Canadian Journal of Health Technologies

Canada's Drug Agency
L'Agence des médicaments du Canada

October 2024 Volume 4 Issue 10

Drugs **Health Technologies** Health Systems

Methods and Guidelines

Development of an Evaluation Instrument on Artificial Intelligence Search Tools for Evidence Synthesis

Key Messages

What Was the Question?

There are conflicting calls for evidence synthesis producers to adopt or avoid recent artificial intelligence (AI) technologies. Our question was: How do we evaluate rapidly evolving AI tools to enhance the production of evidence syntheses and maintain quality standards?

What Did We Do?

To advance information retrieval science for producing evidence syntheses at Canada's Drug Agency (CDA-AMC), the Research Information Services team developed a process to evaluate promising AI search tools. We inventoried 51 tools in the fall of 2023, established selection criteria, assessed specific attributes, and built a standalone instrument to support continuous monitoring and evaluation of new tools.

What Did We Find?

Rapid development of AI search tools requires a flexible evaluation instrument to inform adoption decisions and enable comparison between tools. We identified mandatory and desirable characteristics for suitable AI tools to assist with information retrieval tasks conducted by CDA-AMC. This work enabled the development of a flexible instrument to evaluate novel AI search tools for evidence synthesis.

What Does This Mean?

CDA-AMC operationalized a replicable process to monitor and evaluate AI search tools. Our approach to evaluating AI technologies will advance information retrieval methods by CDA-AMC, and our evaluation instrument will assist any evidence synthesis producer interested in adopting AI search tools.

Table of Contents

Background 5

Evaluating the Performance of Automation or AI Tools for Evidence Synthesis 5

Project Overview 5

Defining AI Search Tools 6

Methods 6

Generating an Inventory of AI Search Tools 6

Results 10

Development of an Evaluation Instrument 10

Establishing a Replicable and Scalable Process 12

Discussion 12

Conclusion 13

References 14

Appendix 1: Inventoried Search Tools 16

List of Tables

Table 1: Selection Criteria 7
Table 2: Scoring Attributes 9

List of Figures

Figure 1: Flow of Tools Through Selection 8
Figure 2: A Sample Evaluation Created With the AI Search Tool Evaluation Instrument 11

Background

Evaluating the Performance of Automation or AI Tools for Evidence Synthesis

Evidence synthesis producers use a range of automation tools to expedite review tasks, including search alerts for identifying updated evidence sources, deduplication routines for managing search results, machine learning classifiers for screening studies, and more.[1-4] Novel generative AI tools may further assist with review tasks, such as search strategy development;[5] however, just as the preceding generation of technologies required performance testing and validation,[2,4] this new generation of tools also needs evaluation.[6,7]

Evaluation studies of automation tools that assist with information retrieval (i.e., "search") tasks are time-consuming to conduct and publish.[8] With developers constantly updating their tools and releasing new products, performance evaluations also grow outdated quickly. Aside from the challenges to produce and maintain performance evaluation of automation tools for search tasks in evidence synthesis, information specialists may also encounter lacklustre performance in real-world settings.[9,10]

As observed by comparing searches conducted with and without automated search tools, performance may not improve over usual practice, or may only improve in 1 area (e.g., time to search) but decrease in another (e.g., time to screen).[11] While generative AI search tools may surpass their predecessors' performance, evaluating these tools remains a necessary undertaking. Until we can demonstrate reliable performance in real-world settings, searches conducted with AI tools may supplement — but not replace — our usual practice. In addition, given the risk of autonomous technologies missing key evidence or reaching incorrect conclusions, we must avoid overreliance on new AI search tools by operators who cannot validate results.[12]

The Research Information Services (RIS) team at CDA-AMC comprises information retrieval experts with experience evaluating automation technologies for evidence synthesis.[13,14] The RIS team had both the incentive and the expertise to embark on a project in 2023 — following the release of ChatGPT by OpenAI and during the successive AI boom — to develop an efficient and replicable evaluation process to keep pace with technological innovations.

Project Overview

For CDA-AMC to leverage the potential benefits of automation technologies for evidence synthesis, we needed a continuous approach to monitor and assess new tools. Our objective was to advance the science for information retrieval work conducted by CDA-AMC by developing an evaluation approach for current and future AI search tools.

As an evidence synthesis producer and health technology assessment agency, CDA-AMC is not unique in needing to conduct individualized evaluations of automation tools. Our evaluation methods and the instrument developed through our project will be relevant for all evidence synthesis producers interested in using AI search tools.

Defining AI Search Tools

For the purposes of the evaluation project, we defined “AI search tools” to include established automation technologies and machine learning that had not been previously evaluated by CDA-AMC, in addition to newer generative AI technologies. There is a long history of automation use in the information sciences,[15] and we recognized AI as an umbrella term for technologies that performs tasks “that would ordinarily require biological brainpower to accomplish, such as making sense of spoken language, learning behaviours or solving problems.”[16] By adopting a broad definition that included successive generations of technologies, we aimed to recognize the potential value of both older and newer tools. For evaluating AI technologies to assist with search tasks, we further classified tools as “narrow” (i.e., intended to perform a specific task or set of related tasks)[17] or as “generative” (i.e., intended to produce content).[18]

We focused our evaluation on search tasks conducted by RIS at CDA-AMC and by all information specialists who work on evidence syntheses, as follows:

- interviewing: clarifying and refining the topic request
- search strategy development: testing terms, developing queries, determining sources (e.g., databases, websites, web search engines), and translating search queries for each source
- executing search queries in sources to generate results
- documenting the search strategy
- managing, deduplicating, and delivering results
- writing search methodology sections for manuscript.[19]

For information specialists working on systematic reviews, these tasks take, on average, 30 hours to complete but can range from 2 hours to 219 hours.[19] AI search tools already improve (or show promise to help improve) performance or reduce the amount of time spent on these tasks,[4] particularly search strategy development.[6]

To inform the development of our evaluation process, we inventoried a broad range of AI tools that included both narrow tools (e.g., citation analysis, text mining) and generative tools (e.g., large language-model chat bots) to assist with these search tasks.

Methods

Generating an Inventory of AI Search Tools

Tool Identification

To develop our evaluation process, we identified potential AI search tools by consulting 2 curated collections:

- **SR Toolbox**, an online catalogue maintained by the York Health Economics Consortium, the Evidence Synthesis Group at Newcastle University, and the School of Health and Related Research at the University of Sheffield; between November 2023 and December 2023, SR Toolbox contained 248 software tools to assist with any stage of the evidence synthesis process

- York Health Economics Consortium's list of citation analysis tools.

We also consulted the Royal Danish Library's project files on Artificial Intelligence and Literature Seeking.

These sources were supplemented with regular, automated updates from Semantic Scholar. Our team added 10 "seed articles" (i.e., known, relevant publications about automation tools for evidence synthesis) to a Semantic Scholar library. The library generated a weekly research feed of related publications that we screened and selected to further build our library. As of December 18, 2023, the Semantic Scholar research library contained 56 articles that we consulted for our inventory.

Although AI search tools continued to launch over our investigational period, we stopped adding to the inventory once we reached 51 tools (Appendix 1) so we could proceed with building our evaluation process. The increasing availability of new tools further reinforced our need for a standalone instrument to help evaluate future AI search tools.

Data Collection and Tool Selection

We consulted published studies that evaluated a range of databases or search systems[20-22] to design a data collection spreadsheet (in Microsoft Excel). Our team recorded 12 characteristics of the AI tools identified by the search (e.g., tool name, URL, developer, terms of use) to aid selection and scoring.

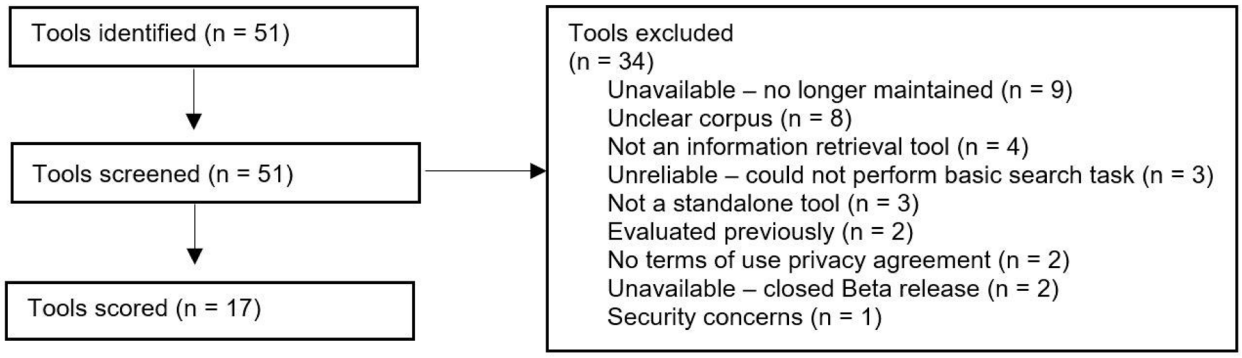
Using the same data sheet, we screened all identified AI tools according to inclusion and exclusion criteria (Table 1).

Table 1: Selection Criteria

Inclusion (must answer yes)	Exclusion (must answer no)
1. Does it support information retrieval tasks?	1. Is it a citation manager?
2. Is it oriented toward biomedicine?	2. Is it a full-text retrieval tool?
3. Is it currently available? Does it support a basic search?	3. Is it a screening tool?
4. Was it created or updated within the past 2 years?	4. Is it a data abstraction tool?
5. Is it a standalone tool?	5. Does it require programming knowledge?
6. Does the tool support English-language requests?	6. Could the terms of use restrict our abilities to incorporate this tool into review production?
7. Does the tool include clear terms of use?	7. Is it cost-prohibitive to use?
8. Are the corpus's contents clear, transparent?	8. Have we investigated this tool previously?

Seventeen selected tools (for which all inclusion criteria answers were "yes" or "unclear," and all exclusion answers were "no" or "unclear") proceeded to be pilot-tested and to have their attributes scored. In instances where our reviewers marked criteria as "unclear," we relied on the information from the tools' webpages. We did not contact tool developers to request additional information. Figure 1 shows the flow of tools through selection.

Figure 1: Flow of Tools Through Selection



Source: Figure adapted from: Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n7

Tool Characteristics

Collected information about our inventoried AI search tools helped us to conceptualize an overview of these tools available in 2023. We used the information to develop the attributes for scoring each tool.

AI Types

We classified 20% (10 of 51) of the AI search tools identified as “generative” (i.e., capable of performing a range of tasks, producing content such as article summaries) or having a generative AI component. The remaining 80% (41 of 51) were identified as “narrow” or “limited” (i.e., capable of performing a specific task).

Search Tasks

The tools claimed to perform a range of 1 or more search-related tasks. Overall, 61% of the tools (31 of 51) claimed to assist with paper identification, 33% (17 of 51) claimed to assist with network identification (e.g., the creation of a network of related publications), and 9% (5 of 51) claimed to assist with summary generation (e.g., summarizing a group of papers).

Costs

In total, 63% of the tools (32 of 51) were freely available at the time of evaluation, and 16% (8 of 51) offered a tiered pricing model with a free (“freemium”) version offering limited tasks. Subscription-based tools offered various pricing models based on institutional types (e.g., academic, business) or number of users (e.g., individual versus team). When published on their websites, listed subscription costs ranged widely from US$55 to US$8,340 annually. Many tools requested that subscribers contact the company for pricing information.

We did not exclude tools based on cost, but the criteria contributed to our scoring.

Developers

The tool developers were a mix of individuals, academics, and corporations. Developers based in the US represented 47% of the tools (24 of 51); but developers came from a wide range of countries including Australia, Bulgaria, Chile, Denmark, Germany, Hungary, India, Israel, Italy, Japan, the Netherlands, New

Zealand, Norway, Poland, Portugal, Spain, Switzerland, the UK, and Ukraine. Developers with international partnerships represented 12% of the tools (6 of 51).

Corpora

Overall, 86% (44 of 51) of the tools relied on corpora (i.e., data collections), and 12% of the tools (6 of 51) relied on the user to upload data. Commonly reported corpora included 1 or more open access, scholarly literature datasets of bibliographic records (e.g., PubMed, PubMed Central) or aggregated datasets (e.g., Lens.org contains records from PubMed, Crossref, Open Alex, Microsoft Academic Graph, and others).

Overall, 18% of the tools (9 of 51) used Crossref data for their corpus; 18% of the tools (9 of 51) used Semantic Scholar data for their corpus; and 6% (3 of 51) used Lens.org data for their corpus. Tools commonly provided data for other tools (e.g., Lens.org is the corpus for SpiderCite); less commonly, the tools operated with subscription databases (e.g., Citespace works with Web of Science records).

We excluded 8 tools for having an unclear corpus (i.e., we could not determine their data sources).

Access

We accessed 78% of the tools (40 of 51) via a website application and 12% of the tools (6 of 51) via a desktop application. For 7% (4 of 51), we could not determine how to access the tools, primarily because the tools were no longer available.

We excluded 9 tools for being unavailable. The breadth of tools in our inventory allowed us to observe this life cycle, where developers can suspend tool maintenance or development.

Pilot Testing and Scoring

To score our 17 selected tools, our team pilot-tested searches in 2 topic areas: *yoga* as an intervention, and *long COVID* as an indication. We picked *yoga* as a test search topic because of our experience with this intervention term yielding large quantities of variable-quality studies, and we picked *long COVID* as a challenging indication with many new studies and emerging terminology.

For pilot testing, we scored each tool according to 12 attributes (Table 2). We developed the attributes through discussion and consensus after exploring the characteristics of all the inventoried tools. We scored selected tools for each attribute using a scale of 0 to 3 (0 = insufficient; 3 = performs well). Published studies for evaluating technology tools for systematic reviews informed our scoring scale.[20,22]

Table 2: Scoring Attributes

Attribute	Scoring questions for reviewer
Novelty	Is it novel? Is this newer AI (e.g., generative) or older (e.g., narrow)? Do other tools offer the same functionality?
Replicability	Does it appear we can replicate methods to search the tool? Do sessions in different browsers (1 in incognito mode) return the same results? Will 2 people get the same results if using the tool?
Efficiency	Does the tool offer an increase in efficiency over our current processes?

Attribute	Scoring questions for reviewer
Compliance	Does the tool align with the GoC Guide on the use of Generative AI? Does the tool comply with copyright laws?
Comprehensiveness	Is the corpus large enough? Does the training set provide comprehensive coverage of biomedical scholarly literature? Is there a gap for specific publication types (e.g., conference proceedings, theses, trial registry records)?
Coverage	Are the coverage dates for the corpus sufficiently large for CDA-AMC products? Does the tool cover publications from at least the past 10 years?
Currency	Is the dataset updated frequently? Is there a coverage gap for recent publications?
Usability	Can we use this tool with our current methods and systems? Does the export function support our use of the tool? Does the file format correspond with our existing citation managers and screening tools?
Reliability	Was the tool available during testing? Did it operate unpredictably? Are we aware of user complaints?
Adoptability	Does it appear relatively easy to learn and to use? Are there training materials available?
Support	Is user support available?
Value	Would we use this for our work? Is this a potentially valuable tool?

AI = artificial intelligence; CDA-AMC = Canada's Drug Agency; GOC = Government of Canada.

Results

Development of an Evaluation Instrument

Our data collection spreadsheet allowed us to compare tools, but it proved unwieldly to maintain as new tools became available for evaluation. We refined our approach through reflection on the process to evaluate our inventory of tools. An adapted evaluation instrument enables rapid evaluation of individual tools by guiding the evaluator through 8 characteristics questions, 13 selection criteria, and 10 scoring attributes. The instrument then guides the evaluator to score the tool out of 30.

Our standalone evaluation instrument is available as a Microsoft Excel worksheet for use by evidence synthesis producers.

To validate the evaluation instrument, we conducted 2 exercises: we used the instrument to evaluate 3 tools from our project, then compared the selection decisions and scores; and we also used the instrument to conduct a new evaluation for a tool that we had not previously evaluated. Further enhancements to the tool resulted from our validation exercises, but we found that the tool performed comparably to the data collection instrument for our evaluation project. Figure 2 presents a sample evaluation using the instrument.

Figure 2: A Sample Evaluation Created With the AI Search Tool Evaluation Instrument

AI Search Tool Evaluation Instrument			
Version 1.0			
	Evaluated by:	Sample evaluator	
	Evaluation date:		06-Sep-24
STEP 1			
Record tool information			
	Tool name:	Sample tool	
	Source URL (if applicable):	https://sampletool.com	
	Task the tool claims to perform:	Network identification	
	Cost:	Free	
	Corpus - source of the data (if applicable):	Sample corpora	
	Terms of use (include URL):	https://sampletool/terms.com	
	Developer name:	Sample developer	
	Developer country/jurisdiction of usage terms:	Sample country	
	Proceed to step 2		
STEP 2			
Apply criteria (must answer yes or unsure)	**Yes**	**No**	**Unsure**
Does it support information retrieval tasks?	X		
Is it currently available?	X		
Was it created or updated within the last 2 years?	X		
Does it support English language requests?	X		
Does the tool include clear terms of use?	X		
Are the terms of use acceptable?			X
Does the tool comply with copyright laws?			X
Are the corpus' contents clear and transparent?	X		
Is the currency of the corpus adequate for our needs?			X
Can we afford the cost to evaluate this tool?	X		
Can we afford the cost to incorporate this tool into our work?	X		
Do we have the technical knowledge to use this tool?	X		
Can we verify its performance?	X		
Proceed to step 3			
STEP 3			
Select to continue evaluation	**Evaluate**		
Proceed to step 4 if decision is "Evaluate"			
STEP 4			
Conduct test searches and score each attribute out of 3 (0=insufficient; 3=very good)	**Score**		**Testing notes**
EFFICIENCY: Does the tool offer an increase in efficiency over our current processes?	1 /3		
USABILITY: Can we use this tool with our current methods and systems?	3 /3		
VALUE: Would we use this tool for our work? Is this a potentially valuable tool?	1 /3		
COMPREHENSIVENESS: Is the corpus large enough for our work? Are there gaps for topics or publication types?	1 /3		
COVERAGE/CURRENCY: Is the corpus frequently updated? Are there gaps for older or newer information?	1 /3		
RELIABILITY: How did the tool perform during testing? Are we aware of user complaints?	1 /3		
ADOPTABILITY: Does it appear relatively easy to learn/use? Are there available training materials?	3 /3		
SUPPORT: Is user support available?	1 /3		
NOVELTY: Does this tool offer unique value? Do other tools offer the same functionality?	1 /3		
REPLICABILITY: Can we reliably replicate methods to search the tool?	0 /3		
Total	**13 /30**		
Scoring guide			
0-10: Revaluate if warranted by future development, recommendations, or validation studies			
11-20: Proceed with caution and consider only to supplement usual practice			
21-30: Consider incorporating into search methods			

AI = artificial intelligence; Sep = September.

Establishing a Replicable and Scalable Process

To monitor the availability of new tools, we continuously receive automated alerts of methods papers on information retrieval and alerts from Semantic Scholar using the seed evaluation studies that we identified for this project. As we learn of new AI search tools, we rapidly screen and score those tools using our evaluation instrument.

If a tool meets our selection criteria, it is pilot-tested and scored out of 30 using the 10 refined attributes that our team identified through our evaluation of our inventory tools. To aid in decision-making for scored tools, we added a guide to the instrument that suggests 3 possible actions: reevaluate if warranted by future development, recommendations, or validation studies (for tools that score 0 to 10); proceed with caution and consider only to supplement usual practice (for tools that score 11 to 20); and consider incorporating into search methods (for tools that score 21 to 30).

Using the evaluation instrument, we select and score tools as they become available to us. We use the scores to compare available AI search tools and to inform our adoption decisions.

Discussion

A strength of the inventory project was that it identified key questions to determine the appropriateness of automation tools to support the work of CDA-AMC. For example, many tools claimed to deliver comprehensive information, but often we could not determine the tools' data sources. Tools promoted the number of items they contained (e.g., "over 7 million articles") without sharing details about the items' scope, coverage dates, currency, and so on. A major take-away for our team was to always check a tool's corpus to determine its adequacy and to check for overlap with existing databases and collections.

Another take-away for our team was to always assess the tools' terms of use to determine what rights we retain to the information we input (e.g., search strategies, PDFs) and the output generated by the tool (e.g., data visualizations, lists of articles). Some tools required us to upload full-text PDF articles in violation of licence agreements. In other cases, tools generated outputs that we were restricted to reproduce or use. We learned to verify any risks or restrictions before inputting or outputting data to these tools.

In addition to learning key questions for evaluating AI tools (e.g., What is its corpus? What are its terms of use? Can we export the results?), CDA-AMC gained essential knowledge about the range of available AI tools to assist with search tasks for evidence synthesis. We benefited from having the opportunity to test a wide range of AI tools and identified multiple tools that may be useful for our work.

While we aimed to identify all AI search tools available to our team, the rapid development of automated tools for evidence synthesis forced us to limit the number of tools we screened and scored. It is a limitation of our inventory that we could not identify all automation tools to assist with evidence synthesis production. The summary of the characteristics of AI search tools presented previously may not represent all available tools in the fall of 2023.

We also acknowledge that our definition of AI search tools may be too inclusive of older automation technologies, but this definition enabled RIS to evaluate a wider range of potentially useful technologies. Incorporating more technologies into our definition also allowed us to develop an inclusive instrument that evaluates all automated search tools for evidence syntheses and not just for a narrow subclass of tool types.

We further recognize the limitation that our selection criteria and scoring attributes did not account for potential sources of bias (e.g., language bias). We also did not attempt to evaluate bias using our standalone evaluation instrument. Inherently, all corpora (data sources) used for our usual practice (e.g., PubMed, ClinicalTrials.gov) and to train AI search tools contain bias. Determining the underlying biases of these corpora proved too large an undertaking for our project scope. We acknowledge the need for future investigations by producers and increased transparency by tool developers to uncover biases in the automation tools used for evidence syntheses.

Finally, we acknowledge the limited validation of the evaluation instrument available via this report. As we continue to evaluate new tools, we anticipate that the instrument will undergo additional modifications. We also welcome feedback and suggestions on the instrument and encourage audiences to adapt it as needed for their own purposes. Our team will update our evaluation approach or make new versions of the tool available as the AI technologies space evolves.

As internal evaluations of AI tools are justified, it follows that internally tailored evaluation instruments will be the most useful for such evaluations. We share our instrument knowing that evidence synthesis producers all work in unique environments but share common goals. The evaluation instrument aims to help serve a common goal: to advance our methods while maintaining standards for high-quality evidence syntheses.

Conclusion

Our inventory showed the breadth of available automation search tools for consideration by CDA-AMC and other evidence synthesis producers. An evaluation of these inventoried tools identified key characteristics to inform selection and scoring of tools. This evaluation approach resulted in a standalone instrument for use by CDA-AMC and other evidence synthesis producers for monitoring and evaluating search tools as AI technologies evolve.

Automation technologies, including generative AI, can assist with search tasks for evidence syntheses, but internally conducted evaluations of the utility and appropriateness of automation tools are needed now and in the future. As CDA-AMC identified and evaluated promising AI search tools, developers continued to release new products. To address the volume of novel products, we operationalized a replicable and scalable process. This continuous monitoring and evaluation will ensure that CDA-AMC leverages the potential of AI search tools to improve the performance of current and future work.

References

1. Cochrane. Artificial intelligence (AI) technologies in Cochrane [webinar recording]. 2024: https://training.cochrane.org/resource/artificial-intelligence-technologies-in-cochrane/. Accessed 2024 Sep 9.
2. Beller E, Clark J, Tsafnat G, et al. Making progress with the automation of systematic reviews: principles of the International Collaboration for the Automation of Systematic Reviews (ICASR). *Systematic Reviews.* 2018;7(1):77. 10.1186/s13643-018-0740-7 PubMed
3. Tsafnat G, Glasziou P, Choong MK, Dunn A, Galgani F, Coiera E. Systematic review automation technologies. *Systematic Reviews.* 2014;3(1):74. 10.1186/2046-4053-3-74 PubMed
4. van Altena AJ, Spijker R, Olabarriaga SD. Usage of automation tools in systematic reviews. *Research Synthesis Methods.* 2019;10(1):72-82. 10.1002/jrsm.1335 PubMed
5. Alaniz L, Vu C, Pfaff MJ. The Utility of Artificial Intelligence for Systematic Reviews and Boolean Query Formulation and Translation. *Plast Reconstr Surg Glob Open.* 2023;11(10):e5339. 10.1097/GOX.0000000000005339 PubMed
6. NICE. Use of AI in evidence generation: NICE position statement. Version 1.0. London (UK). 2024: https://www.nice.org.uk/about/what-we-do/our-research-work/use-of-ai-in-evidence-generation--nice-position-statement. Accessed 2024 Sep 13.
7. Wildgaard LE, Vils A, Sandal Johnsen S. Reflections on tests of AI-search tools in the academic search process. *LIBER Quarterly: The Journal of the Association of European Research Libraries.* 2023;33(1). 10.53377/lq.13567
8. O'Connor AM, Tsafnat G, Thomas J, Glasziou P, Gilbert SB, Hutton B. A question of trust: can we build an evidence base to gain trust in systematic review automation technologies? *Systematic Reviews.* 2019;8(1):143. 10.1186/s13643-019-1062-0 PubMed
9. Suppadungsuk S, Thongprayoon C, Krisanapan P, et al. Examining the Validity of ChatGPT in Identifying Relevant Nephrology Literature: Findings and Implications. *J Clin Med.* 2023;12(17). 10.3390/jcm12175550 PubMed
10. Blum M. ChatGPT Produces Fabricated References and Falsehoods When Used for Scientific Literature Search. *J Card Fail.* 2023;29(9):1332-1334. 10.1016/j.cardfail.2023.06.015 PubMed
11. Paynter RA, Featherstone R, Stoeger E, Fiordalisi C, Voisin C, Adam GP. A prospective comparison of evidence synthesis search strategies developed with and without text-mining tools. *J Clin Epidemiol.* 2021;139:350-360. 10.1016/j.jclinepi.2021.03.013 PubMed
12. Hersh W. Search still matters: information retrieval in the era of generative AI. *Journal of the American Medical Informatics Association.* 2024;31(9):2159-2161. 10.1093/jamia/ocae014 PubMed
13. Glanville J, Wood H. Text Mining Opportunities: White Paper. CADTH. 2018: https://www.cda-amc.ca/sites/default/files/pdf/methods/2018-05/MG0013_CADTH_Text-Mining_Opportunites_Final.pdf. Accessed 2024 Sep 14.
14. Canada's Drug Agency. *Ovid Multifile Searches with our Filters.* 2023: https://www.cda-amc.ca/ovid-multifile-searches-filters.
15. Cox AM. The impact of AI, machine learning, automation and robotics on the information professions: A report for CILIP. CILIP: the Library and Information Association. 2021: https://www.cilip.org.uk/page/researchreport. Accessed 2024 Sep 13.
16. Government of Canada. Directive on Automated Decision-Making. Ottawa (ON). 2023: https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592. Accessed 2024 Jul 24.
17. DeepAI. Understanding Narrow AI: Definition, Capabilities, and Applications. 2024: https://deepai.org/machine-learning-glossary-and-terms/narrow-a. Accessed 2024 Mar 7.
18. McKinsey & Company. What is generative AI? 2023: https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai. Accessed 2023 Sep 13.
19. 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. 10.5195/jmla.2018.323 PubMed
20. Bethel A, Rogers M. A checklist to assess database-hosting platforms for designing and running searches for systematic reviews. *Health Info Libr J.* 2014;31(1):43-53. 10.1111/hir.12054 PubMed

21. Gusenbauer M, Haddaway NR. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research Synthesis Methods.* 2020;11(2):181-217. 10.1002/jrsm.1378 PubMed

22. Cooper C, Brown A, Court R, Schauberger U. A Technical Review of the ISPOR Presentations Database Identified Issues in the Search Interface and Areas for Future Development. *International Journal of Technology Assessment in Health Care.* 2022;38(1):e29. 10.1017/S0266462322000137 PubMed

Appendix 1: Inventoried Search Tools

- 2Dsearch
- Author Name Disambiguation
- Anne O'Tate
- AntConc
- Apples-to-Apples
- Article Galaxy
- BeCAS
- Buhos
- CADIMA
- Carrot2
- ChatGPT
- citationchaser
- CiteSeer[x]
- CiteSpace
- Connected Papers
- Consensus
- Copilot
- Dimensions
- Elicit
- iCite
- Inciteful
- Infomap Online
- Iris.ai

- Lens.org
- Litmaps
- Local Citation Network
- Nested Knowledge
- Network Workbench
- Omnity
- OpenAlex
- Oyster
- Paperfetcher
- PaperQA
- Perplexity
- Research Rabbit
- Result Assessment Tool
- Sci2
- ScienceOpen
- SciSpace (formerly Typeset)
- Scilit
- Scite.ai assistant
- Scite.ai search
- Scopus AI
- Semantic Reader
- Semantic Scholar
- SnowGlobe
- SpiderCite

- System Pro
- VisualBib
- Yewno
- Zeta Alpha

Canada's Drug Agency
L'Agence des médicaments du Canada
Drugs. Health Technologies and Systems. Médicaments, technologies de la santé et systèmes.

ISSN: 2563-6596

Canada's Drug Agency (CDA-AMC) is a pan-Canadian health organization. Created and funded by Canada's federal, provincial, and territorial governments, we're responsible for driving better coordination, alignment, and public value within Canada's drug and health technology landscape. We provide Canada's health system leaders with independent evidence and advice so they can make informed drug, health technology, and health system decisions, and we collaborate with national and international partners to enhance our collective impact.

Questions or requests for information about this report can be directed to Requests@CDA-AMC.ca.