what are some examples of automated evidence-synthesis tools?

We found a daunting range of over 50 available tools that were identified by scoping and mapping reviews. Over half in each review were devised for searching and screening (Cierco Jimenez et al., 2022- 63 tools; Khalil et al., 2022- 55 tools). Aside from data extraction, these can be some of the time consuming and information intensive review tasks. While most readers will be familiar with these, a brief overview is given in figure 2.

Figure 2. Outline of searching and screening tasks.

Searching	Screening
• Devise key terms and subject headings. • Combine terms using Boolean logic. • Execute across different electronic databases. • Explore grey literature. • Export records and de-duplicate. • Retrieval of full-text documents of relevant records.	• Devise inclusion and exclusion criteria. • Assess relevance of records against inclusion criteria iteratively by title, abstract and full text. • Document reasons for exclusion and negotiate conflicts. • Compile body of relevant articles marked for data extraction and synthesis.

Selecting seven tools from recent reviews (see table 1 for further detail), we found that their potential applications across the continuum of searching and screening tasks was diverse.

Three of the tools from Bond University's Systematic Review Accelerator project (Clark et al., 2020) have potential as exploratory tools, or adjuncts to our typical methods in search strategy development. Though, researchers note an important a trade-off in lower sensitivity versus time saved in search strategy development (Paynter et al., 2021). While lower sensitivity might be an untenable trade-off for more systematic methods, such tools may have greater applicability for more pragmatic reviews. These tools could also reduce the level of administrative work in translating search strings across databases[1], as well as time taken in testing retrieval of initial searches. Time savings achieved through these tools may enables us to be more responsive to client need, or ad hoc requests in the initial planning stages of reviews.

Search evaluation tools which test sensitivity and precision may have less of a role for most of our work, typically focused on rapid-turnaround reviews. Though, *SearchRefinery* could be useful when we have a starting point of key research papers. PubMed IDs from key papers could be used to evaluate retrieval of potential search strings. This could be used alongside tools which analyse word frequency to test how well search terms retrieve relevant sources. However, the tool relies on *PubMedIDs* to test retrieval, and so it may be less useful for reviews focused on social care or social sciences.

---

[1] Previously, Health Databases Advanced Search (HDAS) enabled us to conduct a single database search and translate this to the native database syntax with relative ease and efficiency. Closure of this service in 2022 (National Institute for Health and Care Excellence, 2022) led to us having to manually translate searches within each native database. While we search across fewer databases than in systematic reviews, translating syntax and search field terms across databases is both time-consuming and cognitively demanding. Particularly for the kinds of complex searches we tend to undertake. In this sense, we can see utility in adopting Polyglot Search Translator as an adjunct tool for reducing the burden of this stage.

Most excitingly, we identified two grey literature searching tools. Devised as open-source tools as part of the *Evidence Synthesis Hackathon* (Evidence Synthesis Hackathon, undated), we'll be taking a close look at these to understand how they might reduce the time to conduct and compile grey literature search results.

Despite the high potential value of these tools to our work, it is worth noting that many tools exist as proof of concept or prototypes. Trying a few rapid tests of the *Shiny* interface for *Greylitsearcher* (Haddaway, 2022), and the chrome browser extension for Grey Literature Reporter (Webpage: Penkin & Haddaway, 2022; Chrome web store link), we hit some barriers, including error messages. This dovetails with experiences of others including Marshall and Wallace (2019) and Khalil et al. (2022). With this research noting that if there is to be more widespread adoption and usage among evidence and knowledge professionals, the field of research must move beyond available prototypes to professionally maintained platforms which enable seamless use.

Screening tools such as *Abstrackr* (Wallace et al., 2011) may introduce efficiencies in some of our larger projects, where we have a large body of citations which necessitates multiple screeners. One feature caught our interest the most. The tool periodically uses relevance ranking data to re-order citations so that the most relevant sources appear first. This could be invaluable in navigating complex research topics, enabling us to build a coherent mental map of the unfamiliar territory more rapidly.

We anticipate that any reductions in screening workload and time spent may be more modest than for systematic reviews. Due to a lower burden of articles to screen, earlier stopping criteria and fewer people spending their time screening. Before adoption, we would also want to know more about how the tool performs and the scale of any potential drawbacks, given that the European Centre for Disease Prevention and Control (ECDC, 2022) highlights several studies which indicate potential performance concerns. A published evaluation of *Abstrackr* found the proportion of relevant sources that were excluded ranged from 0 to 22 per cent (Gates et al., 2020). The upper range of this figure represents a high error rate. As such, we need to be cautious in our adoption and testing, given that the authors were unable to assess how a high error rate might impact the weight of review findings. Ideally, to drive greater adoption of tools in rapid review contexts, more granular information is needed on how error rates vary, how they impact review findings, and ways to mitigate this impact.

So, what might these tools mean for our work?

While this is a promising area of development, AI-based evidence synthesis tools should not be considered a 'panacea' or 'cure all' for the pressures imposed by an ever-expanding evidence base. Given potential trade-offs, and a lack of information on unintended consequences, it is important that tools aren't applied uncritically to resolve workload pressures. As ECDC (2022) have highlighted in mixed-methods research with public health professionals on the use and impact of automated ES tools: adoption and deployment of tools within the evidence synthesis workflow requires considerable time and effort.

Realising any potential time or workload efficiencies will be contingent on having sufficient time and space to experiment, to build familiarity, confidence and skills (ECDC, 2022). And, at this stage of development, efficient tool adoption and optimisation within individual organisations may require domain knowledge in computer science, programming and information science. For example, some of the tools we have identified are packages or add-ons to the programming language and software environment *'R'*. While this comes with a higher barrier to entry for the novice or programming-naïve adopter, such tools have the benefit of being open source. Where shared code repositories on *GitHub* enable curious researchers to 'look under the hood', this may enable greater researcher collaboration and further tool refinement.

Table 1. Tabulation of seven automated evidence synthesis tools according to area of use, features, benefits and drawbacks, and links to further information.

Tool (Area of use)	Features/ function	Reported benefits / drawbacks	Links to further information
Word Frequency Analyser (Search)	Web-based app which tabulates most frequent words in title, abstract and keyword fields. Researchers must upload a reference library (XML, RIS or BibTex format). Settings may be modified to alter reporting parameters.	Clark et al. (2020a) used WFA (amongst a suite of other tools) to complete a systematic review in two weeks. Anecdotally, WFA may save time when used inductively to devise potential search terms based on their frequency within a corpus of seminal articles.	Guidance for use is available here.
SearchRefinery (Search)	Web-based interface for formulating and visualising Boolean queries within PubMed. The tool tests search strings based on whether they retrieve key pre-identified citations, displaying the number of relevant records retrieved and those yet to be retrieved.	Scells and Zuccon (2018) report that the tool enables experts to better understand query retrieval by providing a visual interface. The tool may also lessen the time taken to refine and test query retrieval.	Guidance for use is available here. More detailed information concerning the potential uses of the tool and its impact may accessed via a conference paper by Scells and Zuccon (2018).
Polyglot Search Translator (PST) (Search)	Web-based interface translates PubMed/ MEDLINE search strings to the correct syntax for other native database platforms. Does not translate subject heading syntax, which varies across databases.	RCT evidence by Clark et al. (2020b) shows modest efficiencies in time saved (31 vs 45 min) and lower mean number of translation errors (8 vs. 16) for this tool compared to manual translation.	Guidance for use is available here.
Abstrackr (Screen)	Web-based platform for collaborative citation screening involving multiple reviewers. The tool can process multiple forms of citation data (PubMedIDs; RIS; XML). Reviewers assign relevance labels to entire abstracts or specific terms which pertain to likelihood of relevance. The system uses this to rank each article according to relevance, iteratively re-ordering the citations in the reviewer's list so that the most relevant appear first.	ECDC (2022) note variation in tool performance across some studies report low false negative rates, others report some tendency towards overinclusion and performance limitations. Some report more favourable performance for mixed-methods and qualitative studies. Gates et al. (2020) compares the tool across different review	Guidance for use is available here. For more detailed information on the development and deployment of Abstrackr, please refer to Wallace et al. (2011).

		methodologies reporting on time saved and the proportion of relevant articles missed.	
LitSearchr (Search)	R package which quasi-automates search strategy development using text-mining and analysing networks of keywords. Users devise a "naïve" search strategy to capture relevant results. By importing the results as .ris or .txt file, different functions support: keyword identification, building a network of where key words occur, identifying where key words may be removed or excluded, and grouping terms into concepts. The tool uses these prior stages to build Boolean search strings across different languages and database syntax, with the capability to assess precision and recall. Web-based graphical user interface available in beta form - though this form excludes several more extensive functions.	Time/ manual workload savings for systematic reviews (<2 hours, compared to 17-34 manual hours)- though this was for environmental science reviews (Grames et al., 2019). The method may identify articles missed by conventional methods.	Further information is available here. Detailed guidance for use is available via the following vignette (Grames & Hennessy, 2020). Published research (Grames et al., 2019) has assessed performance relative to manual systematic review methods.
Grey Literature Reporter and Greylitsearcher (Search)	Two tools developed as part of the Evidence Synthesis Hackathon to increase grey literature search efficiency and transparency. Grey Literature Reporter is a Google Chrome browser extension where users 'train' the tool on search results of a current website, manually tagging relevant data or keywords within the source. The tool automatically 'scrapes' the remaining webpages identified, recording results URL, title and search string used. Greylitsearcher is an R package and Shiny app which uses Google site:search functionality to search all of the pages of the websites specified, with advanced search capability. Users may save HTML data of the website pages, uploading these to the "scrape data" tab to return a CSV file with search results.	No published or unpublished research has sought to quantify the benefits of these tools. Anticipated or intended benefits include: efficiency and time saving in research data management (e.g. reducing time to devise, execute and compile search results for grey literature searching); more efficient and transparent practices for searching and results reporting.	Grey Literature Reporter: Further information is available here A link to download the Chrome browser extension is available here. Greylitsearcher: Shiny App Further information is available via Github