



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

A test-and-learn program to advance information retrieval methods with Artificial Intelligence

IRM 2026

Robin Featherstone, MLIS

Acting Director, Research Information Services

April 24, 2026

Disclosures



Robin Featherstone

- Full-time employee of Canada's Drug Agency; no relevant disclosures

Canada's Drug Agency

- The organization is funded by contributions from the Canadian federal, provincial, and territorial ministries of health.
- We receive application fees from the pharmaceutical industry for:
 - Our Reimbursement Review processes, including those used for:
 - oncology drugs
 - non-oncology drugs
 - plasma protein and related products reviewed through the interim process
 - Scientific Advice

Inflated expectations in 2023



ChatGPT Tutorial: Write a systematic review under 1 hour

177K views · Jan 21, 2023
YouTube › Benjamin Tran, MD



Do hours of Research Literature review in minutes using this Ne...

110.9K views · 5 months ago
YouTube › Advanced ChatGPT



Best AI tools for research paper writing and systematic review b...

3.3K views · 4 months ago
YouTube › Insights4UToday



Followed by disillusionment

ARTIFICIAL INTELLIGENCE

Why Meta's latest large language model survived only three days online

Galactica was supposed to help scientists. Instead, it mindlessly spat out biased and incorrect nonsense.

JCF IGNITE! | VOLUME 29, ISSUE 9, P1332-1334, SEPTEMBER 2023

Download Full Issue



Purchase Subscribe

ChatGPT Produces Fabricated References and Falsehoods When Used for Scientific Literature Search

Moritz Blum, MD

Published: July 03, 2023 • DOI: <https://doi.org/10.1016/j.cardfail.2023.06.015>

Beware of references when using ChatGPT as a source of information to write scientific articles

Luis Sanchez-Ramos, MD • Lifeng Lin, PhD • Roberto Romero, MD, DMedSci

Published: April 06, 2023 • DOI: <https://doi.org/10.1016/j.ajog.2023.04.004> • Check for updates

How we responded to the AI hype



Research Information Services (RIS) undertook a phased approach to leverage the potential of new search tools for HTA information retrieval



*Phase 1 FY 2023-2024 - **Inventory***

- Built an inventory of automated or AI search tools
- Ranked and selected tools for testing
- Produced an evaluation instrument



*Phase 2 FY 2024-2025 - **Testing***

- Performance tested selected tools using a retrospective study design
- Developed implementation recommendations



*Phase 3 FY 2025-2026 - **Development***

- Developed custom search tools
- Initiated performance testing of app-assisted grey literature searching using a prospective study design

How we defined AI tools in 2023

AI:

Umbrella term for information technology that performs tasks that would ordinarily require biological brainpower²

Generative AI:

produces content³

e.g., produces reference lists,
generates a search strategy,
makes network graphs, etc.

Narrow AI:

performs a specific task or set of
related tasks⁴

e.g., identifies potential references
through citation analysis, finds
duplicate references, etc.

2. Government of Canada. *Directive on Automated Decision-Making*. Ottawa 2023.

3. McKinsey & Company. What is generative AI? 2023; <https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai>. Accessed September 13, 2023.

4. DeepAI. Understanding Narrow AI: Definition, Capabilities, and Applications. 2024; <https://deepai.org/machine-learning-glossary-and-terms/narrow-a>. Accessed March 7, 2024.



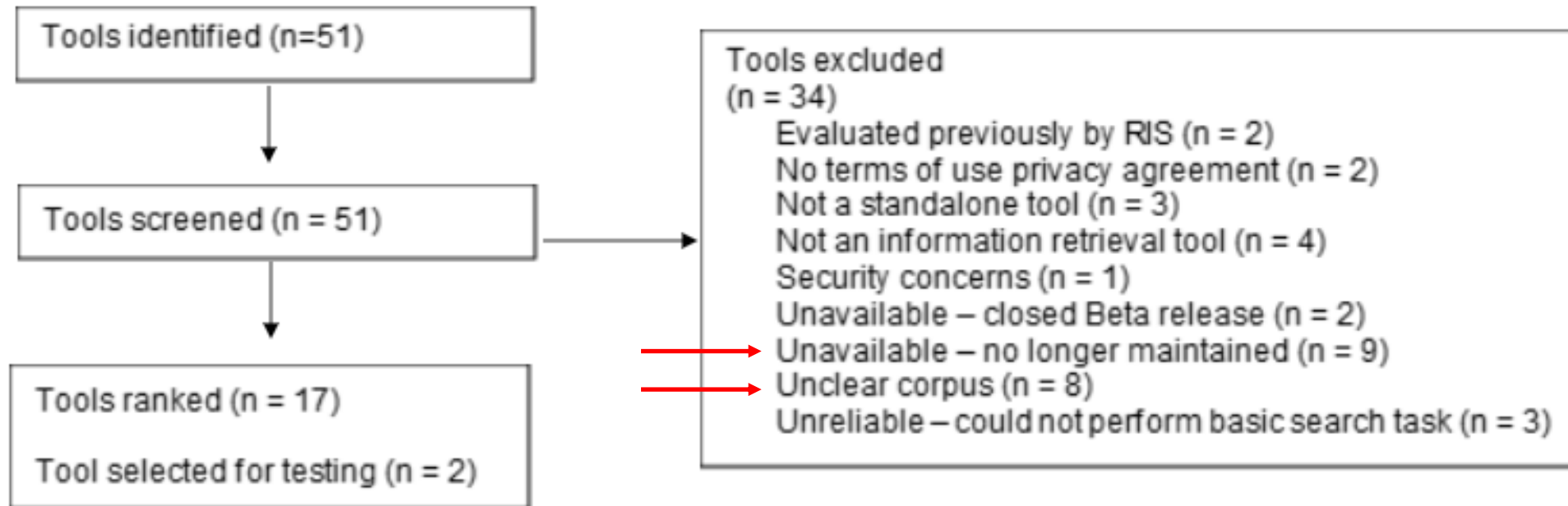
Canada's Drug Agency
L'Agence des médicaments du Canada
Drugs. Health Technologies and Systems. Médicaments, technologies de la santé et systèmes.



Phase 1: Inventory & Evaluation Instrument



Flow of tools through screening and selection





How we ranked eligible tools

Attribute	Ranking questions for reviewer	Weight (max 1.0)
A1. NOVELTY	Is it novel? Is this newer AI (e.g., generative) or older (e.g., narrow)? Do other tools offer the same functionality?	0.25
A2. REPLICABILITY	Can we 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?	0.25
A3. EFFICIENCY	Does the tool offer an increase in efficiency over our current processes?	1.0
A4. COMPLIANCE	Does the tool align with the GoC Guide on the use of Generative AI? Does the tool comply with copyright laws?	1.0
A5. COMPREHENSIVENESS	Is the corpus large enough? Does the training set large provide comprehensive coverage of biomedical scholarly literature? Is there a gap for specific publication types (e.g., conference proceedings, theses, trial registry records, etc.)	0.75
A6. COVERAGE	Are the coverage dates for the corpus sufficiently large for CADTH products? Does the tool cover publications from at least the last 10 years?	0.75
A7. CURRENCY	Is the dataset updated frequently? Is there a coverage gap for recent publications?	0.75
A8. 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?	1.0
A9. RELIABILITY	Was the tool available during testing? Did it operate unpredictably? Are we aware of user complaints?	0.75
A10. ADOPTABILITY	Does it appear easy to learn/use? Are there training materials available?	0.75
A11. SUPPORT	Is user support available?	0.5
A12. VALUE TO RIS	Would we use this for our work? Is this a potentially valuable tool?	1.0



Most important attribute



How the tools did

Tool name	Score (/36)	Weighted Score (/100)
Lens.org	27.5	85.95
SpiderCite	28	85.71
iCite	29.25	84.52
SciLit	26.5	84.05
Citationchaser	27	83.81
LitMaps	29.25	83.10
Semantic Scholar	29.5	82.14
Connected Papers	28.5	80.95
Inciteful	28.25	80.95
2dSearch	27.25	79.29
Scite.ai	27.5	77.38
System Pro	26.75	76.90
Iris AI	26.75	74.76
Consensus	26.75	74.52
Local Citation Network	24.75	74.52
Anne O'Tate Citation Cloud	23.5	72.86
Elicit	25.75	70.71

Selected for testing



What we learned

1. Always check the corpus (i.e., the data sources)
 - Where does the data come from?
 - How comprehensive is its coverage? Are there gaps?
2. Read the terms and conditions
 - Who own the inputs? Does using the tool break our license agreements?
 - Who owns the outputs? Does the tool assert ownership?



Evaluation instrument



Methods and Guidelines

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

Selection

Scoring

Artificial Intelligence Search Tool Evaluation Instrument			
Version 1.1			
Evaluated by: Robin Featherstone		Evaluation date: 06-Sep-24	
STEP 1			
Record tool information			
Tool name: Copilot		Source URL (if applicable): https://copilot.microsoft.com	
Task the tool claims to perform: Provides AI-powered chat for the web		Cost: Copilot is part of our M365 enterprise	
Corpus - source of the data (if applicable): Bing Search		Terms of use (include URL): https://learn.microsoft.com/en-us/c	
Developer name: Microsoft		Developer country/jurisdiction of usage terms: United States	
Proceed to step 2			
STEP 2			
Apply criteria (must answer yes or unsure)			
		Yes	No
	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)			
	EFFICIENCY: Does the tool offer an increase in efficiency over our current processes?	Score	Testing notes
	USABILITY: Can we use this tool with our current methods and systems?	1 / 3	Unclear if we a
	VALUE: Would we use this tool for our work? Is this a potentially valuable tool?	3 / 3	
	COMPREHENSIVENESS: Is the corpus large enough for our work? Are there gaps for topics or publication types?	1 / 3	Further testing
	COVERAGE/CURRENCY: Is the corpus frequently updated? Are there gaps for older or newer information?	1 / 3	Further testing
	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: Reevaluate 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			



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



Phase 2: Testing



What we did for Phase 2

We followed a protocol to determine the performance of Lens.org, SpiderCite, and Microsoft Copilot by

- Reviewing evidence reports that used these tools
- Running a retrospective comparative analysis using our projects
- Conducting a focus group to gather user feedback on the tools

Anticipating AI for Information Retrieval: Phase 2 Testing Protocol Research Information Services Aug 8, 2024

Overview: For phase 2 (testing), we will investigate the performance of 3 tools to assist with RIS information retrieval methods:

- Lens.org <https://www.lens.org>
- SpiderCite <https://sr-accelerator.com/#/spidercite>
- Microsoft Copilot <https://www.microsoft.com/en-ca/microsoft-copilot>

Our objective is to determine if AI tools should supplement current RIS information retrieval methods given the estimated contribution of eligible and unique studies, and the time and cost to search and screen results.

Methods: We will compare test searches and reference searches for completed projects to measure tool performance for time, cost, search precision, search sensitivity, unique contribution, and impact on effect estimates. We will also conduct a literature review to estimate the reported use of these tools in research synthesis publications.

Reference searches: Our reference standard for comparison will be search results from completed/ongoing projects. We identified a convenient sample of projects (ongoing or published in the past 24 months) for a range of topics and methodologies with at least five eligible articles (i.e., selected articles incorporated into the reviews' findings):

Sample projects:



Comparative analysis

- Assigned an Information Specialist (IS) to one of 7 selected completed projects
- Reran the “usual practice” search for the project (i.e., our “reference search”) and executed test searches using the 3 tools
- Compared search performance using values for
 - Sensitivity
 - Number Needed to Read (NNR)
 - Time
 - Unique contributions

Code	Type	Topic	Title
RC1504	Simple	Device	Radiofrequency ablation for chronic knee, hip, and shoulder pain
RC1522	Simple	Drug	Ketamine for adults with treatment resistant depression or PTSD: a 2023 update
OP0553	Complex	Health System	ED overcrowding – overview of reviews
RC1476	Simple	Procedure	Interventions for the presurgical management of knee osteoarthritis: a rapid qualitative review
HC0043	Complex	Health System	Considerations of access and inclusion in adolescent eating disorder care: a custom rapid report
HC0080	Complex	Health System	Culturally appropriate care for the treatment or management of substance or opioid use
SR0827	Simple author search	Indication	Clinical author identification: Schizophrenia

Why Lens.org, SpiderCite, and *Microsoft Copilot*?

Lens.org

- Comprehensive scholarly works content set with available AI-driven discovery and analytics tools
- Potential to incorporate into our work with support for complex queries (2K character limit), mass export of search results in .ris format



SpiderCite

- Reliable tool for citation searching with comprehensive coverage (Lens.org content)



Microsoft Copilot

- Opportunity to test a large language model (LLM) with additional data security through our license and enterprise data protection
- Uses a model built upon OpenAI's GPT-4 LLM with a similar interface to the ChatGPT chatbot studied for search strategy development





What we found

Key finding from the literature review:

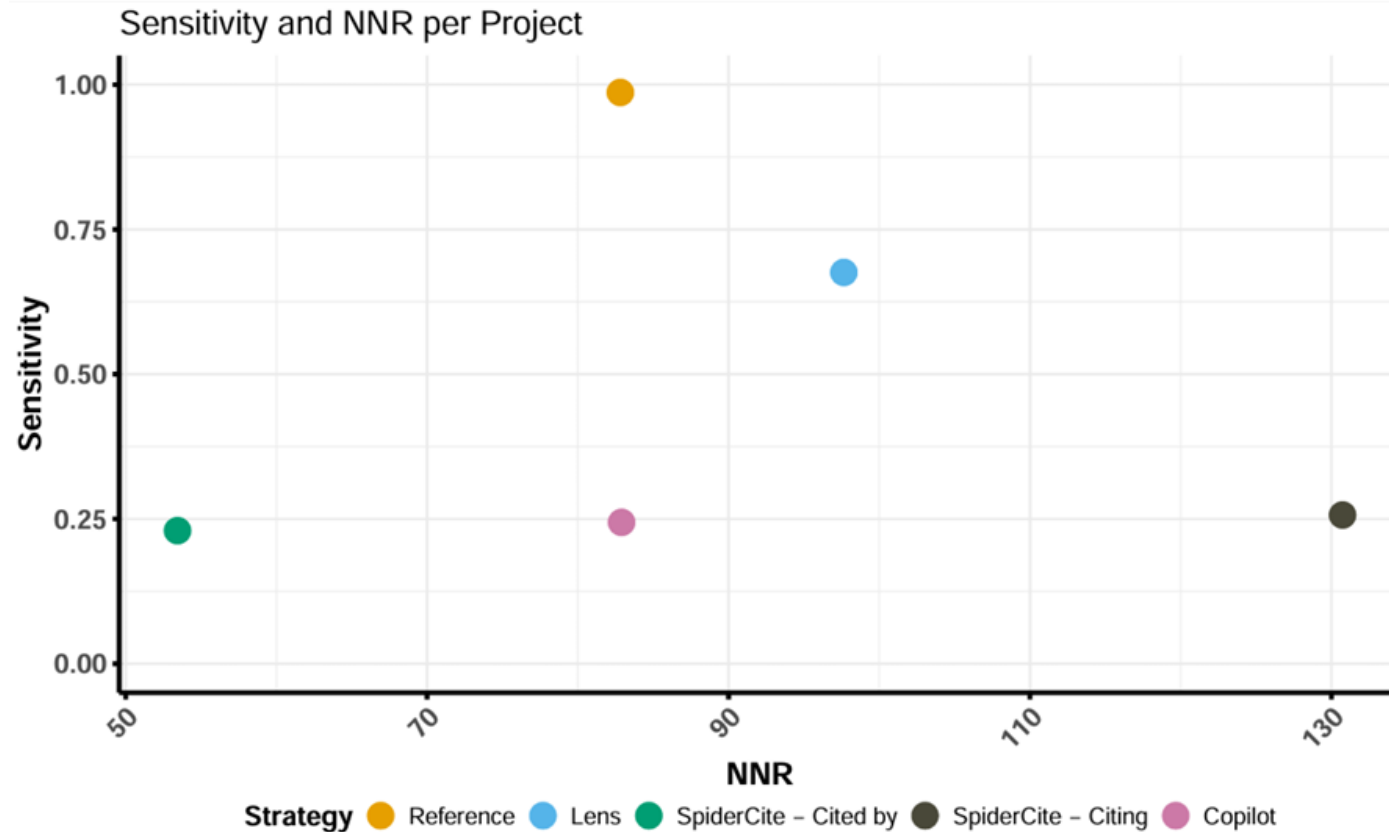
- AI or automated search tools were very rarely used as the only source of evidence

Tool	Total Results	Used as Sole Source	Used with Other Sources	Unclear
Lens.org	109	25 (22%)	79 (72%)	5 (4%)
SpiderCite	41	0	41 (100%)	0
Copilot	8	0	8 (100%)	0



What we found

AI or automated search tools have low sensitivity (do not capture all eligible articles) and rarely decrease screening burden



What we found

Unique studies captured by AI or automated search tools for complex studies are more likely to impact the projects' findings

Projects	Likelihood that unique articles not captured by the reference searches could impact the project's findings ^a			
	Lens.org	SpiderCite – Cited by	SpiderCite – Citing	Copilot
Simple projects				
1	Likely: 2-3 items meet criteria for further assessment	Likely: 2-3 items meet criteria for further assessment	Unlikely: <2 items meet criteria for further assessment	Very likely: ≥4 items meet criteria for further assessment
2	Unlikely: <2 items meet criteria for further assessment	Very unlikely: 0 items meet criteria for further assessment	Very unlikely: 0 items meet criteria for further assessment	Very unlikely: 0 items meet criteria for further assessment
4	Unlikely: <2 items meet criteria for further assessment	Unlikely: <2 items meet criteria for further assessment	Very unlikely: 0 items meet criteria for further assessment	Very unlikely: 0 items meet criteria for further assessment
Complex projects				
3	Likely: 2-3 items meet criteria for further assessment	Very unlikely: 0 items meet criteria for further assessment	Very likely: ≥4 items meet criteria for further assessment	Very likely: ≥4 items meet criteria for further assessment
5	Very likely: ≥4 items meet criteria for further assessment	Very likely: ≥4 items meet criteria for further assessment	Very likely: ≥4 items meet criteria for further assessment	Very likely: ≥4 items meet criteria for further assessment
6	Very likely: ≥4 items meet criteria for further assessment	Very likely: ≥4 items meet criteria for further assessment	Very likely: ≥4 items meet criteria for further assessment	Very unlikely: 0 items meet criteria for further assessment
Simple author search				
7	Likely: 2-3 items meet criteria for further assessment	NA	NA	Very likely: ≥4 items meet criteria for further assessment



What we found about generative AI

LLMs can help us in unanticipated ways

Surprising Uses

An IS described their surprise at Copilot being able to append an agency filter to a search strategy without first providing the chatbot with the filter. It appears that the agency's publicly available search filters were part of the Large Language Model's training set. Prompting Copilot to, "add the CDA search filter for randomized control trials," turned out to be a viable approach to developing a search strategy.

For the simple author search, Copilot retrieved some real Canadian authors. While more searching was needed to validate these individuals and to locate their publications, Copilot also helped us improve our geographic filter by providing a large list of domains for Canadian institutions and agencies.



What we found about LLMs for *database* searching

	Microsoft Copilot	Usual Practice
Viability	Only as a supplemental approach	Functions reliably
Accuracy (sensitivity)	Found 24% of the studies that we wanted it to find*	Found 99% of the studies that we wanted it to find*
Efficiency	0.96 hours/search*	2.88 hours/search*
User Experience	Mixed: “kind of fun”; “worked fine”; “ my nemesis ”; “the most frustrating process I’ve ever been through”	Very acceptable: “I would probably go to the tools that I’m familiar with and I trust and I understand”

* Average values for 6 projects

Minimal efficiency gains (1.92 hours) by adopting Copilot for database searching with unacceptable losses in accuracy



What we learned from phase 2

1. Viable tools are more likely to complement (not replace) usual practice
2. Tool acceptance depends on trust
3. For efficiency, target tasks with *potentially larger gains*
4. If tool adoption is limited by our ability to change processes, we may have to build tools if we want them to work for us



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

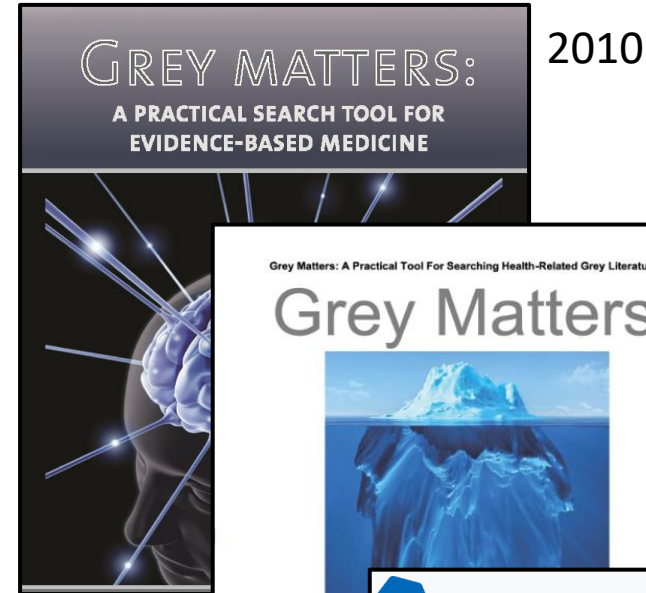


Phase 3: Development

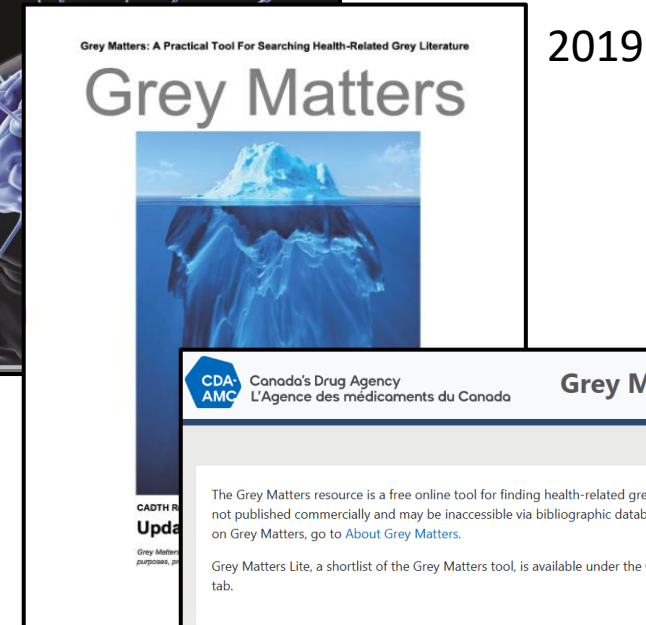


Grey literature checklists

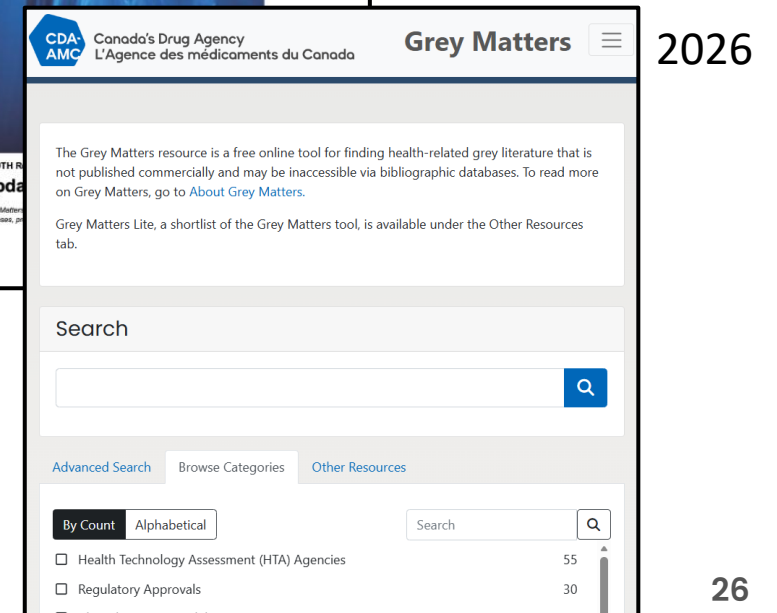
- Grey literature checklists were first created at our organization > 25 years ago
 - Simple lists of websites to search for grey literature
 - Not meant to be exhaustive but aims to be comprehensive for Canadian and international HTA agencies
 - Originally meant for internal use only, with different checklists for each of our product types
- We started sharing a public checklist (Grey Matters) > 15 years ago



2010



2019



2026



Search

 Search

For this product type, as many as 79 websites are manually checked for grey literature taking approx. 12.5 hours

Grey Literature Checklists	Category	Topic Specific Checklists	Advanced Search
Grey Matters			169
Health Technology Review			193
Horizon Scan			97
Non-Sponsored Drug Review – Backgrounder			79
Non-Sponsored Drug Review – Trial Registries			5
Rapid Review			35
Rapid Review – Mandatory			10
Sponsored Drug Review – Guidelines			18
Sponsored Drug Review – HTA Agencies			6

▲ RIS Searches, Screening, Article Ordering & Delivery	55h
Complete IS initial tasks	3h
Complete backgrounder	12.5h ← Grey lit search
Complete literature search for CL review	9h
Complete NMA search	2h ← Database search



New possibilities

	Phase 1	Phase 3
Partnership opportunities	Limited	Internal and external
Available AI tools	Traditional, non-deterministic	Agentic, deterministic

With the emergence of agentic AI tools, we could target app development to assist with grey literature searching

Traditional GenAI	Agentic GenAI
<p>Direct answer generation At least not more than a few steps, no lanes</p>	<p>Multi-step problem processing Complex processing flows possible, multiple lanes</p>
<p>Dependent Can only act, when being prompted by a human</p>	<p>Autonomous Can start actions on its own based on internal or external triggers</p>
<p>Can answer Answer a user's questions or generate an output (image, movie)</p>	<p>Can reason and decide Can reason before acting, decide on next steps and explain results</p>
<p>Stateless No memory or very short-term memory in session (minutes)</p>	<p>Stateful Can keep a context for days, weeks, months, years</p>
<p>No tool use Generate the answers based on input and pre-training data.</p>	<p>Tool use Uses databases, RAG, APIs, to process a problem.</p>
<p>Non-deterministic Responses are generated almost fully based on stochastic</p>	<p>Large deterministic parts Extensive deterministic chains help to process, check, safe, verify results</p>



Development process

1. RIS established user requirements
2. IT Services prepared a technical proposal (including a budget)
3. RIS & IT Services developed a testing plan
4. IT Services initiated development work
5. RIS reviewed and provided feedback on prototype tools

User Requirements: Grey Literature Search Assistant

Date created: July 23, 2025

Robin Featherstone, David Kaunelis, Danielle MacDougall

Requirements

1. Retrieve reports from specified web sites

As a user, I want to find reports on topics provided as keywords from web sites specified in [Grey Literature Checklists](#) (pre-existing lists of sites).

As a user, I will prompt the tool by entering text words for the topic and selecting one or more checklists. From a list of potential results, I will select results to download.

2. Generate machine-readable metadata of reports

As a user, I want to retrieve a single [RIS-formatted text file](#) with metadata for all selected results that I can upload to Endnote citation management software or other systematic review software tools (e.g., Covidence, DistillerSR).

3. Save query prompt

As a user, I want the option to save query prompts so I can report these search methods, share my approach with other team members, and replicate the search for future updates.

4. Monitor available reports from specified websites

As a user, I want the option to rerun my saved query prompts at an interval that I specify (e.g., daily, weekly, monthly). I want to be notified when new reports are available.



System overview: Grey Literature Search Assistant

- Automates discovery and curation of grey literature sources from the CDA-AMC Rapid Review checklist + Google
- Aims to replace a manual 2-4 hour search process with a sub-10 minute workflow
- Uses Microsoft, open-source, Bright Data, and Anthropic components

Grey Literature Search Assistant (Beta)

Step 1: Enter research question

Optional — skip to Step 3 if you already have keywords. Used for AI curation scoring and keyword suggestions.

Single question Multiple questions

RESEARCH QUESTION

Step 2: Generate keywords

Optional — skip to Step 3.

> [★ Suggest keywords \(optional\)](#)

Step 3: Select source and build query

SCOPE

All sources

Google only (general search)

Includes Core HTA plus all additional site groups.

QUERY MODE

Simple

Advanced Boolean (Two concepts)

TOPIC / KEYWORDS

Step 4: Run discovery

> [Run discovery](#)

Step 5: Review results



Columns ▾

+ Add result



Technology stack: Microsoft components

Component	Description	Purpose
Azure Container Apps	Microsoft managed hosting service	Production deployment environment
Azure Blob Storage	Microsoft object storage	Stores Excel and .ris formatted text file outputs
Azure Entra ID	Access security via single sign-on	Enables secure access for employees



Technology stack: Open-source components

Component	Description	Purpose
Streamlit	Python framework with a React front end and a Flask back end (layers talk via a standard API)	Web User Interface (UI) and Flask Python web server handles search logic, API, running the AI curation layer, and managing the database
PyPDF	Python library for working with PDF documents	Tier 1 deterministic PDF metadata extraction (title, year)
BeautifulSoup	Python library for parsing HTML and XML documents	Tier 2 deterministic HTML parsing (<time>, Dublin Core, JSON-LD structured data)



Technology stack: Bright Data components

Component	Description	Purpose
Bright Data SERP (Search Engine Results Pages)	API provides real-time search results	Site-specific and Google SERP discovery
Bright Data Web Unlocker	API for optimizing search requests	Returning web page metadata for verification



Technology stack: Anthropic components

Component	Description	Purpose
Claude Haiku	LLM designed for efficient handling of simple tasks	Tier 3 AI extraction fallback; Invoked only when deterministic parsing fails
Claude Sonnet	LLM designed for structured analysis and deep reasoning	AI curation and scoring; Invoked only when requested



Cost and performance

Controlled by 3 mechanisms:

1. Deterministic-first extraction avoids Haiku calls for most URLs
2. User-driven selection limits Sonnet AI curation to URLs of interest
3. Prompt caching on the Sonnet system reduces repeated input token costs

Phase / Operation	Estimated Cost / Time
Phase 1: Full 41-site discovery (no Google)	~\$0.025–\$0.040 per search; <60 seconds
Phase 2: Verification per URL (Bright Data fetch)	~\$0.001–\$0.003 per URL
Tier 3 Haiku fallback per URL	~\$0.0001–\$0.0003 (only when deterministic fails)
AI curation, single question, 100 results	~\$0.01–\$0.02 (2–3 batches)
AI curation, 3 questions, 100 results (with caching)	~\$0.02–\$0.04 (system prompt cached batch 2+)
Combined target (search + curation)	~\$0.05 per complete search run



AI Prompts: Simple task

Component	Task	System Prompt
Claude Haiku	Metadata Extraction	<p>You extract bibliographic metadata for a grey literature tracker. Rules: Use ONLY the provided URL and page text. Do NOT invent or guess values. If year cannot be found, set year to "" and explain in notes. Return STRICT JSON only.</p> <p>Required output schema: {"title": "", "year": "", "doc_type": "", "notes": ""} doc_type must be exactly 'PDF' or 'Webpage'. No other values accepted.</p> <p>User payload includes: topic, site name/number, search_query, url, SERP title hint and snippet, and up to 12,000 chars of page_text.</p>



AI Prompts: Complex task

Component	Task	System Prompt
Claude Sonnet	Evaluates SERP results against a research question, assigns a numeric score and brief reasoning to each result	<p>System Prompt for Simple Curator (Slice 1)</p> <p>Situation: You are conducting grey literature searches as part of a health technology assessment (HTA) process. Grey literature includes reports, guidelines, policy documents, and other materials not published through traditional academic channels but essential for comprehensive evidence synthesis in healthcare decision-making.</p> <p>Task: Evaluate search results that have been retrieved from a pre-specified list of trustworthy grey literature sources. Your job is to assess each result based on its title, description (snippet), and source to determine relevance to the supplied research question.</p> <p>You will receive:</p> <ol style="list-style-type: none"> 1. A research question describing what the user is looking for 2. A list of search results with title, URL, snippet, and source site <p>For each result, provide a relevance score (0-10) and brief reasoning that explains your assessment. [...]</p> <p><i>Prompt continues with descriptions of objective, knowledge & expertise, scoring rubric, source hierarchy, assessment approach (8 steps), output format.</i></p>



Core design principles

- **Tiered extraction**
 - Deterministic methods (PDF metadata, HTML parsing) are always attempted before falling back to AI
- **Two-phase workflow**
 - Fast SERP discovery first; users select verification only for sources they choose; AI curation is optional
- **Search-engine first (no web crawling)**
 - All URLs come from SERP results only
- **Auditability**
 - Every result is traceable to its source URL, the query that found it, and the extraction method used



Testing plans

- Measure viability, accuracy, efficiency, user experience (same domains)
- Make recommendations and refine the tool based on our findings
- Add another grey literature checklist to the app
- Repeat testing

Protocol for Evaluating App-Assisted Grey Literature Search Methods

Research Information Services

Jan 2026

Overview: For phase 3 of the project, Anticipating Artificial Intelligence (AI) for Information Retrieval, Research Information Services (RIS) will evaluate the performance of a purpose-built AI app, the Grey Literature Search Assistant (GLSA).

Our objective is to determine if app-assisted information retrieval with the GLSA has comparable accuracy and greater efficiency than usual practice information retrieval methods for grey literature.

Background

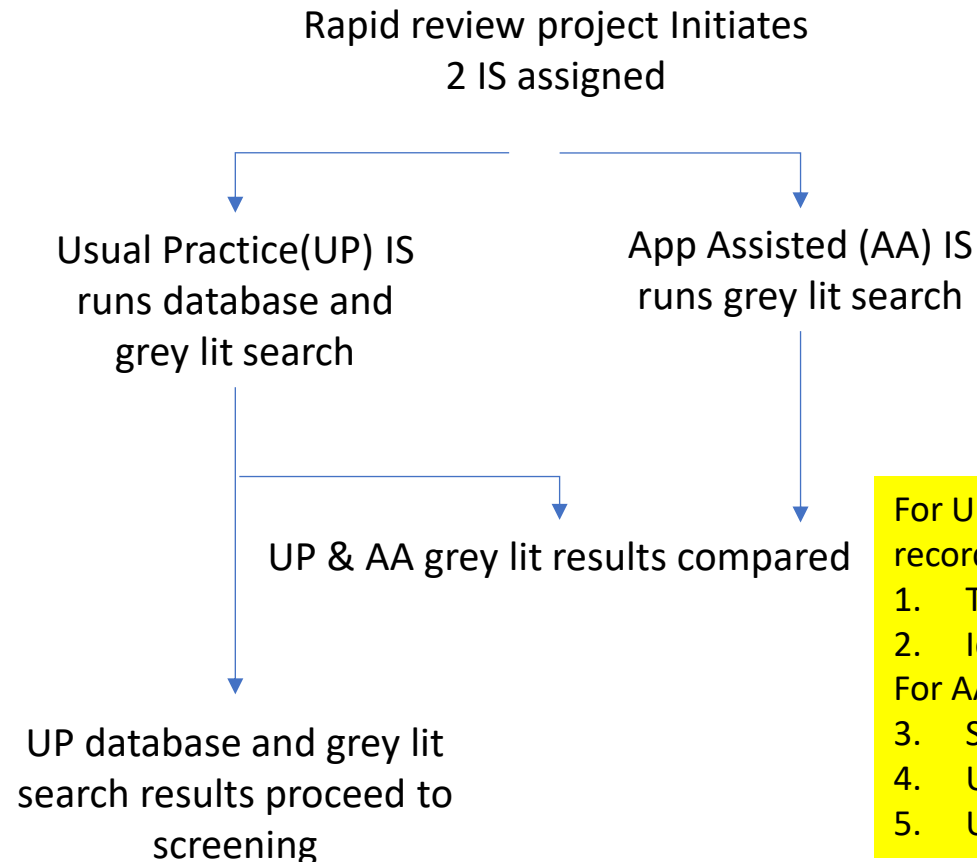
Grey literature is a category of evidence that includes organizational reports, working papers, and government documents.¹ It is referenced or analyzed in all CDA-AMC reports.

Usual Practice Methods to Retrieve Grey Literature: Because grey literature is produced outside of commercial publishing channels¹, CDA-AMC retrieval methods involve Information Specialists (ISs) searching individual websites of producing organizations. This manual search approach for grey literature typically takes much longer than searching for academic articles via scholarly databases.

To ensure that grey literature search approaches are transparent and reproducible, RIS maintains a database of checklists: <https://cda-amc-checklist.andornot.com/>. ISs use

Protocol overview

Prospective design (no risk approach)



We aim to compare UP and AA grey literature searching for 6 projects (3 simple and 3 complex)

For UP and AA grey literature searches, we will record:

1. Time in hours to search
2. Identified search results

For AA searches, we will also calculate or estimate:

3. Sensitivity
4. Unique contributions
5. User experience



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

Vision for AI in HTA information retrieval:

Empowered Information Specialists lead the development of responsible AI systems that integrate with existing HTA processes and target repetitive tasks for large efficiency gains



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

Thank you

Project team members: Sharon Bailey, Robyn Butcher, Caitlyn Ford, David Kaunelis, Hannah Loshak, Danielle MacDougall, Quenby Mahood, Sarah McGill, Eric Morenz, Matthew Morris, Melissa Severn, Melissa Walter

Acknowledgements: Dagmara Chojecki, Amanda Hodgson, Nicole Mittmann, Danielle Rabb



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

Drugs. Health Technologies and Systems. Médicaments, technologies de la santé et systèmes.