State of the Field Report — Literature Searching Methodology & AI

The State of the Field Report synthesizes a curated corpus of 52 papers on literature-searching methodology and AI in evidence synthesis, assembled in November 2026, to answer how a researcher, including a non-academic, should search PubMed well with AI assistance in 2026. It organizes the evidence into a canonical search process (Cooper 2018), a learning ladder of tutorials, quality control via PRESS, measured filter performance and text-mining tradeoffs, the librarian infrastructure, and AI screening, extraction, and reporting standards, closing with an integrated working model, open problems, and a practical checklist.

Table of contents

1. Purpose and method 4
2. Executive summary — ten findings 5
3. The canonical search process — the methodological backbone 6
   3.1 The process model 6
   3.2 The learning ladder — three tutorials, three depths 6
   3.3 The engineering history 7
4. Quality control, filters, and measured performance 7
   4.1 PRESS — the quality instrument 7
   4.2 Filters — the validation evidence 7
   4.3 Strategy-building tools — the SWAR evidence 8
   4.4 Supplementary searching 8
5. The human infrastructure 8
6. The AI wave 10
   6.2 Screening and extraction — the empirical core 10
   6.3 Field maps and orientation 11
7. Risks, integrity, and reporting standards 11
8. Synthesis — an integrated model for PubMed + AI searching (2026) 12
9. Tensions and open problems 13
10. Gaps and research agenda 13
11. Practical checklist (for a non-academic researcher, 2026) 14
12. Limitations of this report 14
13. Addendum — supplementary sources 14
   14. Deep dives — paper-by-paper analysis 15
   14.0 Extraction artefact ledger (consolidated from the six reading agents) 15
   14.1 The guides — plain-language search instruction 16
   Watson (2020): How to undertake a literature search: a step-by-step guide — A 10-step practical literature-search guide for nurses written by an RCN search specialist 16
   Blümle (2019): Systematic literature search in PubMed: a short introduction — German CME tutorial walking clinicians step-by-step through a systematic PubMed search with a live worked example 18
   Motschall (2005): Searching the MEDLINE literature database through PubMed: a short g… — English-language “short guide” demonstrating MeSH-based PubMed searching on a GERD/cancer question 20

Aromataris (2014): Constructing a search strategy and searching for evidence: a guide — Joanna Briggs Institute tutorial on constructing a PICO-anchored, reproducible search strategy for systematic reviews 22
Elston (2019): Getting the most out of online literature searches — tips for advan… — One-page JAAD editor's letter of quick advanced-search tips for clinicians 24
Muin (2006): PubMed interact: an improved interface for MEDLINE/PubMed searches — NLM technical report on PubMed Interact, an Ajax-based alternative PubMed interface with interactive result manipulation 25
Kim (2022): Towards a unified search: improving PubMed retrieval with full text — NCBI study integrating full-text articles into PubMed ranking via PAV-transformed log-odds scores, using click logs 27
Cooper (2018): Defining the process to literature searching in systematic reviews — Literature review of nine guidance documents distilling eight key stages of literature searching in systematic reviews 30
14.2 Filters, quality control, and strategy-building tools 32
Booth (2016): Searching for qualitative research for inclusion in systematic reviews — Structured methodological review of searching for qualitative research (Systematic Reviews 2016) 32
McGowan (2016): PRESS Peer Review of Electronic Search Strategies: 2015 Guideline S… — PRESS 2015 guideline: peer review of electronic search strategies (J Clin Epidemiol 2016) 34
Escobar Liquitay (2023): Search strategies (filters) to identify systematic reviews in MEDLI… — Cochrane methodology review: filters to identify systematic reviews in MEDLINE and Embase (CDSR 2023, MR000054) 36
Beynon (2013): Search strategies to identify diagnostic accuracy studies — Cochrane methodology review: filters for diagnostic accuracy studies (CDSR 2013, MR000022) 38
[[Li (2019): Search strategies to identify observational studies in MEDLINE 39
Lefebvre (2017): Assessing the performance of methodological search filters — HTA monograph: five literature reviews and a qualitative study on search-filter performance (Health Technol Assess 2017;21(69)) 41
Dullea (2026): Search strategies built with the PubReMiner text-mining tool vs cur… — SWAR: PubReMiner text-mining strategies vs conventional searching (Cochrane Evidence Synthesis and Methods 2026) 43
Paynter (2021): Evidence synthesis search strategies with and without text-mining t… — Prospective comparison of search strategies with and without text-mining tools (J Clin Epidemiol 2021) 45
14.3 Supplementary searching, reporting, and knowledge translation 47
Briscoe (2020): Conduct and reporting of citation searching in Cochrane reviews — Cross-sectional study of how backward and forward citation searching is conducted and reported in Cochrane systematic reviews (Res Synth Methods 2020) 47
Luo (2025): GAMER Statement: reporting guideline for generative AI tools in med… — GAMER Statement: international consensus reporting guideline for generative AI tools in medical research (BMJ EBM 2025) 49
Pollock (2022): Methodological quality, guidance, and tools in scoping reviews (pro… — Scoping review protocol on methodological quality, guidance, and appraisal tools for scoping reviews (JBI Evid Synth 2022) 51

Godin (2015): Applying systematic review search methods to the grey literature: a… — Case study applying systematic review search methods to the grey literature: Canadian school breakfast program guidelines (Systematic Reviews 2015) 52
Haby (2016): Best methodologies for rapid reviews for evidence-informed decision… — Rapid review of the best methodologies for rapid reviews in evidence-informed health decision making (Health Res Policy Syst 2016) 54
Grimshaw (2012): Knowledge translation of research findings — Knowledge translation of research findings: five-question framework and evidence on strategies to close the evidence-practice gap (Implementation Science 2012) 56
Godino (2023): How to structure Microsoft Excel documents for systematic reviews — Practical guide to structuring Microsoft Excel documents for systematic review management and PRISMA reporting (Nurse Researcher 2023) 58
Laynor (2024): Identifying meta-research with researchers as study subjects (proto… — Registered Report protocol for a scoping review of meta-research with researchers as study subjects, with a novel PubMed search strategy (PLoS ONE 2024) 60
14.4 The AI core — automation of screening and extraction 62
Lund (2024): ChatGPT in medical libraries: possibilities and future directions — Integrative review: ChatGPT's possibilities in medical libraries (Health Information & Libraries Journal 2024) 62
Li (2025): Enhancing systematic literature reviews with generative AI — Five-module GPT-4 system for systematic literature reviews in HTA (JAMIA 2025) 64
Khan (2025): Collaborative LLMs for automated data extraction in living systemat… — Collaborative LLMs (GPT-4-turbo + Claude-3-Opus) for data extraction in living systematic reviews (medRxiv preprint, posted September 23, 2024) 66
Pilz (2024): Semi-automated title-abstract screening with NLP and machine learning — Semi-automated title-abstract screening with NLP and machine learning (Systematic Reviews 2024) 67
Mahboubi (2025): SciScribe: automating and contextualizing literature reviews — SciScribe: literature-review curation platform in cardiac surgery (J Thorac Cardiovasc Surg 2025) . . 69
Wang (2020): Error rates of human reviewers during abstract screening — Human error rates during abstract screening in systematic reviews (PLoS ONE 2020) 71
Gallo (2025): Best practices in LLM research: repeat prompting — Best practices in LLM research: repeat prompting and correlated outputs (JAMIA 2025) 73
Yarar (2026): Concerns of using LLMs in health care research: umbrella review — Umbrella review: concerns of using LLMs in health care research and practice (JMIR 2026) 74
14.5 Field maps and AI literacy 76
Su (2025): LLMs in medical diagnostics: scoping review with bibliometric analysis — Scoping review with bibliometric analysis of LLMs in medical diagnostics (J Med Internet Res 2025;27:e72062) 76
Ma (2026): AI in implementation research: applications and recommendations — Scoping review: AI in implementation research (Transl Behav Med 2026;16:ibag070) 78
Garry (2026): Exploring variation in research priorities generated by AI tools — Variation in research priorities generated by five AI tools (J Glob Health 2026;16:04037) 80
Demir (2025): AI-driven abstract generating under the PRISMA-A framework — LLM-generated PRISMA-A abstracts: ChatGPT-4o vs Gemini Pro (BMC Oral Health 2025) . . 81

Begagic (2025): The role of reviewers in the era of systematic reviews: a practical… — Practical guide: the role of reviewers of SRs and meta-analyses, with a 23-item checklist (Biomol Biomed 2026;26(1):40–50) 84
Johnson (2022) — The Systematic Review Toolbox 2022 relaunch (Syst Rev 2022; commentary) 86
Verma (2024) — LERRN: free citation database for resource reviews (JMLA 2024;112(2), Virtual Project) 87
14.6 The librarian craft — roles, competencies, and training 89
Lizarondo (2025): Data extraction and transformation in mixed methods systematic reviews — JBI guidance on data extraction and qualitization in convergent integrated mixed-methods systematic reviews (JBI Evid Synth 2025;23(3):429–440) 89
Spencer (2018): Roles for librarians in systematic reviews: a scoping review — Scoping review identifying 18 roles for librarians in systematic reviews (JMLA 2018;106(1):46–56) .. 91
Townsend (2017): A competency framework for librarians involved in systematic reviews — Competency framework for librarians involved in systematic reviews (JMLA 2017;105(3):268–275) 93
Farris (2024): Curriculum framework for systematic review competencies (librarians) — Case report: building the "Essential Searching Skills" curriculum for MLA's Systematic Review Services Specialization (JMLA 2024) 95
Boden (2018): Learning while doing: MLA Systematic Review Project evaluation — Program evaluation of the MLA Systematic Review Project: learning by doing at scale (JMLA 2018;106(2)) 97
Westrick (2021): Medical librarians' practices in locating clinical trials — Survey: medical librarians' knowledge and practices in locating clinical trials for systematic reviews (JMLA 2021) 99
Minuti (2018): Teaching searching skills with a flipped classroom — Flipped-classroom teaching of PubMed searching skills to medical students (Med Ref Serv Q 2018;37(2):119–131) 101
Parker (2022): Literature reviews: tips from knowledge synthesis librarians — Literature reviews: tips from knowledge synthesis librarians; the PIECES framework (JGME 2022) .. 103
15. Addendum — supplementary sources 105
16. Addendum II — process, quick-reference, and appraisal sources 105

1. Purpose and method

This report synthesizes a curated corpus of **48 papers** on literature-searching methodology and the use of artificial intelligence in evidence synthesis, assembled 15–17 November 2026. It answers one practical question: **how should a researcher — including a non-academic — search PubMed well, with AI assistance, in 2026?**

Accountability. Every claim is anchored to a source page in this workspace; all 48 papers are @-mentioned below, and each has its own page with a structured summary, full citation, DOI, and transcribed full text. Corpus construction was itself a documented pipeline: a 273-entry Zotero bibliography was verified per-DOI against PubMed (each article checked individually; abstracts transcribed), then screened in two passes (relevance tiers, then practicality tiers), then curated by hand (clinical-topic LLM applications removed; quality-assessment and reporting-standards papers retained). Vocabulary for the underlying searches was farmed from three seed PMID sets, verified against NLM MeSH records, and the search strings were tuned against 420 empirically

measured PubMed combination counts (documented in the workspace search archive). Full texts were transcribed from Zotero PDFs (pdftotext; Windows OCR for the one scanned PDF) so that every claim below can be traced to source text on the article pages.

Reading path. Section 2 is an executive summary. Sections 3–6 map the corpus thematically; section 7 distills an integrated working model; section 8 names the structural tensions; section 9 gives gaps and a research agenda; section 10 is a practical checklist; section 11 declares limitations; section 12 links supplementary sources.

2. Executive summary — ten findings

1. **The search process is standardized in consensus, not in statute.** Nine guidance documents share one tacit model with eight stages Cooper (2018): Defining the process to literature searching in systematic reviews — but the least standardized stage (supplementary searching) is where much retrieval innovation happens.
2. **Human screening error is ~11%.** Paired expert reviewers err in **10.76%** of abstract-screening decisions Wang (2020): Error rates of human reviewers during abstract screening — so automation tools do not need to be perfect to be useful; they need to be human-comparable.
3. **LLM pipelines now reach human-comparable performance on SLR subtasks.** GPT-4-based abstract screening: sensitivity ~90%, $\varkappa$ 0.71 Li (2025): Enhancing systematic literature reviews with generative AI; two-LLM concordance for data extraction: 94–99% accuracy on concordant responses, ~0% hallucinations Khan (2025): Collaborative LLMs for automated data extraction in living systemat….
4. **Text-mining strategies are faster but not yet a replacement.** PubReMiner strategies: faster to build, significantly lower sensitivity across 12 reviews Dullea (2026): Search strategies built with the PubReMiner text-mining tool vs cur…; same direction in the US setting Paynter (2021): Evidence synthesis search strategies with and without text-mining t….
5. **Methodological filters degrade under validation.** DTA filter sensitivity fell from median 86% (development) to 73% (evaluation) Beynon (2013): Search strategies to identify diagnostic accuracy studies; SR-retrieval filters derive mostly from pre-PRISMA review sets Escobar Liquitay (2023): Search strategies (filters) to identify systematic reviews in MEDLI….
6. **PRESS remains the quality instrument.** Six checklist elements, unchanged since 2008, confirmed by 117 experts McGowan (2016): PRESS Peer Review of Electronic Search Strategies: 2015 Guideline S….
7. **LLM evaluation methodology is itself immature.** Repeat-prompting correlation (ICC 0.69) inflated a published study's effective sample size >100-fold and reversed its conclusions Gallo (2025): Best practices in LLM research: repeat prompting.

Reporting standards exist but are barely endorsed. GAMER's nine-item checklist (Luo) defines the disclosure standard for generative-AI use, but journal endorsement of AI-specific reporting guidelines remains rare — single-digit endorsement rates across surveyed specialties.

The librarian multiplier is evidence-backed, with 18 documented roles (Spencer 2018), six competencies (Townsend 2017), and curriculum implementations (Farris 2024).

AI literacy is now a baseline skill. The AI-methods papers in this corpus cluster around image analysis, NLP/LLMs, and drug discovery — the same triad that organizes the field's literacy guides.

3. The canonical search process — the methodological backbone

3.1 The process model

Cooper (2018): Defining the process to literature searching in systematic reviews audited the literature-searching chapters of nine guidance documents (Cochrane Handbook, CRD, Collaboration for Environmental Evidence, JBI Reviewers' Manual, IQWiG, the social-sciences practical guide, EUnetHTA, Campbell Handbook, NICE manual) and extracted methodological stages until consensus stabilized into **eight key stages**: (1) *who* should search — eight documents affirm information specialists/librarians; the supporting literature includes the finding that librarian co-authorship correlates with higher search-quality scores; (2) aims — "thorough, comprehensive, transparent, reproducible," linked by three documents to bias minimization; (3) preparation; (4) search strategy; (5) database searching — consistently first; (6) **supplementary searching — the least standardized stage**, where guidance is diverse (citation chasing, grey literature, trial registers, web, experts); (7) reference management; (8) reporting. Unique guidance appeared in single documents only: *when to stop searching* (one document), quality assurance/peer review (two), text mining (four). The paper explicitly flags that the conventional approach's fitness for new review types (rapid, scoping) is untested — a gap this corpus's AI papers now begin to fill.
Booth (2016): Searching for qualitative research for inclusion in systematic reviews complicates the orthodoxy productively: for qualitative evidence syntheses, comprehensive sampling gives way to **purposive sampling**, sources extend beyond medical databases (theses, books, grey literature), and citation searching can be a *primary* method. Its 7S framework (structured questions, sampling, sources, procedures, strategies/filters, supplementary strategies, standards) organized 113 included items and documented the retrieval challenges specific to qualitative work: non-standard terminology, non-explicit titles, weak abstracts, inadequate indexing.

3.2 The learning ladder — three tutorials, three depths

Watson (2020): How to undertake a literature search: a step-by-step guide is the entry level: a 10-step narrative (question → concepts via PEO/PICOS/SPIDER → synonyms → databases → Boolean combination → limits → run/review → adapt → save → obtain) with a worked example, a glossary, and the RCN planning grid. Blümle (2019): Systematic literature search in PubMed: a short introduction is the intermediate CME level: a fully commented PubMed search history for an RCT question (anti-VEGF in retinopathy of prematurity), teaching PubMed's **database components** (MEDLINE ~90% of ~27M records; In-Process; Publisher; PubMed-not-MEDLINE), **automatic term mapping**, the precision/recall formalism (recall needs a gold standard; precision is computable), MeSH database navigation (explode, subheadings, major topic), filters, and the AND medline[sb] restriction trick. Motschall (2005): Searching the MEDLINE literature database through PubMed: a short g… is the oldest and most granular: a GERD→esophageal-cancer etiology question walked from an 807-hit naive search to 113 focused citations via MeSH database navigation (entry terms, *See also* Heartburn, subheading restriction to *Esophageal Neoplasms/etiology*), history combining (#5 AND #8 → 336), auto-explode design filters (Cohort Studies → Longitudinal/Follow-Up/Prospective; Risk → Logistic Models/Risk Factors), Limits, text-word tags ([tw], [tiab], [ti], truncation), NOT medline[sb], and Clinical Queries — ending in a six-point checklist. Read in sequence, they form a curriculum: concepts → PubMed mechanics → measured strategy.

3.3 The engineering history

Muin (2006): PubMed interact: an improved interface for MEDLINE/PubMed searches (2006) predates the modern AI era but anticipates it: seven slider bars for limits, results-count preview without submitting, in-page related articles, citation deletion, and an "information box" educationally displaying mapped MeSH terms — precisely the affordances later mainstreamed. Kim (2022): Towards a unified search: improving PubMed retrieval with full text explains why "just also search the full text" is a hard problem: naive body-text pooling *degrades* precision (full texts discuss multiple subtopics and mention things in passing); their solution converts BM25 section scores to log-odds via PAV regression and sums the top-3 section scores, yielding ~7% relative improvement over abstracts alone on PubMed click data. The methodological parallel to the LLM papers is direct: heterogeneity of evidence sources requires probabilistic normalization, not naive pooling.

4. Quality control, filters, and measured performance

4.1 PRESS — the quality instrument

McGowan (2016): PRESS Peer Review of Electronic Search Strategies: 2015 Guideline S… built the guideline from an SR + survey of 117 experts + consensus forum. The **six retained checklist elements**, each with sub-questions: (1) *translation of the research question* — does the strategy match the PICO? too many/few concepts? (2) *Boolean and proximity operators* — nesting logic, NOT-exclusion risk, proximity width; (3) *subject headings* — relevance, missing headings, explosion, major-topic ("starring") justification, floating subheadings; (4) *text words* — spelling variants, synonyms and antonyms, truncation placement, acronyms, field selection; (5) *spelling, syntax, line numbers* — orphan lines, wrong-interface truncation symbols; (6) *limits and filters* — relevance, epidemiological bias, filter sourcing. The dropped seventh element (translation to other databases) was removed as searcher's discretion. Implementation standards: peer review the primary search; one review usually suffices; document it; acknowledge the reviewer; five-working-day turnaround. The Teaching Table 4 case is worth replicating in any training: a plausible acetaminophen/migraine search with a seeded error in every element.

4.2 Filters — the validation evidence

Four corpus papers measure methodological filters, and the consistent finding is **optimism decay**:

- Beynon (2013): Search strategies to identify diagnostic accuracy studies: 19 studies, 70 filters (57 MEDLINE, 13 Embase). Where development and evaluation could be compared (13 MEDLINE filters), sensitivity fell from a median **86%** at development to **73%** at evaluation; precision from 9.5% to 4%. No filter met the a-priori "acceptable" bar (sensitivity >90% with precision >10%).
- Escobar Liquitay (2023): Search strategies (filters) to identify systematic reviews in MEDLI…: 8 MEDLINE-filter studies, 3 Embase; three MEDLINE filters exceeded 90% sensitivity, one had >90% specificity; gold standards dated (some pre-PRISMA 2009), small (70–387 records) except Wilczynski's (27,769–49,028); heterogeneous reporting.
- Li (2019): Search strategies to identify observational studies in MEDLINE and …: only 2 eligible studies, 18 filters; the validated specificity-terms filters reached 99.5–100% sensitivity with 16.7–21.1% precision (surgical-intervention gold standard — limited generalizability); comparative-NRS filters performed far worse (48–100% sensitivity, 0.09–4.47% precision).

Term-level data are practical gold: *compara**, *postoperat*/preoperat*, Comparative Studies[MeSH] most sensitive; cohort.tw, nonrandom*.tw most precise.

- Lefebvre (2017): Assessing the performance of methodological search filters (five reviews + qualitative study, searches 2010/11): Review A (23 studies) — internal gold standards mostly hand-searched; validation mostly internal; Review B (18) — comparisons report best/precision-optimal/balanced strategies; Review C (47) — imports DTA metrics (spectrum bias, ROC, 2×2 tables, likelihood ratios) into filter appraisal; Review D — **no studies on how searchers choose filters**; Review E — clinicians often misunderstand test-performance parameters. The 12 interviews found filters used to *reduce record counts, add focus, isolate study types*; the Cochrane RCT and McMaster Hedges filters most cited; the **ISSG Search Filters Resource** the discovery hub; adaptation and reporting practices non-uniform.

4.3 Strategy-building tools — the SWAR evidence

Paynter (2021): Evidence synthesis search strategies with and without text-mining t… (AHRQ/VASP/Cochrane; 7 reviews, simple vs complex): two librarians per review built MEDLINE strategies simultaneously — usual practice vs a free TMT toolkit (AntConc, PubReMiner, MeSH on Demand, Yale MeSH Analyzer). UP searches: sensitivity 92% (85–99), NNR 83, 12 h build. TMT searches: 84.9% (74.4–95.4), NNR 90, **5 h**. Complex topics: TMT searches found unique eligible citations UP missed. Dullea (2026): Search strategies built with the PubReMiner text-mining tool vs cur… (HIQA Ireland; 12 consecutive reviews, 2023–2025, prospectively registered SWAR): PubReMiner-only strategies vs PRESS-reviewed conventional — sensitivity significantly lower overall (Friedman p = 0.0065) though higher in one review (83.87% vs 58.06%); no precision difference; faster; inconsistent across librarians. Both studies conclude: **adjunct, not replacement** — and both call for tool-comparison best-practice research.

4.4 Supplementary searching

Briscoe (2020): Conduct and reporting of citation searching in Cochrane reviews: of 215 reviews, 172 (80%) reported backward citation searching, 18 (8%) forward; Web of Science most used; backward typically from included studies, forward from selected key studies (one-third). MECIR mandated backward (C30) but not forward. Reporting met MECIR minima but lacked index names/dates for full reproducibility. Citation searching matters most where keywords fail — qualitative synthesis, historical terminology — where it can be primary. Godin (2015): Applying systematic review search methods to the grey literature: a… supplies the grey-literature template: four complementary strategies (grey-literature databases; **customized Google search engines**; targeted websites; **expert consultation**), screening on abstract/executive summary/table of contents when abstracts are absent, 302 candidates → 15 included, PRISMA-style documentation throughout. Haby (2016): Best methodologies for rapid reviews for evidence-informed decision… (a rapid review of rapid reviews, PROSPERO-registered) found **no agreed definition or methodology** for rapid reviews; shortcuts in use (narrower question, fewer sources, English-only, no grey literature, single reviewer) with little empirical evidence of impact on conclusions; one low-risk RCT suggested rapid reviews improve clarity/accessibility for decision makers.

5. The human infrastructure

Roles. Spencer (2018): Roles for librarians in systematic reviews: a scoping review catalogued **18 roles** from 310 documents — searching, source selection, teaching, plus planning, question

formulation, protocol development, citation management, de-duplication, screening support, bias assessment, data synthesis, document supply, peer review of searches, methodology writing, co-authorship.

Competencies. Townsend (2017): A competency framework for librarians involved in systematic reviews (University of Michigan informationists) defined **six competencies** — systematic review foundations; process management & communication; research methodology; comprehensive searching; data management; reporting — with indicators assessed on a Miller's-Pyramid adaptation ("knows → knows how → does → teaches"), deliberately avoiding novice/expert labels because librarian roles vary by project.

Curriculum. Farris (2024): Curriculum framework for systematic review competencies (librarians) implemented the competencies as MLA's SRSS Level I course *Essential Searching Skills for Librarians on Systematic Review Teams*: four modules (consultation & pre-searching; searching — question frameworks, concept nesting, documentation, PubMed as teaching platform; PRESS peer review; PRISMA-S reporting), ~30 participants per cohort, **eleven cohorts 2022–2024**, OSF-hosted open materials.

Scale. Boden (2018): Learning while doing: program evaluation of the Medical Library Association Systematic Review Project evaluated ~200 volunteers in 15 teams answering Delphi-ranked profession questions: by Jan 2017 three teams published (one won the Eliot Prize), eight progressing, four stalled; **84% would repeat it**; themes: learning, networking, teamwork, outcomes, barriers (attrition, leadership turnover).

Practice surveys. Westrick (2021): Medical librarians' practices in locating clinical trials (120 respondents): half had been asked to limit searches to trials; trial *types* specified only half the time; **no consistent locating method** — hedges 55%, own strings ~35%, database filters 27%, 23% don't search trials separately; 45% self-rated expert searchers; and they wanted a definitive study-design methodology resource. Parker (2022): Literature reviews: tips from knowledge synthesis librarians packages the workflow via **PIECES** (Planning, Identifying, Evaluating, Collecting/combining, Explaining, Summarizing) with a phase→tools→team table (communication, project management, review software, PRISMA-flow generators) and pro tips (choose the venue early; don't forget ERIC for education topics).

Discipline case (dental reviews): dental SRs are numerous and often weak — 95% of prosthodontics search strategies were \u201cnot systematic\u201d, and periodontal practice guidelines were built on low-quality reviews. Librarian roles stayed traditional (planning and retrieval for all respondents); the most frequent challenges lay with the review teams, not the librarians — more librarian involvement as methodologists could improve quality.

Infrastructure. Johnson (2022): The Systematic Review Toolbox: tools to support evidence synthesis relaunched the community catalogue (systematicreviewtools.com) with **235 tools and 112 guidance documents** coded by review family and stage; gaps: stakeholder engagement (2 tools), qualitative-review guidance (19). Verma (2024): LERRN: a free citation database for resource reviews demonstrates the low-cost community-database model (Zotero RSS → Librarika ILS; ~920 resource-review citations) — the same architecture this workspace uses for its own corpus.

Retrieval frontier. Laynor (2024): Identifying meta-research with researchers as study subjects (proto… poses the hardest problem: when the study population is named by a common noun ("researcher" — tens of millions of hits), when the field is young and unindexed ("meta-research" lacks MeSH, and is confused with meta-analysis), standard search construction fails; the protocol publishes its preliminary-search test data and restricts scope to health disciplines as a feasibility decision.

The appraisal side. Begagic (2025): The role of reviewers in the era of systematic reviews: a practical… equips peer reviewers of SRMAs with a **23-item checklist in four sections**: A. Initial evaluation (justification vs existing reviews; PRISMA/MOOSE/PRIOR compliance); B. Methodology (PROSPERO registration verified against the protocol; PICOS; reproducible search — queries, databases, dates, minimum two databases, Embase+PubMed+Scholar/WoS trio; PRISMA flow diagram consistency; dual independent extraction; RoB 2/ROBINS-I; pre-specified analysis); C. Results (appropriate effect measures; I^2 interpretation with the 25/50/75% conventions and Q-test power caveats; pre-specified vs data-driven subgroups with interaction testing; forest-plot reading; funnel plot/Egger/trim-and-fill caveats at small n; sensitivity analyses; GRADE summary-of-findings); D. Emerging issues and ethics — including **D3 checking included studies for paper-mill origin, D5 verifying "hallucinated" references against original sources, and D7 AI-disclosure compliance**. The guidance explicitly warns against vote-counting narrative syntheses, unplanned subgroup fishing, and meta-regression at small n.

6. The AI wave

6.2 Screening and extraction — the empirical core

Wang (2020): Error rates of human reviewers during abstract screening established the benchmark: 25 AHRQ EPC reviews, 139,467 citations, 329,332 decisions, 86 reviewers, DistillerSR-tracked; **total error rate 10.76% (95% CI 7.43–14.09)**, varying 5.76–21.11% by clinical area and question type; abstract-inclusion rate 18.07%, final inclusion 5.48%. Implication stated plainly: automation tools achieving human-comparable error rates may be adequate as one reviewer paired with a human.

Pilz (2024): Semi-automated title-abstract screening with NLP and machine learning is the transparent classical-ML response: full NLP preprocessing (lowercasing, lemmatization, Porter stemming, stop-word removal) into a unigram matrix; train/test split with human include/exclude labels; multiple classifiers; **iterative training-set growth** to find the minimum size after which one human can stop; applied to two real reviews with meta-analysis (INTRISSI colorectal prophylaxis; perforated peptic ulcer). Positioned explicitly between proof-of-concept papers and inflexible commercial tools (Rayyan, EPPI Reviewer, RobotReviewer, ASReview, SWIFT, DistillerSR, Colandr — active-learning designs that interrupt reviewer workflow).

Li (2025): Enhancing systematic literature reviews with generative AI is the zero-shot LLM pipeline: five modules (search → PICOs protocol → GPT-4 screening with rationale capture → targeted extraction → summarization); **human-in-the-loop**: users review ≥10% of screened abstracts, disagreements refine the PICOs criteria iteratively; evaluation on RRMM (3,713 abstracts) and advanced melanoma (1,720): screening sensitivity ~90%, F1 82, accuracy 89%, κ **0.71**; exclusion-rationale accuracy 97%/84%; extraction F1 93.

Khan (2025): Collaborative LLMs for automated data extraction in living systemat… is the most methodologically detailed: 10 trials, 22 publications, 23 variables from a living SR on metastatic castration-sensitive prostate cancer; PyMuPDF extraction with nonessential-section removal; 1,000-token chunk overlap; **table/figure pages rendered as 500-dpi images**; GPT-4-turbo and Claude-3-Opus at temperature 0; exact-match concordance for numerics; discordant responses cross-critiqued (each model sees the other's answer). Results: development set 96% concordant, accuracy 0.99; test set 87% concordant, accuracy **0.94 for concordant responses vs 0.41–0.50 for discordant**; cross-critique converted 51% of discordances, raising their accuracy to 0.76; **hallucinations 0.25% among concordant responses** (2.3–2.8% for single models); concordant responses significantly outperformed both single models. The transferable

design principle: **agreement between two LLMs is itself a usable accuracy signal** — a statistical second reviewer.

Mahboubi (2025): SciScribe: automating and contextualizing literature reviews (Cleveland Clinic + IBM Discovery Accelerator) demonstrates the curation platform: PMC ingestion (5M+ documents) plus PDF pipelines with table decomposition; **contextualization by linking papers to ClinicalTrials.gov (NCT numbers) and CMS Open Payments (author-industry payments)** — surfacing conflict-of-interest signals inline; keyword search plus **RAG-based natural-language Q&A** over personal collections; demonstrated on a structural-valve-deterioration review question in cardiac surgery.

Demir (2025): AI-driven abstract generating under the PRISMA-A framework tests the *output* side: ChatGPT-4o vs Gemini Pro generating PRISMA-A-compliant structured abstracts from 162 orthodontic SRs/MAs with a tailored prompt; ChatGPT-4o won (OQS 21.67 vs 21.00, p<.001), largest gaps in "Included Studies" and "Synthesis of Results".

6.3 Field maps and orientation

: 95 studies (JBI + PRISMA-ScR; WoS/PubMed/Embase/IEEE/ACM 2022–2025); VOSviewer three-cluster mapping (assessment; specialty applications; ethics); GPT-4 in 74% of studies; high reported accuracy in radiology/psychiatry/neurology; documented biases in race, gender, cost predictions; hallucination and regulatory fragmentation as barriers; future agenda: explainable chain-of-reasoning frameworks, adversarial testing, prospective LLM-vs-workflow trials, multimodal integration.: dual-database design (WoSCC 665 records; PubMed 1,106; strict subsets 532/898) showing growth 66→169→242→(188 partial 2026); nine CiteSpace clusters; emerging terms DeepSeek, clinical reasoning; author-keyword structure stable across definitions ($\rho = 0.973$) — and the explicit methodological lesson that WoSCC and PubMed are complementary, non-equivalent sources.: 29 studies after CASP appraisal; ChatGPT as complementary library support (reference queries, navigation, cataloguing) with documented accuracy and currency limitations.: only 7 studies; AI for implementation monitoring, strategy selection, literature synthesis, dissemination; ethics (privacy, transparency, equity); the gap — AI applications rarely grounded in implementation frameworks.: five tools (DeepSeek, ChatGPT, Claude, Perplexity, Gemini) generating treatment-research priorities for four diseases, compared by SBERT-BioBERT cosine similarity: Gemini most representative; DeepSeek vs Perplexity most divergent (technical vs public-health emphasis), stable across prompt variations. anchors the literacy layer (section 6.1). documents the governance layer (section 7).

7. Risks, integrity, and reporting standards

The measurement problem. Gallo (2025): Best practices in LLM research: repeat prompting reanalyzed an affiliation-bias study (GPT-3.5 judging 30 abstracts × 30 university affiliations × 250 repeat prompts = 225,000 observations): mixed-effects logistic regression with random effects for abstract and affiliation found **ICC 0.69**; the effective sample size per combination fell from 250 to ~1.45 (>100-fold inflation ignored by the original analysis); the original OR 1.04 (P<.001) for top-tier acceptance became OR 1.07 (P=.27); power fell 100%→33%. The recommendation: repeat prompting is good for measuring stability, but analyses must model the clustering (mixed models, cluster-randomized-trial power methods).

The baseline problem. Wang (2020): Error rates of human reviewers during abstract screening: automation is typically benchmarked against human reviewers as "gold standard" — but the gold standard errs at 10.76%. Any evaluation framework (cf. the UK AI Safety Institute's

Inspect framework, referenced by Gallo) should therefore treat human-level error as an acceptable threshold, not perfection.

The reporting layer. GAMER (EQUATOR-registered; Delphi, 51 experts from 26 countries) defines nine reporting items for generative-AI use in medical research — general declaration, tool specifications, prompting techniques, the tool's role, sections touched, verification, privacy, and impact on conclusions — proposed as a minimum standard for journals.

The appraisal layer. Begagic (2025): The role of reviewers in the era of systematic reviews: a practical… operationalizes integrity checking for reviewers: beyond the methods/statistics appraisal (section 5.5 of its checklist), section D targets emerging issues — paper-mill provenance checks (D3), plagiarism screening (D4), **verification of "hallucinated" references against original sources (D5)**, COI declarations (D6), journal-policy compliance including AI disclosure (D7). Context: retraction scandals, paper mills selling authorship, a 72% increase in suspected AI-generated content.

The consumer layer. Standards for online patient health information — 74 unique evaluation items across Eysenbach's seven domains (accuracy, completeness, technical elements, design, usability, accessibility, readability) — apply equally to LLM-generated health content; Yarar's umbrella review aggregates the associated risk literature.

8. Synthesis — an integrated model for PubMed + AI searching (2026)

Layer 1 — Vocabulary before syntax. Every methodological paper presupposes good search terms. The corpus workflow demonstrates the empirical answer: anchor the domain (institutions, key papers), farm MeSH from the field's own literature, verify against NLM records, and measure — the 420-combination matrix in the workspace archive turned "which terms?" into data. This bootstrapping step is *not yet described* in the methodological literature (section 9) — it is this project's methodological contribution.

Layer 2 — Boolean + MeSH is the skeleton. Watson/Blümle/Motschall/Aromataris converge: OR within synonyms, AND across concepts, MeSH + free text together, field codes for precision, truncation for recall, history for documentation. The filter papers bound expectations: filters are aids whose quoted performance shrinks under external validation (Beynon, Escobar, Li, Lefebvre) — never the sole approach.

Layer 3 — Supplementary searching is mandatory and plural. Citation searching (Briscoe: 80% backward, 8% forward, MECIR gap on forward), grey literature via the four-strategy template (Godin), rapid-review shortcuts declared honestly (Haby), experts and registers (Cooper). For qualitative and mixed-methods syntheses, supplementary methods can be primary (Booth), with JBI guidance on extraction and qualitization (Lizarondo).

Layer 4 — Automation assists; humans arbitrate. The error benchmark (Wang 10.76%) frames what "good enough" means. Three automation architectures coexist: classical ML pipelines (Pilz), zero-shot LLM pipelines with human-in-the-loop refinement (Li), and multi-LLM concordance with cross-critique (Khan). Strategy-building tools (PubReMiner) save time but reduce sensitivity (Paynter, Dullea). The consistent design rule across all of them: **the tool proposes; the human disposes** — with Khan's concordance heuristic the most elegant arbitration mechanism (agreement as an accuracy signal, cross-critique as rescue).

Layer 5 — Reporting is the price of admission. . The triad: : PRESS (peer review of the strategy), PRISMA-S (search reporting), GAMER (AI-use reporting) — plus GRADE-style certainty assessment and Begagić's reviewer checklist on the appraisal side. .

: 18 roles, six competencies, working curricula, and survey evidence that even expert searchers want better methodology resources. . And the purpose layer (Grimshaw): : none of this matters unless syntheses — built on good searches — actually reach practice.

9. Tensions and open problems

1. **Comprehensiveness vs timeliness.** Haby shows rapid-review shortcuts lack outcome evidence; Dullea/Paynter show tool-accelerated searching costs sensitivity. No corpus paper resolves when speed is worth the recall — a genuine open decision problem for every project.
2. **Human error as gold standard.** Wang's 10.76% undermines the benchmark logic of every automation evaluation: if humans err at that rate, an automation matching them is "as good as two humans"? The corpus raises but does not settle the question.
3. **Correlated instruments.** Gallo's ICC critique applies to search too: the 13 strings (or N LLM runs) are correlated instruments; coverage claims that treat them as independent are statistically inflated. Multi-instrument estimation (capture–recapture over overlaps) is the missing method.
4. **Tool opacity vs method transparency.** Pilz argues for inspectable pipelines; commercial tools and LLM APIs are opaque; PRESS assumes a human-readable strategy. How to peer-review a prompt or an embedding-based retrieval remains unsolved.
5. **Vocabulary bootstrapping.** The methodological literature starts after terms exist. The affiliation-anchored, first-learning MeSH-farm step used to build this corpus's vocabulary — seed the domain from ~20 institutes' output, harvest MeSH frequencies, then build strings — is a described-but-unpublished procedure that would formalize the cold-start problem (and was flagged in the workspace archive as a candidate SWAR).
6. **Integrity at scale.** Begagić's D-section (paper mills, hallucinated references) plus GAMER's disclosure items address AI-era integrity, but detection tooling and adoption data are absent from the corpus.

10. Gaps and research agenda

- **LLM-assisted search-strategy building** is unevaluated: the corpus covers screening, extraction, and abstract *generation* (Demir) — but no study tests whether an LLM can build a PRESS-passing strategy.
- **Filter revalidation** against post-PRISMA-2020 gold standards (Beynon/Escobar both flag dated reference sets).
- **Corpus-level measurement models**: applying Gallo's clustered-analysis standard to multi-tool retrieval, and capture–recapture estimation of unobserved relevant literature across the 13 strings.
- **The affiliation-anchored vocabulary farm** (this project's Phase 0 proposal) as a registered SWAR: institute-mapped seed corpora, MeSH-frequency profiles, and string validation against anchor-set recall.
- **Non-academic audiences**: the corpus has no plain-language pathway; the practical reading order in this workspace is an implicit first draft.

11. Practical checklist (for a non-academic researcher, 2026)

1. Frame the question with a PICO-family framework (Watson, Aromataris, Parker).
2. Learn the field's vocabulary from its own literature — seed with an institution-anchored set and farm MeSH frequencies before writing any string (Cooper, Spencer; this project's workflow).
3. Build the string: MeSH + free text, OR within concepts, AND across them; truncate; use validated filters only as aids (Blümle, Motschall, Beynon, Escobar, Li).
4. Peer-review the strategy against PRESS — even informally (McGowan, Westrick, Parker).
5. Supplement: backward/forward citation searching (Briscoe), grey literature via a multi-strategy plan (Godin), experts and registers (Cooper).
6. Automate screening/extraction with human arbitration — expect human-level error rates, not perfection (Wang, Pilz, Li, Khan); prefer architectures with concordance checks or human-in-the-loop refinement.

Benchmark any LLM tool with clustered statistics (Gallo); ; disclose use under GAMER (Luo); ;

1. Structure everything in an auditable tool — Godino's Excel method as the floor, Zotero + Capacities as this workspace's implementation — and report under PRISMA-S.
2. Keep a librarian in the loop (Spencer, Townsend) — measurably better searches (Cooper).
3. Remember the purpose: syntheses, not single studies, change practice (Grimshaw); and translate them (Haby, Ma).

12. Limitations of this report

The corpus over-represents English-language, methodologically self-aware literature (screened for practicality and methodological relevance) and deliberately excludes specialty-specific LLM application studies. Under-represented: non-indexed 2026 preprints (though 2025–2026 papers are well represented), non-English evidence, and commercial-tool evaluations. Summaries were written from abstracts and full texts as available; in the underlying screening, six records had no abstract and two corpus papers (Godino, Liu) relied on late-arriving or scanned PDFs (Godino's full text is OCR-derived and carries minor artefacts). PubMed hit counts in the workspace archive were measured on specific dates and drift with database updates. The statistical framing of the corpus as a mixture of retrieval distributions (discussed separately in this workspace) remains an interpretation, not an established model.

13. Addendum — supplementary sources

— the four-strategy grey-literature template behind checklist item 5. — the rapid-review methods evidence. — the JBI extraction/qualitization guidance that AI extraction tools automate. — learning-at-scale evidence (section 5). — the community tool catalogue (section 5, section 8 gap 6). — the 74-item standards compendium (section 7, consumer layer).

Six corpus papers support claims in sections 3–9; they are linked here for completeness.

Aromataris (2014): Constructing a search strategy and searching for evidence: a guide — the systematic-review-grade tutorial behind section 3.2′s learning ladder: the PICO logic grid, pilot searches to harvest keywords and MeSH terms from retrieved articles, field codes ([tw], [tiab], [mh]), truncation, and the worked animal-assisted-therapy search string built column by column.

Elston (2019): Getting the most out of online literature searches — tips for advan… — the one-page quick reference (Boolean operators, field tags, truncation, Limits, Clinical Queries) cited as the cheat-sheet layer of the learning ladder.

Pollock (2022): Methodological quality, guidance, and tools in scoping reviews (pro… — the registered protocol for mapping appraisal tools for scoping reviews (section 4′s reporting/quality apparatus; its completed companion is Gebrye et al. 2023, cited in the Pollock page's text).

Grimshaw (2012): Knowledge translation of research findings — the KT framing paper behind the synthesis: the five questions (what/to whom/by whom/how/with what effect), the argument that the unit of translation should be the up-to-date systematic review, and the barrier/facilitator logic for choosing KT strategies.

Minuti (2018): Teaching searching skills with a flipped classroom — the flipped-classroom education case in section 5 (online tutorials + gamified reinforcement under one-hour constraints).

Godino (2023): How to structure Microsoft Excel documents for systematic reviews — the free six-step Excel audit-trail method (OCR-transcribed on its page) cited in checklist item 8 as the no-cost floor for auditable review management.

14. Deep dives — paper-by-paper analysis

The deep dives were produced by six parallel reading agents, each assigned eight papers' full texts (~2.4 MB). Each agent read its papers end-to-end and produced a structured analysis: what the paper is, methods in detail, key numbers, conclusions, limitations, practical implications for a non-academic PubMed+AI user, verbatim quotes, and within-batch connections. Consolidated extraction caveats in 14.0; per-batch analyses follow.

14.0 Extraction artefact ledger (consolidated from the six reading agents)

Batch 1 (guides): Elston — bullet characters render as stray "d", title dash garbled; source quirk: online publication 2017 vs print 2019. Blümle — German text; search-history table hit counts misaligned by extraction (final step count hedged). Aromataris — logic-grid tables fragmented, figures lost.

Batch 2 (filters): Beynon's PDF extraction is severely truncated (title page, abstract, plain-language summary only — no methods/results); its deep dive rests on those sections alone, and the Beynon page transcription in this workspace is likewise incomplete. Lefebvre's exhibit tables and Escobar's study-characteristics tables partially garbled; largest files read in chunks. Minor source inconsistencies: Booth reports the "Qualitative Research" MeSH introduction as 2002 vs 2003; Dullea reports the Friedman p as 0.0065 vs 0.007.

Batch 3 (supplementary/reporting): Godino — OCR transcription; figure/flowchart numbers and some DOIs unreliable. Grimshaw and Briscoe — scrambled tables (numbers sourced and flagged where used).

Batch 4 (AI core): Lund — contradictory PRISMA counts (166/196/146/136). Wang — abstract says 86 reviewers, results say 85. Li — Tables 2 and 4 partially garbled (ranges taken from the Discussion). Yarar — 448 vs 449 retrieved reviews; Table 1 header says "39 included studies" though 42 were included. Khan — non-peer-reviewed preprint (later published in JAMIA; cite the journal version). Mahboubi — demonstration paper with no quantitative performance metrics.

Batch 5 (field maps): Johnson — Tables 2–3 garbled (row values detached from labels). Demir — internal SD inconsistency (abstract 0.58/0.71 vs tables 0.53/1.00) and garbled percentage

strings. Ma — numbering typo and text-order scrambling in Tables 1–2. Begagic — minor artefacts (columns interleaved, "e2,000" for €2,000, one mid-sentence column break). Su — table values appear only in prose.

Batch 6 (librarian craft): Spencer — Table 1 channel counts missing. Townsend — Miller-pyramid grid garbled. Farris — Table 2 pre/post-test raw values lost (differences recoverable). Boden — Figure 1 absent, one block-quote displaced. Westrick — Figures 1–4 absent. Parker — text order badly jumbled (quotes reconstructed).

Cross-cutting implication: primary sources in this corpus contain their own internal inconsistencies (Wang, Lund, Yarar, Demir, Dullea). Any automated pipeline that ingests these texts will inherit and silently propagate those errors — human arbitration is not optional (cf. Khan's concordance heuristic and Gallo's measurement-error warning).

14.1 The guides — plain-language search instruction

Eight papers on how to search biomedical literature, ordered by file number.

Watson (2020): How to undertake a literature search: a step-by-step guide — A 10-step practical literature-search guide for nurses written by an RCN search specialist

What the paper is and why it matters: This is a practice-oriented tutorial in the *British Journal of Nursing* (2020, Vol 29, No 7) by Mandy Watson, a Literature Search Specialist at the Royal College of Nursing (RCN) Library and Archive Service. It is aimed at nurses undertaking a search for the first time — "whether for an assignment, a dissertation, an interview or to help you care for a patient." Its importance for a synthesis on literature-searching methodology is that it represents the novice end of the instructional spectrum: it translates librarian craft knowledge into a checklist a non-specialist can follow without any theory. It also shows the institutional scaffolding behind searching (librarians, guides, CPD training), which later papers (Cooper 2018) treat as a formal stage of the systematic-review process.

Methods in detail: The paper is not an empirical study but a structured teaching article that decomposes searching into 10 sequential steps: (1) think about your search question (and check you understand all terms); (2) identify key concepts, ignoring "process" words like *evaluate* or *compare*, optionally using the PEO framework (Population/Exposure/Outcome) — the worked example is "the effectiveness of a mindfulness intervention on health-related quality of life of rheumatoid arthritis patients" — with PICOS and SPIDER mentioned as alternatives; (3) generate synonyms, alternative spellings (anaemia/anemia) and truncation/wildcard conventions (* and ?); (4) choose appropriate databases — a library discovery tool for quick needs, CINAHL, British Nursing Index, MEDLINE, Maternity and Infant Care, AMED, Social Policy and Practice for deeper work; (5) combine terms with Boolean operators, synonyms joined with OR and concepts with AND, illustrated by Venn diagrams and a RCN planning template (Figure 3) mapping concepts to search numbers; (6) apply limits (dates, peer review, age, document type); (7) run the search and review titles/abstracts; (8) adapt the strategy if too many or too few results; (9) save the strategy and results, recommending reference managers (EndNote, Mendeley, Zotero); (10) obtain full text via library links, interlibrary loan, Google, or open-access sources such as PLOS One. Box 1 defines basic terminology (Boolean operators, database, limits, search strategy, subject headings), based on De Brún and Pearce-Smith (2014).

Key results and numbers: Being a guide, it offers process numbers rather than experimental results: exactly 10 named steps; 2 example question frameworks (PEO in depth; PICOS and

SPIDER named); 3 Boolean operators with explicit warnings that AND narrows while OR broadens; a concrete example search topic (mindfulness/rheumatoid arthritis); a searchable-example for truncation ("alcohol* would pick up alcohol, alcoholic, alcoholism, etc. Sm?th would find Smith and Smyth"); 3 recommended reference managers; lists of remedies for too many results (limit to recent material, add terms with AND, limit geography) and too few (widen date range, remove geographic limits, drop the least important concept); a subject heading "major heading" option in CINAHL/MEDLINE; and the RCN help offer (guides, video tutorials, one-to-one training face-to-face or via Skype). It also notes that librarian-run search services "are generally available only to qualified healthcare staff." Key Points restate: careful planning, saving strategy and results, and a structured approach yield "a more focused set of search results."

Conclusions and takeaways: The core argument is that searching is manageable when broken into small, planned steps: "Breaking the exercise down into smaller steps will make the process more manageable." The paper closes that a structured approach "will ultimately save time and effort" and produce a more focused result set usable for a successful assignment. A subtle but important takeaway is the dual recommendation to combine controlled subject headings with self-generated keywords — "A comprehensive search, however, will usually include both subject headings from databases and terms that you have thought of yourself" — which is exactly the keyword+MeSH logic formalized in the more technical papers of this batch. CPD reflective questions push the reader to apply a framework, synonyms, and limits to their own topic.

Limitations and caveats: The article is a single-author practice guide, not research: there is no evaluation of whether the 10 steps improve recall, precision, or speed; the claim that structure saves time is asserted, not tested. Its database advice is institution-specific (RCN membership access) and UK-centric; interlibrary loans carry charges. It is also pre-LLM and even pre-"Best Match"-era in tone: nothing on algorithmic ranking, automation, or AI. Minor extraction artifact: the figure/table content (Venn diagrams, the planning template) appears as scattered fragments in the text, so Figure 2/3 must be reconstructed from captions.

Practical implications for a non-academic researcher using PubMed and AI: Adopt the 10 steps as a literal worksheet: write the question, strip it to 2–4 concepts, list synonyms and spellings before touching the search box, and decide limits up front. Keep a record of every database, term combination, and hit count — this is precisely the audit trail that AI tools and reproducibility checklists later demand, so forming the habit early pays off. Use a reference manager (Zotero is free) from day one. Finally, the paper legitimizes asking a librarian for help; even a non-academic can request database guides or training, and an AI assistant can play the role of a first-pass "synonym generator" for Step 3, provided the human verifies the terms.

Notable quotes:

- Opening: "The key is to break the process down into smaller steps and think carefully about each step in advance." (introduction, before the 10-step list)
- On the NOT operator: "Be careful with NOT because it will exclude any results containing that search term regardless of whether other parts of the article might have been of interest." (Step 5)
- On breadth of material: "It might be worth looking at non-research materials too because a general article might provide a useful overview of your topic." (Step 6)

Connections: Watson is the novice-level counterpart of Aromataris (2014): Constructing a search strategy and searching for evidence: a guide: both teach concept-splitting plus Boolean logic, but Watson uses PEO for general nursing questions while Aromataris builds a full PICO logic grid for systematic reviews; Watson's RCN planning template is effectively a simplified logic grid. Cooper (2018): Defining the process to literature searching in systematic reviews's eight key

stages can be seen as the professional, evidence-based formalization of Watson's informal 10 steps (e.g., "save your strategy" → "documenting the search"). Its Boolean/truncation basics are the same material Motschall (2005): Searching the MEDLINE literature database through PubMed: a short g… and Blümle (2019): Systematic literature search in PubMed: a short introduction teach for PubMed specifically.

Blümle (2019): Systematic literature search in PubMed: a short introduction — German CME tutorial walking clinicians step-by-step through a systematic PubMed search with a live worked example

What the paper is and why it matters: A CME-accredited continuing-education article (*Z Rheumatol* 2019;78:155–172, 3 points; originally published in *Der Ophthalmologe* 2018;115:243–260) by Cochrane authors A. Blümle, W. A. Lagrèze, and E. Motschall (the same Edith Motschall of the 2005 guide). It teaches systematic literature searching in PubMed/MEDLINE through a fully documented worked example, and its editors' preface makes the case that "not only the scientist, but also the clinician, who wants to make diagnostic and therapeutic decisions in an evidence-based manner, needs knowledge of the technique of literature searching." It matters because it is one of the most granular public walkthroughs of PubMed mechanics ever printed — every search step carries its exact syntax and hit count, dated 22.11.2017.

Methods in detail: The didactic device is a clinical question: "Are there randomised studies on the therapy of retinopathy of prematurity (ROP) with anti-VEGF antibodies?" The question is decomposed via PICO (aspect 1: ROP [population]; aspect 2: anti-VEGF [intervention]; comparison and outcome deliberately omitted; study design: RCT). The tutorial then demonstrates: (a) a naive simple search ("ROP AND anti-VEGF") whose Automatic Term Mapping (ATM) translation is inspected under Search Details — and shown to fail for the abbreviation "ROP"; (b) finding terms from known relevant articles ("Similar articles", viewing "Publication types, MeSH terms" of indexed records, the AND medline[sb] trick to restrict to indexed items); (c) a MeSH Database search per aspect — "Retinopathy of Prematurity"[Mesh] (#3, n=5,211) and, for anti-VEGF, climbing the MeSH tree from the 2016 term Bevacizumab up to "Antibodies, Monoclonal"[Mesh] (#10, n=206,931) with the default Explode function kept enabled; (d) complementary truncated textword searches ([tw]) — retinopath*[tw] AND prematur*[tw] (#4), ROP[tw] (#5), vegf*/vascular endothelial growth factor* (#11) and the substance names (#12); (e) combining aspect blocks with OR (→ #6, #13) and the blocks with AND (#14 = 843 records on 22.11.2017); (f) narrowing with the RCT article-type filter (#15) and, because that filter only finds indexed records, re-running with rando*[tw] (#16, final result reported as 67 in the prose); and (g) managing the History display, filters, and saving. It ends with checklist-style lists for narrowing vs expanding results and a summary of PubMed fundamentals.

Key results and numbers: PubMed composition (as of 15.09.2017): 27,593,215 records total, split MEDLINE 24,295,306 / In-Process 526,629 / Publisher 1,978,457 / PubMed-not-MEDLINE 792,823; ~750,000 clinical trial records with ~40,000 added per year; MeSH contains ~26,000 headings in 19 main categories, updated annually; truncation with * matches up to 600 word endings. Precision/Recall are defined algebraically (precision = relevant found / all found; recall = relevant found / all relevant present) with the caveat that recall "is a theoretical quantity" since the database's total relevant set is unknown. Worked search metrics (22.11.2017): ROP MeSH 5,211; Antibodies, Monoclonal MeSH 206,931; aspect-2 block #13 = 284,507; combined ROP AND anti-VEGF (#14) = 843; the RCT filter and final rando*[tw] step (history #15/

#16) yield the small final set, with the prose reporting 67 records for the closing step. A CME quiz question even drills precision arithmetic (100 hits, 20 irrelevant → precision 80%).

Conclusions and takeaways: The Fazit für die Praxis states that systematic searching is necessary for reliable retrieval; a PICO-formulated question anchors a reliable strategy; MeSH searching catches synonym/spelling variation in the indexed portion, while "the non-indexed portion of PubMed can only be searched with text words," and textword searches also recover wrongly or inconsistently indexed articles. The recurring trade-off is precision vs recall: each narrowing device (subheadings, Major Topic, explode off) raises precision at the risk of losing relevant records. The tutorial's implicit thesis — echoed in the editors' preface — is that PubMed competence is now a core clinical skill, learnable by following a protocol rather than intuition.

Limitations and caveats: All hit counts are snapshots (22.11.2017) and the authors warn that a reader's own history numbering will differ after deletions or detours, and that PubMed even reorders history steps when filters are applied and cleared — the extraction itself shows garbled table alignment around the final hit counts. The example is ophthalmology-specific; one clinical scenario cannot prove generality. It is interface-bound to a 2017-era PubMed (Advanced search builder layout) and predates Best Match default sorting nuances discussed only in passing. The text is largely German (English abstract provided), which limits accessibility for some readers.

Practical implications for a non-academic researcher using PubMed and AI: Treat the paper as a script to replay: run one naive search, read Search Details to see what PubMed did, then build one search line per PICO aspect (MeSH + truncated textwords OR'd together) and AND the lines. Never trust a single MeSH term — climb the tree and note the "Year introduced." If an AI assistant drafts your search, use this paper's discipline as the acceptance test: ask it to show each block separately with hit counts, and always add a textword fallback like the rando*[tw] pattern for anything PubMed indexes with delay. The lists of ways to widen or narrow an over/under-performing search are directly usable as prompts for AI-assisted query tuning.

Notable quotes:

- From the English abstract: "In order to identify current (and relevant) evidence for a specific clinical question within the unmanageable amount of information available, solid skills in performing a systematic literature search are essential." (Abstract)
- Margin maxim (German): "Eine Recherchestrategie ist nur so gut wie die Fragestellung, auf der sie basiert" — "A search strategy is only as good as the question it is based on." (margin note to the PICO section)
- On unindexed material: "Der nichtverschlagwortete Teil von PubMed kann nur mit Textwörtern durchsucht werden" — "The non-indexed part of PubMed can only be searched with text words." (Fazit für die Praxis)

Connections: This is effectively the 2019 update and German-language expansion of Motschall (2005): Searching the MEDLINE literature database through PubMed: a short g… — they share the author Motschall, the same quick-vs-comprehensive arc, and even the same clinical-example style. Its precision/recall framing supplies the quantitative vocabulary that Cooper (2018): Defining the process to literature searching in systematic reviews uses when discussing "comprehensive" search aims. Its step-by-step History mechanics and MeSH-tree navigation are the manual procedures that Muin (2006): PubMed interact: an improved interface for MEDLINE/PubMed searches tried to automate in its interface, and its narrowing/widening checklists become urgent in light of Kim (2022): Towards a unified search: improving PubMed retrieval with full text's finding that ~15% of PubMed queries return nothing.

Motschall (2005): Searching the MEDLINE literature database through PubMed: a short g... — English-language "short guide" demonstrating MeSH-based PubMed searching on a GERD/cancer question

What the paper is and why it matters: A "Clinical Information" tutorial published in *Onkologie* (Karger, published online 12 August 2005; DOI 10.1159/000087186) by Edith Motschall (information retrieval specialist, Freiburg) and Yngve Falck-Ytter (Deutsches Cochrane Zentrum / Case Western Reserve). It is one of the earliest widely circulated concise English walkthroughs of PubMed's sophisticated features, and a direct ancestor (in method and in authorship) of the 2019 Blümle tutorial. Its historical importance: it captures PubMed as it was in 2005 (12+ million MEDLINE citations, History, Limits, Clinical Queries, Auto-Explode) and articulates, before "reproducible search" became standard rhetoric, that quick searches and comprehensive searches serve different purposes.

Methods in detail: The guide starts with a real clinical question — "Is a 30-year-old man with typical acid reflux symptoms for many years (gastroesophageal reflux disease, GERD) more likely to develop esophageal cancer than people without reflux symptoms?" — split into patient (GERD), outcome (esophageal cancer: etiology/risk) and study-design components (cohort, case-control). Phase 1, "Quick and Simple Search": typing "gastroesophageal reflux esophageal cancer" into the start page triggers Automatic Term Mapping to MeSH; clicking "Details" exposes the internal translation; the result is 807 citations (accessed 11/02/2005). Phase 2, "Comprehensive Search": in the MeSH Database, "Gastroesophageal Reflux" is located (entry terms like GERD and reflux map to it), combined with the related term "Heartburn" via "Search Box with OR" → 13,304 citations; "Esophageal Neoplasms" is then narrowed with the subheading etiology → 6,115; the History feature combines the two intermediate sets (#5 AND #8) → 336 citations. Study-design vocabulary is added through MeSH terms Cohort Studies, Case-Control Studies and Risk, exploiting Auto-Explode so that narrower terms (Longitudinal, Follow-Up, Prospective Studies; Retrospective Studies; Logistic Models, Risk Assessment, Risk Factors) are included automatically → 1,002,993 citations; ANDed with the 336 → 113 citations, which contain Lagergren's landmark NEJM 1999 reflux/adenocarcinoma study. Further sections cover Limits (publication types, human, dates), textword searching ([tw], [tiab], [ti], truncation neoplas*), the NOT medline[sb] trick for non-MEDLINE records, Clinical Queries (etiology category, narrow vs broad), and saving in MEDLINE format for Reference Manager, Procite, EndNote.

Key results and numbers: MEDLINE at the time: "more than 12 million bibliographic citations from over 4,600 international biomedical journals"; over 20,000 MeSH headings. Hit-count ladder: naive search 807 → MeSH reflux OR heartburn 13,304 → esophageal neoplasms/etiology 6,115 → combined 336 → study-design term search 1,002,993 → final combination 113. Clinical Queries: narrow/specific etiology search = 27 citations; broad/sensitive = 369; crucially, the specific search misses the relevant Lagergren 1999 article. The seven-step summary at the end (formulate PICO → MeSH Database → History combination → inspect indexed citations' MeSH → Limits → textwords → Clinical Queries with Details check) functions as the paper's checklist deliverable. It also notes the SARS example: the thesaurus is updated annually, so "new diseases and substances may not yet have MeSH terms."

Conclusions and takeaways: The paper's analytic core is the pros-and-cons comparison of MeSH vs textword searching: MeSH describes content precisely and exploits hierarchy (explode, subheadings raise precision), but suffers from human indexing inconsistency and vocabulary lag; textword searching reaches unindexed records but demands manually listing synonyms and

spellings. “In practice, both textword and MeSH searching are complementing each other.” The quick approach is endorsed only as a starting move for harvesting candidate MeSH terms; designed strategies using the MeSH Database, limits and textwords are “preferable because it allows more flexibility.” Humility is explicit: several strategies may be needed, and “we will never know how many relevant articles were missed.”

Limitations and caveats: Every number is a 2005 snapshot of both the database and a pre-2005 interface (Limits page, Preview/Index, “Go” button) that no longer exists; screenshots referenced in figures are lost or garbled in the text extraction. The single worked example (etiology question) favors subheading-based narrowing, which the authors themselves flag as inconsistent when applied by indexers. No evaluation of the guide’s effectiveness is attempted, and the authors admit the recall ceiling problem — the true number of missed relevant articles is unknowable from within PubMed.

Practical implications for a non-academic researcher using PubMed and AI: Use the paper’s hit-count ladder as a mental model: a trustworthy answer-level search should end in a result set small enough to read (113 in their case), reached by documenting each combination step, not by tuning a one-box query. When AI tools propose PubMed queries, apply the two checks this paper pioneered: inspect the “Details”/Search-Details translation the engine actually ran, and confirm that known relevant papers (the “pearl” test, here Lagergren 1999) are retrieved — Clinical Queries-style narrowed searches failed that test. Finally, borrow the practice of ending with the authors’ 7-step summary as a personal SOP.

Notable quotes:

- On the naive approach: “Although this low-effort approach may be all it takes to find the approp

riate evidence fast, the proportion of irrelevant articles may be quite high.“ (Introduction, after the 807-citation figure)

- Mid-tutorial reassurance: “Don’t be intimidated by this large number. Remember, combining this result with the 336 citations from step #9 with the Boolean operator AND limits the considerably…” (Combining Several Search Steps section, after the 1,002,993-citation set)
- On the limits of knowledge: “In practice, several approaches and different strategies may be necessary to vary precision and recall rates and to compare the results. Unfortunately, most of the time, we will never know how many relevant articles were missed.” (Simple or Comprehensive Search? section)

Connections: Direct ancestor of Blümle (2019): Systematic literature search in PubMed: a short introduction: same Freiburg/Cochrane lineage, same quick-then-comprehensive pedagogy, same MeSH/textword complementarity — the 2019 paper refreshes its numbers and adds the History table format. Its Clinical Queries and field-tag material reappears in compressed form in Elston (2019): Getting the most out of online literature searches — tips for advan…. Its explicit MeSH-vs-textword pros/cons and humility about missed articles anticipate the evaluation concerns that Cooper (2018): Defining the process to literature searching in systematic reviews systematizes, and its “quick search as term-harvesting” move is exactly the iterative term discovery that Aromataris (2014): Constructing a search strategy and searching for evidence: a guide prescribes.

Aromataris (2014): Constructing a search strategy and searching for evidence: a guide — Joanna Briggs Institute tutorial on constructing a PICO-anchored, reproducible search strategy for systematic reviews

What the paper is and why it matters: The third article in an American Journal of Nursing series teaching nurses to conduct systematic reviews "one step at a time," written by Edoardo Aromataris (director of synthesis science, Joanna Briggs Institute, Adelaide) and Dagmara Riitano. It bridges the gap between the basic how-to guides (Watson, Motschall) and formal methodological guidance (Cooper): it explains not just mechanics but why a search must be comprehensive, how to document it, and how the whole enterprise rests on a review protocol. Because JBI is one of the nine guidance bodies later audited by Cooper (2018), this article is a readable window into that guidance tradition.

Methods in detail: The article constructs the search strategy for a running example question: "Is animal-assisted therapy more effective than music therapy in managing aggressive behavior in elderly people with dementia?" Steps: (1) start from a review protocol with defined question and inclusion criteria, and decide data types (quantitative/qualitative/economic), study designs, and limits; (2) build a logic grid/concept map with columns aligned to the four PICO elements; (3) populate columns with keywords harvested from an initial low-effort search ("dementia AND animal-assisted therapy") by scanning titles/abstracts and reading a few relevant papers; (4) add index terms by inspecting how retrieved articles were MeSH-indexed, using the MeSH database trees (16 main trees noted) and "Entry Terms" as synonym hints; (5) qualify every term with field descriptors — [mh] for MeSH, [tiab] for title/abstract, [tw] preferred over [tiab] for systematic reviews because it searches more fields — and apply truncation (therap*); (6) search each column separately with OR within columns (e.g., the dementia column "could yield more than 100,000 citations"), watching for contaminating terms (aggress* also describes cancer spread); (7) combine columns with AND via the Advanced search history, which auto-parenthesizes sets; (8) refine with filters and justify any date limits in the protocol. The full final PubMed strategy for all four columns is printed verbatim. It then covers where to search (MEDLINE via PubMed, Embase/Emtree, CINAHL, PsycINFO, Web of Science, Scopus…), warns against federated search engines for reviews, and extends the method to supplementary techniques: reference-list screening, hand searching (no more than two or three journals), expert contact, and three routes to grey literature (organization websites; OpenGrey and the Grey Literature Report; web engines with site:gov tricks, MedNar, CADTH's Grey Matters checklist). Reporting closes the loop: PRISMA requires the full strategy for at least one database as an appendix.

Key results and numbers: There are no outcome metrics; the "numbers" are structural: 4 PICO concepts; 4 logic-grid tables evolving from bare concepts to fully field-coded terms (e.g., Dementia [tw] OR Alzheimer [tw] OR Huntington* [tw] … OR Dementia [mh]); one complete combined search string of ~60 terms with parenthesized blocks; an estimate that a single-concept dementia search yields ">100,000 citations"; recommended hand-searching ceiling of 2–3 journals; 3 named grey-literature approaches and 2 named grey databases (OpenGrey, Grey Literature Report) plus CADTH Grey Matters; a named list of core databases (MEDLINE, Embase, CINAHL, plus Web of Science, PsycINFO, Scopus, JSTOR, Academic Search Premier/ Onefile, Cochrane Nursing Care Field register); and platform differences (asterisk truncation in PubMed vs $, ?, # in Ovid; character-substitution wildcards unsupported in PubMed).

Conclusions and takeaways: Searching for a review is framed as disciplined iteration: "Developing a search strategy is an iterative process — that is, it involves continual assessment and refinement," and no two searchers will produce identical strategies. Sensitivity and specificity must be balanced knowingly, with the default bias toward sensitivity because "if you don't find the evidence, it can't be reviewed!" Two hard rules stand out: comprehensive strategies must contain BOTH free-text keywords and controlled index terms (because indexing lags and varies — the Bonding, Human-Pet example), and reporting is inseparable from searching, since reproducibility — "A hallmark of a systematic review is its reproducibility" — depends on transparent publication of databases, platforms, dates, limits, and full strategies (PRISMA).

Limitations and caveats: It is guidance, not evidence: no evaluation of the logic-grid method's performance is offered, and the authors concede there is "no prescriptive approach to conducting a comprehensive search." Some specifics are dated — OpenGrey and the Grey Literature Report are long defunct; MEDLINE/PubMed's 2014 interface and Limits predate Best Match; and the PMCID/PMC full-text scale discussed by Kim (2022) has since tripled. The example is deliberately narrow (a therapy-comparison question), and the claim that broad terms like "safety" add little beyond specific terms is asserted without citation. Table contents in the extraction are fragmented across the pdftotext output, so the logic grids need careful reconstruction.

Practical implications for a non-academic researcher using PubMed and AI: Build a four-column table (or spreadsheet) before searching: even without PICO formalism, columns of synonyms + MeSH terms per concept make searches reproducible and give AI assistants something structured to expand and audit. Run each column as a separate search line and check hit counts before ANDing — a habit that also gives you the data to tell whether an AI-generated query block is pulling its weight. Adopt the article's known-article check ("are already identified relevant articles being captured?") as the acceptance test for both manual and AI-drafted strategies, and always export results into a reference manager with duplicates removed.

Notable quotes:

- The guiding maxim: "One rule of thumb we use when beginning a search for evidence to support a systematic review: if you don't find the evidence, it can't be reviewed!" (opening section, GETTING STARTED)
- On craft: "Consequently, searching for evidence is sometimes considered more of an art than a science. It's therefore unlikely that two people, whether they are clinicians or librarians, will develop an identical search strategy or yield identical results from a search on the same review question." (GETTING STARTED)
- On why free text matters: "Because new articles in PubMed are not indexed immediately, and because indexing is a manual, subjective process susceptible to human variation, it's important to also search for the key terms in the titles and abstracts of articles." (Index terms or subject headings section)

Connections: The Joanna Briggs Institute Reviewers' Manual tradition it represents is one of the nine guidance documents audited in Cooper (2018): Defining the process to literature searching in systematic reviews, whose key stages 2 (aims), 4 (PICO/SPICE/SPIDER structures) and 8 (PRISMA reporting) map almost one-to-one onto this article's arc. It shares the concept/synonym/Boolean scaffolding with Watson (2020): How to undertake a literature search: a step-by-step guide but at systematic-review rigor, and its keyword-plus-MeSH doctrine operationalizes the MeSH-vs-textword analysis of Motschall (2005): Searching the MEDLINE literature database through PubMed: a short g… and Blümle (2019): Systematic literature search in PubMed: a short introduction.

Elston (2019): Getting the most out of online literature searches — tips for advan... — One-page JAAD editor's letter of quick advanced-search tips for clinicians

What the paper is and why it matters: A single-page "Letter from the Editor" in the *Journal of the American Academy of Dermatology* (2019;81:905) by Dirk M. Elston (Medical University of South Carolina), titled "Getting the most out of online literature searches — tips for advanced searches" (the em dash renders as "searchesdtips" in the extraction). It compresses the essential PubMed syntax into a format a busy clinician can absorb in minutes, and explicitly promises a follow-up letter. In this batch it serves as the minimal viable curriculum — the irreducible core (Boolean logic, term mapping order, field tags, truncation, limits) that every longer tutorial elaborates.

Methods in detail: The letter is pedagogy by example rather than systematics. It opens by contrasting natural-language querying (typing a question like "what treatments are effective for cold urticaria but not ordinary urticaria?") with Boolean searching, noting that translation of questions into terms is variable — "slightly different phrasing can produce very different results." It then itemizes: the three Boolean operators (OR expands across citations; AND narrows with no positional constraint; NOT excludes but risks losing relevant articles) with the requirement of UPPERCASE, and left-to-right processing unless parentheses force innermost-first evaluation. It describes PubMed's term-resolution order — "PubMed maps terms to the MeSH Translation Table of key words first. If it does not find a match, it moves on to the Journals Translation Table and then the Author Index" — and recommends last name plus initials for authors. Field tags are introduced with abbreviations ([jn], [tw], [au], [ti], [ot]) and the Ovid equivalents (.jn, .au). Truncation is demonstrated with lichen* retrieving lichen, lichenoid, etc. Two troubleshooting moves close the technical content: the Details button to see how the search was performed, and wrapping an author name in double quotes with [au] to switch off automatic truncation. Limits (fields, language, age, gender, human vs animal) and Clinical Queries filters (therapy, diagnosis, etiology, prognosis) are mentioned, followed by a one-paragraph motivational conclusion.

Key results and numbers: As a letter, it contains no dataset results. Its enumerable content: 3 Boolean operators; 3-step term-mapping order (MeSH table → Journals table → Author Index); 5 field tags (jn, tw, au, ti, ot); 1 worked search question (cold urticaria vs ordinary urticaria); 1 truncation example (lichen*); 4 Clinical Queries categories (therapy, diagnosis, etiology, prognosis); and 1 troubleshooting technique (quoted author names). Publication metadata itself is a small anomaly: the 2019 volume carries DOI 10.1016/j.jaad.2017.11.017 and a "© 2017" copyright line, suggesting online-first publication in late 2017 ahead of the 2019 print issue.

Conclusions and takeaways: The letter's thesis is a cost-benefit claim about learning search syntax: the operators and tags "take a little time to learn initially but can save countless hours when trying to find treatment options for a difficult patient." It reassures that the techniques "may seem a bit daunting at first but are actually quite easy to apply" and can "make you a hero to a patient in need of more effective treatment." The takeaway for methodology historians is that by 2017/2019, editorial boards of major clinical journals considered Boolean/field-tag literacy part of authorship and clinical competence — and that natural-language input, while easy, is explicitly flagged as less predictable than constructed queries.

Limitations and caveats: By design it is only a teaser: no worked-through results, no hit counts, no discussion of precision/recall, no mention of systematic search requirements (PRISMA, PRESS), and it excludes MeSH-tree navigation, history building, or textword-vs-MeSH complementarity. Some claims are impressionistic — e.g., that lowercase "and" or simply strung-together

terms "often produces similar results" is anecdotal ("in my experience"). The pdftotext extraction mangles bullet characters into stray "d" letters and the title's dash into "d," and the letter's promise of a "future letter" cannot be verified from this text alone.

Practical implications for a non-academic researcher using PubMed and AI: Memorize the letter's three essentials before anything else: uppercase AND/OR/NOT, the meaning of [tw]/[ti]/[au]/[ot] tags, and checking Details after every search. When an AI assistant writes queries for you, insist it outputs them in exactly this style (uppercase operators, parenthesized blocks, field tags) — that makes its output auditable in raw PubMed rather than locked inside a chat tool. Use the author-name quoting trick when tracking down a specific paper, and treat the letter's warning as the default stance: natural-language one-box searches are fine for orientation, but any decision-relevant search should be rebuilt with explicit Boolean structure.

Notable quotes:

- On input sensitivity: "You can simply type in your question, but to complete the search, the system translates the question to search terms and categories, and slightly different phrasing can produce very different results." (opening paragraphs)
- On NOT: "NOT excludes terms to eliminate unwanted references, but may also eliminate relevant articles and should be used carefully." (Boolean logic section)
- On the payoff: "They may seem a bit daunting at first but are actually quite easy to apply and can make you a hero to a patient in need of more effective treatment." (introduction)

Connections: It is the compressed, single-page rendition of the same material taught at length in Motschall (2005): Searching the MEDLINE literature database through PubMed: a short g… and Blümle (2019): Systematic literature search in PubMed: a short introduction — Clinical Queries, field tags, truncation, Details — with the identical 2005-era warning about NOT that Watson repeats. Its natural-language-versus-Boolean framing foreshadows the ranking issues in Kim (2022): Towards a unified search: improving PubMed retrieval with full text, where query phrasing and token placement determine retrieval outcomes, and its assumption that clinicians self-search is the same premise underlying Muin (2006): PubMed interact: an improved interface for MEDLINE/PubMed searches's novice-friendly interface project.

Muin (2006): PubMed interact: an improved interface for MEDLINE/PubMed searches — NLM technical report on PubMed Interact, an Ajax-based alternative PubMed interface with interactive result manipulation

What the paper is and why it matters: A software article in *BMC Medical Informatics and Decision Making* (2006;6:36) by Michael Muin and Paul Fontelo of the National Library of Medicine's Office of High Performance Computing and Communications. It documents the technical development of PubMed Interact, an alternative Web interface to MEDLINE/PubMed descended from their earlier SLIM ("Slider Interface for MEDLINE/PubMed searches," 2005). Its significance is twofold: historically, it is an early NLM experiment in rethinking PubMed's user experience — trying to surface filters "that often remain hidden from novice users" — and conceptually, it anticipates the modern expectation that search results be manipulable objects (delete, tag, augment) rather than a static ranked list, a trajectory that continues today in AI-era search UX.

Methods in detail: PubMed Interact is a client-server Web application: PHP 4.4.1 backend on Windows Server 2003 with Apache 2.0.52 and MySQL 4.1.18, HTML/JavaScript front end.

Server-side object-oriented PHP classes act as modified XML parsers connecting to NCBI's Entrez Programming Utilities (ESearch, EFetch, ELink); retrieved XML for the top 200 citations is cached in a local MySQL database to reduce load on E-Utilities, so subsequent actions (deletions, related-article fetches) are served locally. The front end's interactivity is built on DHTML, Document Object Model (DOM) tree manipulation and Ajax (XMLHTTPRequest, restricted to the local domain, with PHP proxies to E-Utilities). The search form has seven JavaScript slider bars — publication date (separate start/end-year sliders), journal subset (PubMed vs MEDLINE subset vs Core Clinical Journals, further limitable to abstracts or full/free full text), age group, methodology filter (case reports via publication type; the Haynes RB et al.-derived clinical study categories, i.e., PubMed Clinical Queries; and the pre-configured systematic reviews subset), a MeSH-mapping slider for advanced users that rewrites search tags from the mapped MeSH terms, and a results-per-page control (10–100) — plus a dropdown limiting to human studies and/or English. Notable functions: a "Preview Count" button that asynchronously fetches the hit count before submitting; an information box displaying total results plus mapped MeSH terms, mapped subheadings and unmapped terms (explicitly "an educational tool for new users"); toggleable abstracts with structured-abstract rendering via regular expressions; deletion of individual citations (brief red flash, then DOM removal); the Auto-Append Article (A3) feature that backfills the list from the cache as citations are deleted; two color-coded user relevance lists (high = light green, low = light yellow); and per-citation retrieval of the top 10 related articles in a floating DIV, with "Add to Main List" inserting them into the working set.

Key results and numbers: The evaluation is technical rather than behavioral. Deployment: alpha went live late November 2005, beta in February 2006. Scale: the beta's local MySQL cache held "over 29,900 records of citations in XML format" using 54 MB of disk, with a maintenance script deferred so MySQL benchmarks could be recorded. Efficiency-by-design: one E-Utilities query stores the top 200 citation details regardless of display count; the related-articles routine stores details for all 10 related articles per citation locally. Compatibility: full functionality in Windows Firefox 1.5+, IE 5.5+, Opera 8.5+ and Linux Firefox 1.5+; formatting inconsistencies (but no functional failures) in Mac OS X Firefox 1.5+ and Safari 2.0+. Standards testing via the W3C Markup Validation Service flagged an unsupported element attribute for each slider and one recurring error per citation (numerical strings used as id attributes); the authors retained these "invalid codes" because removing them "degraded the functionality of the application." No precision/recall, timing, or user-performance figures are reported.

Conclusions and takeaways: The authors position PubMed Interact as "an experiment in user-computer interface," arguing that DOM manipulation plus Ajax bridges "an effective search strategy with a highly-interactive interface": users can set parameters, refocus searches, label, delete, and enrich the citation list without page reloads. The project's stated long-term aim is an interface that "empowers novice users with interactive tools for search parameters and provides expert users with easy access to advanced search filters." Future work explicitly plans comparative usability studies against Entrez PubMed with users of varying skill levels, assessing performance, search outcomes, educational impact, and effects on searching attitudes and strategies. The tool is offered freely ("anyone may use the service," no non-academic restrictions).

Limitations and caveats: The paper candidly limits itself: "This paper is limited to the development, implementation and technical evaluation of PubMed Interact. It does not provide empirical evidence to show increased efficiency in searching or better precision and recall for results," and "a formal user evaluation … is needed." No evidence is given that sliders or interactive lists actually improve retrieval. The architecture is a period piece (PHP 4, IE

5.5, XMLHTTPRequest same-origin constraints), and the deliberate retention of W3C-invalid markup is acknowledged as a compatibility gamble. The project home page given (http://pmi.nlm.nih.gov/interact) is long dead, and no comparative study against Entrez PubMed appears to have been published from this exact project — so its actual impact cannot be verified from the text.

Practical implications for a non-academic researcher using PubMed and AI: Although the tool itself is defunct, its design logic is directly reusable: (1) preview hit counts before committing to a query — modern PubMed's number display and any AI workflow should follow the same cheap-preview discipline; (2) read the mapped MeSH terms and unmapped terms for your query (the information box idea) to learn controlled vocabulary from your own searches; (3) curate results actively — delete irrelevant items, tag relevance, and pull in "related articles" of promising hits as a supplementary search method, which Cooper (2018)'s batch would classify as legitimate searching. When using AI assistants, request these same affordances: per-block hit counts, MeSH mapping transparency, and iterative list-building rather than one-shot output.

Notable quotes:

- On the problem addressed: "The goal of the project is to advance the development of a Web-based medical search tool that can enhance user interaction with the MEDLINE/PubMed database and push to the forefront the different strategies and filters in Entrez PubMed that often remain hidden from novice users, such as age groups, clinical study filters and systematic reviews." (Background)
- On the UX philosophy: "Many interactive features occur at client-side, which allow instant feedback without reloading or refreshing the page resulting in a more efficient user experience." (Abstract, Results)
- The honest boundary: "A formal user evaluation of the application is needed to validate the usability and benefits of an alternative PubMed search interface." (Limitations)

Connections: Muin and Fontelo's NLM project is the interface-layer sibling of Kim (2022): Towards a unified search: improving PubMed retrieval with full text: both are NCBI/NLM efforts to improve PubMed retrieval, one through interactive UI (filters surfaced, results manipulable), the other through the ranking engine itself (full-text scoring) — together framing the two levers of search quality. Its methodology-filter slider operationalizes the Clinical Queries material of Motschall (2005): Searching the MEDLINE literature database through PubMed: a short g…, Blümle (2019): Systematic literature search in PubMed: a short introduction and Elston (2019): Getting the most out of online literature searches — tips for advan…, and its MeSH-mapping slider automates the manual "Details/term harvesting" step those tutorials teach.

Kim (2022): Towards a unified search: improving PubMed retrieval with full text — NCBI study integrating full-text articles into PubMed ranking via PAV-transformed log-odds scores, using click logs

What the paper is and why it matters: An original-research paper in the *Journal of Biomedical Informatics* (2022;104211) by Won Kim, Lana Yeganova, Donald C. Comeau, W. John Wilbur and Zhiyong Lu — the NCBI/NLM team behind PubMed's relevance search. It addresses a structural blind spot: PubMed's 34+ million abstracts versus ~5 million full texts in PubMed Central (PMC) that cannot be queried simultaneously. The paper asks how to score full-text articles and merge those scores with abstract-only scores without degrading retrieval.

It matters to this synthesis because it quantifies, at production scale, exactly why naive full-text search hurts precision — the central tension behind every "search the body text, not just the abstract" recommendation in the tutorial papers.

Methods in detail: Data: the PubMed Click Dataset (PCDataset), mined from 2017 PubMed user logs, using the click-through assumption that a clicked document is more relevant than the higher-ranked documents seen and rejected. Crucially, the dataset excludes documents with any query token in the title (title presence is a known strong click signal) and requires all query tokens to be present in the abstract — deliberately creating a hard setting where negatives also contain all query terms. It was split two-thirds training (1,268,941 queries) / one-third test (634,364); on average 6.60 documents per query qualified, ~30% labeled positive, and 2.65 documents per query (~40%) had PMC full text; average abstract length 114 tokens, average body text 2,378 tokens. Full texts came from the PMC collection in BioC (~3 million manuscripts), with section types normalized via BioC identifiers into 13 types (Abstract, Abbreviation, Caption, Discussion, Case, Keyword, Conclusion, Result, Methods, Introduction, Section Titles, Supplement, Appendix). Scoring: BM25 at paragraph level (k1=1.2, b=0.75, k3=1000; average paragraph length 46.4 tokens); section scores built from paragraph scores via max, sum, or top-N sum strategies; then, instead of raw BM25, each section type's BM25 score is transformed by Pool Adjacent Violators (PAV) isotonic regression into a probability of a click, converted to a log odds ratio (relative to the random-click prior), and section log-odds are summed — the Naive Bayes assumption. score(d,K) sums the K highest section scores; a second-layer PAV maps this combined score to a click probability so full-text and abstract-only documents become comparable. Comparators: abstract-only BM25, naïve body-text BM25, article-as-one-text BM25, abstract+body sum, BM25_SumK, and BM25F (Robertson & Zaragoza's weighted-sections extension), optimized greedily. Evaluation: MAP with percentile bootstrap (5% level, 10,000 resamples).

Key results and numbers: Random ranking achieves MAP 0.579; abstract-only BM25 reaches 0.6405 (+10.6% over random). The best configuration of the proposed method, LogOddsRatio_Sum3 — sum of all paragraph scores as section score, K=3 top sections (97% of PMC full-text documents have ≥3 sections) — reaches 0.6448: "we gain another 0.67%, or 7% relative improvement over abstract alone," i.e., 11.36% over random; the difference is statistically significant against all comparators (vs BM25F: ΔMAP 0.0012, 95% CI [0.0007, 0.0017]). Naïve incorporation of body text degrades performance: BM25_Article and basic abstract+body combinations score below abstract-only. BM25F improves over abstract-only but is beaten by the proposed method, at far higher training cost — PAV needs 13×N query-document evaluations vs 7,800×N for BM25F; PAV took ~5 minutes for ~8 million pairs, while BM25F would take ~29 days for only ~66 thousand pairs (completed in under a week on a compute farm). Section analysis (1,683,182 full-text articles, 40% of test documents): Abstract is by far the most often top-3 section (0.0087 frequency-of-top-3 for a ubiquitous section… per Table 1, Discussion 0.2048, Result 0.2625, Caption 0.2679 lead among body sections). Zero-result opportunity: "About 15% of PubMed queries return no results"; re-running zero-result queries logged on Oct 12, 2018 against PMC showed 54% retrieve full-text documents — "equivalent to ~140 K zero-result queries a day that could potentially produce results" (e.g., query "cd40 fmf" fails in PubMed but matches 90+ PMC documents).

Conclusions and takeaways: Two formal conclusions: (1) "BM25 section scores are not directly comparable with each other for making predictions" — the same score means different things in the abstract, Methods, or Discussion, and PAV-based log-odds normalization fixes this; (2) LogOddsRatio_SumK with a second-layer PAV successfully merges full-text and abstract scores into a unified ranking that beats abstract-only. The authors stress the austerity of their test bed

(all query tokens already in every abstract; users see only titles while scoring uses text they never see; click data are noisy) and expect larger gains where query terms appear only in body text. A "particular significance" is claimed for zero-result and low-yield queries, where full-text matching can rescue retrieval. Extensions to title-matching documents and neural retrieval methods are flagged as future work before production implementation.

Limitations and caveats: The authors state the key limitation: "we have excluded documents with query terms in the title… Our results do not generalize directly to documents with query words in the title." The gold standard is implicit relevance from click logs — a known-noisy proxy — and evaluation rewards predicting "the appeal of the title," an indirection the authors concede. Improvements are statistically significant but small in absolute terms (0.67 points of MAP), and the 0.579 random-baseline construction means much of the apparent signal comes from the dataset's design. Production PubMed does not (as of the paper) implement this method. For a reader of the tutorial papers, note this is relevance-ranking research, not search-strategy research: nothing here addresses MeSH, filters, or comprehensiveness.

Practical implications for a non-academic researcher using PubMed and AI: Know that PubMed ranks and retrieves on title/abstract/bibliographic fields only: a paper can discuss your topic extensively in the body and still be invisible to your query — so when PubMed returns nothing, do not conclude the evidence doesn't exist; ~15% of queries return zero results and 54% of such queries have full-text matches in PMC. Practice the paper's implicit lesson for AI-assisted searching: use multiple paraphrases/synonyms (token coverage matters more than phrasing elegance), try Europe PMC or PMC full-text search as a body-text complement, and use "similar articles"/related-article features to cross from abstract-matched to full-text-relevant documents. For AI tools that promise "full-text search," ask whether they weight sections — Kim et al. show naive body-text search actively lowers precision.

Notable quotes:

- The counterintuitive core finding: "Naïvely using the body text of articles along with abstract text degrades the overall quality of the search." (Abstract, Results and conclusions)
- The headline gain: "By including full text where available, we gain another 0.67%, or 7% relative improvement over abstract alone." (Abstract)
- The practical bombshell: "About 15% of PubMed queries return no results [18] due to one or more query tokens not being present in the title or abstract." (Conclusions and discussion, zero-result queries paragraph)

Connections: Kim et al. quantify the recall-vs-precision trade-off that Blümle (2019): Systematic literature search in PubMed: a short introduction and Motschall (2005): Searching the MEDLINE literature database through PubMed: a short g… teach qualitatively (textword searches raise recall and lower precision; here body-text search does the same), and their zero-result finding gives hard numbers to the "too few hits — how to expand" checklists of those tutorials. It is the ranking-engine counterpart of Muin (2006): PubMed interact: an improved interface for MEDLINE/PubMed searches's interface experiment — two NLM generations attacking the same user-experience problem from different layers. Its log-based evaluation mindset contrasts sharply with the guidance-document analysis of Cooper (2018): Defining the process to literature searching in systematic reviews, illustrating how search methodology is studied at both the engineering and the process levels.

Cooper (2018): Defining the process to literature searching in systematic reviews — Literature review of nine guidance documents distilling eight key stages of literature searching in systematic reviews

What the paper is and why it matters: A DEBATE/methodology article in *BMC Medical Research Methodology* (2018; DOI 10.1186/s12874-018-0545-3) by Chris Cooper (Exeter), Andrew Booth (ScHARR, Sheffield), Jo Varley-Campbell, Nicky Britten and Ruth Garside. Its premise: information specialists and review teams "appear to work from a shared and tacit model of the literature search process" that has never been made explicit. The paper's contribution is to articulate that implicit model as eight key stages, derived from nine major guidance documents, and to map supporting empirical studies onto each stage. For this synthesis it is the structural backbone of the batch: it names and orders the activities that Watson, Motschall, Blümle and Aromataris teach piecemeal.

Methods in detail: A literature review with two evidence streams. (1) Guidance: nine documents selected purposively for prominence in UK-informed systematic reviewing — CRD's guidance (2009); the Cochrane Handbook v5.1 (2011); Collaboration for Environmental Evidence guidelines v4.2 (2013); JBI Reviewers' Manual (2014); IQWiG Methods Resources (2014); Petticrew & Roberts, *Systematic Reviews in the Social Sciences* (2006); EunetHTA information-retrieval guideline (2016); the Campbell Collaboration *Searching for Studies* guide (2017); and the NICE guidelines manual (2014). The list was checked by six senior UK information professionals. (2) Supporting studies: identified via authors' topic knowledge, "pearl growing" (Booth's search-tactics pearls), backward and forward citation chasing in Google Scholar, and a PubMed search using the systematic review methods filter — "(literature search*[Title/Abstract]) AND sysrev_methods[sb]" — run in August 2017, returning 586 results, sifted by the lead author. Analysis: relevant sections of each document were read and re-read; a "key methodological stage" was defined as "a distinct step in the overall process for which specific guidance is reported, and action is taken, that collectively would result in a completed literature search"; chapter subheadings were extracted verbatim into a table, compared and contrasted for agreements and unique guidance, and cross-document consensus selected the key stages.

Key results and numbers: Eight key stages emerged: (1) who should search — eight documents say searchers with relevant expertise should "ideally" be in the team; six name information specialists, librarians or trial search co-ordinators; Meert et al. showed librarian co-authorship correlates with higher scores on the searching component of reviews. (2) Aims and purpose — five documents use "thorough", "comprehensive", "transparent", "reproducible"; the social-sciences guide offers "the aim of the literature search is not to retrieve everything. It is to retrieve everything of relevance"; three documents tie searching to minimizing bias. Egger et al.'s operational definition of a comprehensive trial search: not English-only, Cochrane CENTRAL or ≥2 other databases, plus ≥1 method for unpublished trials (conference abstracts, theses, trials registers, expert contact); Lam & McDiarmid found the average number of databases searched rose from 1 to 4 between 1994 and 2014. (3) Preparation — six documents; check existing/ongoing reviews (CDSR, DARE — frozen since 2015 — PROSPERO, MEDLINE) and pilot a preliminary strategy; text mining for term development noted as exploratory. (4) Designing the strategy — PICO dominant (five documents); NICE also endorses SPICE and multi-stranded approaches; BeHEMoTh and SPIDER for qualitative/theory questions; Hausner et al.'s conceptual vs objectively derived strategies; limits introduce bias and should be described. (5) Database searching — the "first step" in all nine documents; Cochrane's minimum is CENTRAL + MEDLINE +

EMBASE; no consensus on how many databases; same database on different hosts yields different returns (Younger & Boddy, AMED); searching PubMed in addition to Ovid MEDLINE improves currency. (6) Supplementary searching — endorsed across documents but with inconsistent ordering/extent; IQWiG treats handsearching as case-by-case; Cooper et al.'s own case study found supplementary methods out-yielded database searching for an environmental review. (7) Managing references — five documents; EndNote de-duplication (Bramer), Zotero, and Gall's finding that EndNote's built-in MEDLINE search is limited. (8) Documenting — only Cochrane names PRISMA explicitly and requires full strategies per database in an appendix; consensus items: databases + hosts, strategies, limits (CRD wants them justified); unique items: number of duplicates (CEE), screening decisions (IQWiG), grey-literature list (Campbell), annotation of tested-but-unused terms (NICE); PRESS is the peer-review instrument; Rader et al.'s survey of 260+ review authors found 86% (153/178) want more reporting guidance.

Conclusions and takeaways: The review concludes that a shared model exists and names it "the conventional approach": database searching first, supplementary methods second, thorough/transparent/reproducible aims throughout. It demonstrates consensus on primary vs supplementary search methods across all nine documents, but "little evidence for a consistent process" for supplementary methods. Two forward-looking arguments stand out: comprehensiveness may not be the right quality marker for all review types (qualitative synthesis, rapid reviews — citing Nussbaumer-Streit et al.'s 2018 finding that abbreviated searches are viable for rapid synthesis when decision-makers accept the trade-off), and preparation (stage 3) is the understudied decisive moment where structures, terms, and sources are chosen. The paper calls for research on whether the same process model suits qualitative evidence synthesis and on when to stop searching.

Limitations and caveats: The authors flag three limitations themselves: the guidance sample is deliberately UK/European/Australian ("Other handbooks exist" — AHRQ, IOM, CADTH excluded, though they believe inclusion would not change findings); the guidance documents may not be independent ("The handbooks are potentially linked to one another" — Cochrane is a reference source for others, risking circular consensus); and the supporting-studies stream "is not a systematic review of the evidence for each key stage," so relevant studies may have been missed. One cannot exclude that the "shared model" is partly an artifact of Cochrane's diffusion. All numbers cited are second-hand from the reviewed studies, not re-verified.

Practical implications for a non-academic researcher using PubMed and AI: Use the eight stages as a self-audit checklist for any evidence project: decide who searches (consider involving a librarian or, failing that, documenting your own expertise honestly); state your aim (retrieving everything of relevance, not everything); scope first (PROSPERO/Google for existing reviews); structure with PICO-like columns; search databases before supplements; log reference management and de-duplication; and — critically — document databases, hosts, dates, strategies and limits even if you never publish, because this is what makes your search checkable by others or by an AI assistant asked to verify it. The paper also warns against over-investing in exhaustiveness when a rapid, fit-for-purpose answer suffices.

Notable quotes:

- The thesis statement: "Information specialists and review teams appear to work from a shared and tacit model of the literature search process. How this tacit model has developed and evolved is unclear, and it has not been explicitly examined before." (Background)
- Quoted aim (from Petticrew & Roberts, in key stage two): "the aim of the literature search is not to retrieve everything. It is to retrieve everything of relevance."

- The structural conclusion: "Bibliographic database searching is consistently the first method of literature searching referenced in each guidance document." (Conclusions)

Connections: This paper is the meta-level synthesis of the batch: Aromataris (2014): Constructing a search strategy and searching for evidence: a guide's JBI manual is literally one of its nine audited documents, and its stages 2, 4 and 8 correspond to Aromataris's aims, logic-grid/PICO method, and PRISMA reporting advice. Watson (2020): How to undertake a literature search: a step-by-step guide's 10 steps are an informal novice rendition of the same process, while Blümle (2019): Systematic literature search in PubMed: a short introduction and Motschall (2005): Searching the MEDLINE literature database through PubMed: a short g… zoom into stage 4–5 mechanics (strategy design, database searching) for PubMed. Its discussion of comprehensiveness, bias and abbreviated searches directly frames where AI-driven rapid-synthesis methods (the AI papers elsewhere in this project) will be judged.

End of Batch 1 deep dives. All eight source texts were read in full; no section was omitted. Minor pdftotext artifacts noted inline where they affect interpretation (Elston title dash rendered as "d"; Blümle history-table hit counts partly misaligned; Aromataris logic-grid tables fragmented across the extraction).

14.2 Filters, quality control, and strategy-building tools

Booth (2016): Searching for qualitative research for inclusion in systematic reviews — Structured methodological review of searching for qualitative research (Systematic Reviews 2016)

What the paper is and why it matters: Booth (ScHARR, Sheffield; Cochrane Qualitative and Implementation Methods Group) presents a structured methodological review of the evidence base for literature searching for qualitative evidence syntheses (QES). Written to underpin updated Cochrane searching guidance, it catalogues methodological guidance, innovative approaches and empirical studies of retrieval for qualitative research. It matters because it is the definitive statement — at the time — that most searching practice for qualitative studies rests on "custom and practice" rather than empirical evidence, and it names the distinctive challenges (non-standard terminology, poor titles/abstracts, sampling debates) that make qualitative searching fundamentally different from trial searching. It also provides the "7S" framework scaffolding used by many later papers.

Methods in detail: The author exploited his 7-year role maintaining the Cochrane Qualitative and Implementation Methods Group study register: a sensitive keyword search of that register (terms such as "search*", "retriev*", "database*") retrieved 654 references. Separately, 15 "citation pearls" (key methodological papers, including Booth's STARLITE, Dixon-Woods, Campbell's meta-ethnography evaluation, the McMaster Hedges qualitative filters) were searched in Google Scholar for citing articles, yielding 1,063 references (with duplicates). Combined 1,717 references; 1,299 unique records were screened by a single reviewer in July/August 2015. Papers were coded (one or more categories) under an eight-part framework: overviews and guidance, sampling, sources, structured questions, search procedures, search strategies/filters, supplementary strategies, and standards — the "7S structure." Data were extracted into a Google Forms spreadsheet with narrative text extracts; no formal quality appraisal was attempted because designs were too heterogeneous — designs were characterised instead (case studies, comparative studies, surveys, validation studies).

Key results and numbers: 113 items were included: 46 overviews, 13 formal guidance papers, 3 narrative reviews of the search process, 1 summary, 7 hybrid overview+case-study papers, and 43 formal study designs (case study n=25; comparative n=7; literature surveys n=5; multiple case studies n=1; survey n=1; validation studies n=4). Distribution across 7S components: Sampling 47, Sources 22, Structured Questions 17, Search procedures 6, Search Strategies/Filters 16, Supplementary Strategies 24, Standards 17 (aggregate >131 due to multi-coding). Notable numbers: Dixon-Woods et al. found 23% of records screened for a breastfeeding qualitative review had no abstract; Wright et al. found QES searched 3–20 databases (37% searched 3–5); a CAM meta-ethnography searched 67 databases yet found 87% of included studies in PubMed alone; Greenhalgh & Peacock's audit found only 30% of included studies came from databases and hand-searches, ~50% from snowballing, 24% from personal knowledge/contact; 81% of published meta-ethnographies used exhaustive searches despite meta-ethnography's purposive-sampling rationale; Shaw et al. found only 4% of retrieved papers relevant at full-text screening; a preferred QES size of "between 6 and 14 studies" and Campbell's ceiling of "about 40 papers" are cited. CINAHL's "qualitative studies" heading dates from 1988 vs MEDLINE's much later Qualitative Research term (the text inconsistently says 2002 and 2003).

Conclusions and takeaways: The review concludes that "poor empirical evidence underpins current information practice" for qualitative searching; sampling should be matched to the synthesis method (comprehensive for meta-aggregation, theoretical/purposive for meta-ethnography); MEDLINE+CINAHL are a minimum for health topics; a parsimonious filter (qualitative OR findings OR interview*) may suffice where extensive supplementary searching is used; reference checking should be "a default for every review"; and ENTREQ/STARLITE/PRISMA should guide reporting, though ENTREQ uptake was low (1 of 19 meta-ethnographies). It proposes a research agenda across all 7S components.

Limitations and caveats: Single-reviewer screening and coding (acknowledged risk); inclusion required search/retrieval terms in titles/abstracts or citation links, so reviews with unlabelled searching innovations may have been missed. Quality assessment was abandoned for design-typing only. Much guidance is derivative of a few heavily co-cited key texts, giving a false impression of consensus. Case studies dominate (n=25) and lessons may not transfer across topics; database yields were shown to be topic-specific. The review was compiled to inform guidance the author was co-writing — a declared dual role.

Practical implications for a non-academic researcher using PubMed and AI: For PubMed-only qualitative searches, the core lesson is that a simple free-text cluster — qualitative OR findings OR interview* —

plus broad subject terms often outperforms long specialist filters, and that supplementary tactics (citation chasing, Google Scholar, reference lists, contacting authors) matter more than adding databases. AI tools can help enact the "expansive" iterative searching and citation-pearl-growing approaches Booth describes (generating synonyms, clustering results), but the sampling decision (comprehensive vs saturation) remains a human methodological choice that AI cannot make transparently for you.

Notable quotes:

- "searching for, and identifying, appropriate qualitative research is characterised as 'frustrating and difficult'"
- "Campbell et al. affirm that 'omission of some papers is unlikely to have a dramatic effect on the results'"

- "exhaustive searching is improbable… the obligation on reviewers is to plan a thoughtful and clearly described plan [sic] to locate the sample of studies most likely to answer their research question[s] reliably."

Connections: Booth's review is the qualitative-studies counterpart to the filter-performance Cochrane methodology reviews in this batch — Beynon (2013): Search strategies to identify diagnostic accuracy studies, Li (2019): Search strategies to identify observational studies in MEDLINE and … and Escobar Liquitay (2023): Search strategies (filters) to identify systematic reviews in MEDLI… — which use gold-standard testing rather than structured overview. It supplies background for McGowan (2016): PRESS Peer Review of Electronic Search Strategies: 2015 Guideline S…'s PRESS quality agenda, and the qualitative-filter performance figures in its Table 5 (McMaster Hedges sensitive/specific/optimal filters) feed the broader filter-evidence base Lefebvre (2017): Assessing the performance of methodological search filters surveys. Its warning that specific index terms ("grounded theory," "ethnography") yielded only false hits contrasts with the text-mining tools trialled in Dullea (2026): Search strategies built with the PubReMiner text-mining tool vs cur… and Paynter (2021): Evidence synthesis search strategies with and without text-mining t…, which automate term discovery — yet both find automated term choice does not yet match expert searching.

McGowan (2016): PRESS Peer Review of Electronic Search Strategies: 2015 Guideline S... — PRESS 2015 guideline: peer review of electronic search strategies (J Clin Epidemiol 2016)

What the paper is and why it matters: This guideline statement updates and expands the original (2008–2010) PRESS publications, providing an evidence-based structured tool for peer review of electronic search strategies — especially the "primary" (usually MEDLINE) search for systematic reviews and HTAs. It matters because search strategies frequently contain silent errors, and PRESS is the de facto international standard for catching them; it institutionalises a quality-assurance step in evidence synthesis that parallels peer review of the science itself. Every major later workflow in this batch — Dullea's conventional comparator, Paynter's peer-reviewed searches, Lefebvre's recommended relative-recall subject-search assessment — operationalises PRESS.

Methods in detail: The update used three evidence streams: (1) a systematic review update searching for checklists evaluating or validating electronic search strategies and elements relating to search quality or errors; (2) a web-based international survey of experts on the importance of search elements; and (3) a consensus development forum with leading experts drawn from CADTH, Cochrane Trial Search Coordinators, the Cochrane Information Retrieval Methods Group, a PRESSforum user, and HTAi's information-resources interest group. Decisions were reached by open (unblinded) voting with consensus defined a priori as a majority vote; in the event, all checklist items received unanimous votes. The guideline delivers four components: an updated PRESS 2015 Evidence-Based Checklist (six elements), six Recommendations for librarian practice, four Implementation Strategies, and a Guideline Assessment Form; a companion Explanation & Elaboration document gives detail.

Key results and numbers: The systematic review found no new search elements to add to the 2008–2010 checklist and no evidence refuting existing elements, and "suggested that structured PRESS could identify search errors and improve the selection of search terms." The survey included 117 experts who answered at least one question beyond demographics; 108 completed it fully. Most respondents felt peer review should occur after the MEDLINE search is prepared but

before translation to other databases. Of the seven original PRESS elements, six were retained — translation of the research question; Boolean and proximity operators; subject headings; text word searching; spelling, syntax and line numbers; limits and filters — and the seventh, "skilled translation of the search strategy to additional databases," was removed and left to searcher discretion. The motivating evidence base: a study of over 100 MEDLINE searches "revealed that most search strategies contained errors." Implementation strategies specify: peer review the primary search; one peer review is acceptable with a second in defined cases; document the process; recognise the reviewer by name; maximum turnaround of five working days. A worked acetaminophen/migraine case study embeds at least one error in each element (e.g., non-exploded MeSH, "Animals/ and Humans/" instead of NOT, wrong line numbers, an inappropriate RCT limit).

Conclusions and takeaways: Literature search strategies for knowledge syntheses "should be peer reviewed using a structured tool such as the PRESS 2015 Evidence-Based Checklist," which improves quality and comprehensiveness and reduces errors. Notably, the core elements of an effective Boolean search remained stable over ~10 years, supporting their durability. The guideline is deliberately scoped to the database search itself — grey literature, registers, citation searching and expert contact lie outside PRESS. The authors flag that text mining may eventually shift practice toward "maximizing recall without regard to precision," making the interface between Boolean searching and text-mining/AI a rich research area.

Limitations and caveats: Grading of the strength of recommendations was not performed because the update did not re-assess risk of bias of the original review's studies. The updated evidence review was restricted to health-science databases. Piloting of the revised guideline was undertaken by a single agency (CADTH), the study's sponsor — a potential conflict acknowledged in the funding statement. All authors had received CADTH funding. Consensus voting was unblinded; formal consensus-methods metrics (e.g., Delphi rounds, agreement statistics) were not reported. Generalisability to non-health disciplines is limited, as uptake appears confined to SR/HTA contexts.

Practical implications for a non-academic researcher using PubMed and AI: A solo researcher can use the public PRESS checklist as a self-review rubric before finalising a PubMed query: check concept translation against PICO, Boolean/nesting, exploded MeSH plus free-text synonyms, truncation and acronyms, syntax/line errors, and inappropriate limits/filters — with hits-per-line displayed. When an AI drafts a query, PRESS is exactly the checklist for auditing it; the six elements map neatly onto common LLM failures (vague concept translation, missing MeSH, wrong operators). Documenting the reviewer's name/credentials, as PRESS requires, also disciplines AI-assisted search reporting.

Notable quotes:

- "a study we conducted of over 100 MEDLINE searches revealed that most search strategies contained errors"
- "this example shows that a convincing-looking search can contain shortcomings that peer review can identify"
- "The name and credentials of the person who undertook the peer review should be reported."

Connections: PRESS is the quality-assurance backbone for this batch: Dullea (2026): Search strategies built with the PubReMiner text-mining tool vs cur… peer-reviews its conventional comparator with the PRESS checklist; Paynter (2021): Evidence synthesis search strategies with and without text-mining t… PRESS-reviews both UP and TMT searches; Lefebvre (2017): Assessing the performance of methodological search filters recommends using PRESS to assess subject searches when building relative-recall reference sets; Escobar Liquitay (2023): Search

strategies (filters) to identify systematic reviews in MEDLI…'s filter-development studies would have benefited from PRESS-style reporting standardisation, which its conclusions explicitly call for. McGowan's forward-looking note on text mining shifting toward recall-maximisation anticipates exactly the Paynter (2021): Evidence synthesis search strategies with and without text-mining t…/Dullea (2026): Search strategies built with the PubReMiner text-mining tool vs cur… evaluations of text-mining tools, and echoes the sensitivity-over-precision theme of Booth (2016): Searching for qualitative research for inclusion in systematic reviews for qualitative searching.

Escobar Liquitay (2023): Search strategies (filters) to identify systematic reviews in MEDLI... — Cochrane methodology review: filters to identify systematic reviews in MEDLINE and Embase (CDSR 2023, MR000054)

What the paper is and why it matters: This Cochrane methodology review systematically compiles empirical evidence on search filters designed to retrieve systematic reviews in MEDLINE and Embase. Systematic reviews are the raw material of guidelines, overviews and evidence maps, so identifying them efficiently is a recurring search task; yet there was "no guidance for finding the best filters for identifying systematic reviews." It matters because it both grades the available filters (some exceed sensitivity >90% and precision >10%) and documents how weak, outdated and inconsistently reported the underlying evidence is — a cautionary meta-message about the whole filter-development field.

Methods in detail: The authors searched MEDLINE, Embase, PsycINFO, LISTA and Science Citation Index from inception to January 2023, plus PubMed Related Articles, reference lists, and the InterTASC ISSG and McMaster websites. They included studies whose primary objective was development, evaluation or comparison of a filter to retrieve systematic reviews in MEDLINE, Embase or both. Two pairs of authors independently screened; two reviewers independently extracted data using the InterTASC ISSG Search Filter Appraisal Checklist. Performance measures (sensitivity, specificity, precision, NNR, accuracy) were calculated from 2×2 data where necessary; a priori acceptability thresholds were sensitivity >90% and precision >10% (adopted from Beynon 2013). Planned subgroup analyses (pre/post-PRISMA-2009, interface, topic) and secondary outcomes were largely abandoned due to scarce data; results were narratively synthesised per database and interface.

Key results and numbers: From 34,134 retrieved records (11,676 duplicates removed; 22,458 screened; 54 full texts), eight studies were included in 10 publications. Eight studies developed MEDLINE filters (Boynton 1998, White 2001, Shojania 2001, Wilczynski 2007, Lee 2012, Boluyt 2008, Avau 2021, Salvador-Oliván 2021); three developed Embase filters (Wilczynski 2007, Lee 2012, Avau 2021). Six studies reported sensitivity, seven precision, six specificity; only Shojania reported NNR and PPV. Three MEDLINE filter groups achieved >90% sensitivity with variable precision: Boynton 1998 (quasi-gold standard of 288 records; most sensitive strategy 99.5% sensitivity, 75.6% specificity, but precision as low as 0.1%); Shojania 2001 (externally validated: sensitivity 93% on DARE-derived and 97% on ACP Journal Club sets, specificity 97.2%, precision 33.2%, NNR 3.01); Wilczynski 2007 (Hedges: best sensitivity 94.6%; on the full 49,028-record validation database 100% sensitivity, 3.41% precision). White 2001 achieved the highest specificity (>90%): its optimised strategy reached 92.3% sensitivity, 87.7% specificity, 5.6% precision; a "balanced query" reached 98% sensitivity, 90.8% specificity, 14.2% precision. Topic-specific filters: Lee 2012 (public health) 86.8% MEDLINE sensitivity, 1.1% precision; Avau 2021 (first

aid) PubMed recall 90%, specificity 97%, precision 9.7%; Embase 91%, 96%, 5.4%. Salvador-Oliván 2021′s precision-maximising filter showed 83.8% precision (72.3–96.7%) and 91.6% recall, but found the stock PubMed SR filter retrieves only 62% of its output. Gold standards were mostly small (70–387 records) except Wilczynski's (27,769; 10,446; 49,028). None of the filters distinguish high- from low-quality reviews.

Conclusions and takeaways: MEDLINE filters showed broadly similar sensitivity and precision, with White 2001 best on specificity; Embase filters showed variable performance with poorly documented development. The authors highlight three filter groups "that have been independently evaluated and reached our pre-defined threshold" (Boynton 1998; Shojania 2001; Wilczynski 2007) for priority use, while acknowledging they predate PRISMA 2009 — which changed how reviews are titled and indexed and may have shifted filter performance. Key recommendations: consensus guidance and standardised reporting for filter development; interface-adaptable strategies; re-evaluation of existing filters against a common reference set; and attention to built-in article-type filters and resources like Epistemonikos whose performance is unmeasured.

Limitations and caveats: Only eight studies, most developed 1995–2008 and over 15 years old, raising transferability concerns to current platforms and indexing. Three filters were topic-specific (first aid, child health, public health), limiting generic applicability; several used quasi-gold standards precluding specificity estimates; gold standards were small and inconsistently reported; unblinded classification of records as "systematic reviews" may bias performance estimates. The review could not perform planned subgroup or secondary analyses due to data scarcity. The companion Navarro-Ruan comparison of SR[pt] versus the Wilczynski filter used unblinded classification and partial verification, further softening the evidence.

Practical implications for a non-academic researcher using PubMed and AI: If you need a defensible set of systematic reviews for an umbrella-style task, the practical route is PubMed's built-in "Systematic Article" filter plus MeSH publication type — but this review shows that filter's own sensitivity was only informally validated, so combine it with a tested hedge such as the Shojania/PubMed strategy. An LLM can assemble these validated filter strings, but the core lesson is that generic "systematic review" phrasing varies over time; ask the AI to check term currency and to log which filter version you used and when.

Notable quotes:

- "Although many had acceptable sensitivity … and specificity …, no single filter can be recommended since most were derived from older sets of reviews that may not reflect current reporting characteristics and standards."
- "None of the filters were designed to differentiate systematic reviews on the basis of their methodological quality."
- "we found highly heterogeneous development methods, accuracy assessments and outcome selection."

Connections: This review is the systematic-reviews sibling of Beynon (2013): Search strategies to identify diagnostic accuracy studies (DTA studies) and Li (2019): Search strategies to identify observational studies in MEDLINE and … (observational studies) — all three Cochrane methodology reviews sharing the same methods template, ISSG appraisal checklist and the Beynon-derived >90%/>10% threshold. It cites Booth (2016): Searching for qualitative research for inclusion in systematic reviews (excluded as qualitative-focused) and Lefebvre (2017): Assessing the performance of methodological search filters as guidance for assessing filters, and its conclusions about heterogeneous reporting echo Lefebvre's reviews A/B. Its discussion of machine-learning filter development frames the empirical tool evaluations in Paynter (2021):

Evidence synthesis search strategies with and without text-mining t… and Dullea (2026): Search strategies built with the PubReMiner text-mining tool vs cur….

Beynon (2013): Search strategies to identify diagnostic accuracy studies — Cochrane methodology review: filters for diagnostic accuracy studies (CDSR 2013, MR000022)

What the paper is and why it matters: This Cochrane methodology review assesses the evidence on methodological search filters intended to identify diagnostic test accuracy (DTA) studies in MEDLINE and EMBASE. DTA studies are notoriously hard to find — methods are poorly reported in titles/abstracts and indexing is inconsistent — so many groups designed filters to boost precision when combined with target-condition and index-test terms. This review is the empirical arbiter: it shows that no published filter combines the high sensitivity needed for systematic reviews with acceptable precision, and it establishes the influential benchmark (sensitivity >90%, precision >10%) reused across the filter-evidence field.

Methods in detail: The review team searched MEDLINE (1950–Nov 2012), EMBASE (to 2012 week 48), the Cochrane Methodology Register, Web of Science, PsycINFO, LISA and LISTA, plus citation searches, reference-list checks, and the ISSG Search Filters Resource website. Eligible studies reported development or evaluation (or both) of a MEDLINE or EMBASE filter aimed at retrieving DTA studies, with a measure of filter performance. The main outcome was filter performance (sensitivity, specificity, precision, NNR with 95% CIs where 2×2 data allowed). Data extraction covered reference-set identification, term identification and combination, internal/external validity testing, number of filters evaluated, dates, and databases/interfaces used. Performance reported in development studies was compared with that from subsequent evaluation studies of the same filters.

Key results and numbers: Nineteen studies were included, reporting 57 MEDLINE filters and 13 EMBASE filters (70 filters total per the plain-language summary); 30 MEDLINE and 4 EMBASE filters were tested in evaluation studies against gold standards. All 19 studies reported sensitivity, nine specificity, fourteen precision. No filter that had original development performance data and was later tested in an evaluation study met the a priori acceptable sensitivity (>90%) and precision (>10%). For MEDLINE filters developed in one study and evaluated in others (n=13), development-study sensitivity ranged 55–100% (median 86%), specificity 73–98% (median 95%), precision 1.1–40% (median 9.5%); evaluation studies of the same 13 filters showed lower estimates: sensitivity 14–100% (median 73%), specificity 15–96% (median 81%), precision 0.2–16.7% (median 4%). MEDLINE filters without an original performance measure: sensitivity 31–100% (median 71%), specificity 13–90% (median 55.5%), precision 1.0–11.0% (median 3.35%). For EMBASE, two development studies reported three filters with sensitivity 74–100% (median 90%) and precision 1.2–17.6% (median 3.7%); evaluation studies gave sensitivity 72–97% (median 86%) and precision 1.2–9% (median 3.7%). EMBASE development-vs-evaluation performance was more consistent than MEDLINE's, but no EMBASE filter met both thresholds either.

Conclusions and takeaways: The headline conclusion: "None of the current methodological filters designed to identify reports of primary DTA studies in MEDLINE or EMBASE combine sufficiently high sensitivity … with a reasonable degree of precision." This supports the Cochrane DTA Handbook's recommendation that a methodological filter should not be added to target-condition and index-test searches as the only search approach. The review also shows filters performing consistently worse when re-tested on new data than in their original development

sets — a systematic overestimation pattern. Practical message: use the disease × index-test combination plus supplementary searching, rather than trusting any single filter.

Limitations and caveats: Reported outcome measures varied across studies; some development studies reported no performance data at all (original data unavailable for 17 filters later tested in evaluations); reference sets and their limitations were described inconsistently; and filter-development methods, term selection and validation practices were heterogeneous. Extraction relied on 2×2 data availability, so some comparisons could not be made. Estimates from relative-recall–type reference sets may misclassify non-gold-standard records, and precision estimates depend heavily on topic prevalence.

NOTE on the source text: The pdftotext extraction of this paper is severely truncated — the file contains only the title page, structured abstract and plain-language summary (roughly 60 lines), with no methods, results, discussion, study-characteristics or reference sections. All analysis above is based on the abstract and summary only; numbers are faithful to that text but far less detail is available than for the other papers in this batch.

Practical implications for a non-academic researcher using PubMed and AI: When searching for diagnostic accuracy studies on PubMed, don't rely on a design filter alone: combine your disease/index-test concepts, consider Clinical Queries' diagnosis subset for a first pass, and use reference chasing and citation searching for the remainder. If an AI proposes a published DTA filter, ask for the filter's validated sensitivity — this review's medians (≈73% on re-testing) mean roughly one in four relevant studies may be missed. PubMed's "sensitivity and specificity" MeSH terms help, but indexing of DTA studies remains inconsistent.

Notable quotes:

- "No filter which had original performance data from its development study, and was subsequently tested in an evaluation study, had what we defined a priori as acceptable sensitivity (> 90%) and precision (> 10%)."
- "The results showed that search filters do not perform consistently, and should not be used as the only approach in formal searches to inform systematic reviews of diagnostic studies."
- "None of the filters reached our minimum criteria of a sensitivity greater than 90% and a precision above 10%."

Connections: Beynon 2013 is the methodological template that Escobar Liquitay (2023): Search strategies (filters) to identify systematic reviews in MEDLI… and Li (2019): Search strategies to identify observational studies in MEDLINE and … explicitly adapt (identical threshold, appraisal checklist, development-vs-evaluation design), and its central finding — filters lose performance on re-testing — is echoed in Li's overestimation warning. Its DTA-filter focus is one of the four study designs synthesised in Lefebvre (2017): Assessing the performance of methodological search filters reviews A and B. The poor title/abstract reporting of DTA studies parallels the qualitative-reporting problems catalogued by Booth (2016): Searching for qualitative research for inclusion in systematic reviews, and both explain why automated approaches like those trialled in Paynter (2021): Evidence synthesis search strategies with and without text-mining t… and Dullea (2026): Search strategies built with the PubReMiner text-mining tool vs cur… struggle: the words signalling a study type are absent or idiosyncratic.

[[Li (2019): Search strategies to identify observational studies in MEDLINE

and …]] — Cochrane methodology review: filters for observational studies in MEDLINE and Embase (CDSR 2019, MR000041)

What the paper is and why it matters: Observational studies underpin reviews of aetiology, risk, prognosis and adverse events, yet — unlike RCTs — there is no mature, validated retrieval machinery for them, because of poor and inconsistent indexing (STROBE notwithstanding). This Cochrane methodology review assembles the (thin) evidence on methodological filters for observational studies in MEDLINE and Embase. It matters because it quantifies just how sparse and dated the evidence is (two eligible studies, both from 2006), and because it shows that where filters do work well (one study), they can achieve >99% sensitivity with useful precision — an existence proof that filter development for observational designs is feasible.

Methods in detail: The review searched MEDLINE (1946–April 2018), Embase (1974–April 2018), CINAHL (1937–April 2018), the Cochrane Library (1992–April 2018), Google Scholar (first 500 records of three searches) and OpenGrey, and scanned reference lists and contacted authors. Included studies used a relative-recall approach — comparing a strategy with a methodological filter against a reference standard derived from systematic reviews — or compared two or more filters. Two authors independently screened, extracted (data on filters, records retrieved, sensitivity/precision, reference-standard methods) and appraised each included study using the UK InterTASC ISSG Search Filter Appraisal Checklist. A narrative synthesis tabulated sensitivity and precision by database (MEDLINE, Embase, combined) and by development vs validation systematic reviews.

Key results and numbers: Only two studies met inclusion, reporting 18 filters (six MEDLINE, six Embase, six combined), all developed using terms from the reference-standard records. Study 1 (Fraser 2006) targeted observational studies of surgical interventions: its reference standard was 217 articles from one LASIK systematic review (83.9% case series; 206 MEDLINE-indexed, 191 Embase-indexed). Its six filters (Precision Terms and Specificity Terms Filters per database) achieved sensitivity 99.5–100% with precision 16.7–21.1%; the Specificity Terms Filter for combined MEDLINE/Embase was preferred (precision 17.1% at 100% sensitivity). External validation on two further reviews (PRK, n=39; tonsillectomy, n=30) retrieved 85.2–100% of the validation sets, with 21.9–39.6% reduction in retrievals. Study 2 (Furlan 2006) targeted comparative non-randomised studies (cNRSs): 12 filters (Fixed method A, Fixed method B, Progressive method (CV), Progressive method (CV or TW)) built from four Cochrane reviews totalling 89 cNRSs achieved much lower sensitivity (48–100%) and precision (0.09–4.47%). Aggregated: MEDLINE filters sens 48–100% (mean 80.94%, SD 14.59), precision 0.12–21.10% (mean 3.55%); Embase sens 81–100% (mean 93.67%), precision 0.08–18.8% (mean 2.50%); combined sens 76–100% (mean 94.25%), precision 0.10–17.1% (mean 1.64%). Most sensitive individual terms: (compare$ or compara$).mp (65.5%); cohort.tw had the best precision (41.7%). Handsearch-based Hedges comparisons elsewhere show 80–100% sensitivity but precision generally below 10%.

Conclusions and takeaways: The evidence for observational-study filters is "limited and heterogeneous": the Fraser filters (surgical, one review) are strong — high sensitivity, decent precision, externally validated — while the Furlan cNRS filters are weaker, unvalidated externally, and topic/method specific. Both studies are >10 years old and indexing and labelling of observational studies have since changed, so the authors call for re-evaluation, robust external validation on large multi-topic sets, evaluation of currently used but unvalidated filters (BMJ Best Practice, ISSG, SIGN, UTHealth), and better indexing of observational designs. The relative-recall approach is endorsed as an economical alternative to handsearching, though it risks confounding.

Limitations and caveats: Only two included studies; Fraser used a single topic (LASIK) with a case-series-dominated reference set (83.9% case series), limiting generalisability; Furlan's four reference standards totalled fewer than the ~100 articles suggested for stable estimates. Both

studies derived filter terms from the same records used to test the filters, which the authors flag as possible overestimation. One Furlan topic-only strategy reached only 85% sensitivity, potentially deflating paired filter estimates. Validation was done for only one of 18 filters. All testing used Ovid, leaving PubMed/other interfaces untested.

Practical implications for a non-academic researcher using PubMed and AI: For observational-study searching on PubMed, combine MeSH (Cohort Studies, Case-Control Studies, Cross-Sectional Studies, Epidemiologic Studies) with proven high-yield free text (cohort, compared/comparative, prospective, retrospective, follow-up) — the individual-term data show these carry most of the signal. An AI assistant can draft such a filter, but note the review's warning that filter performance degrades over time and across interfaces: treat any generated filter as a draft to test against known relevant papers. Expect low precision; a good observational filter here still means reading ~5–6 records per relevant one.

Notable quotes:

- "identifying observational studies as part of systematic reviews efficiently is challenging due to poor and inconsistent indexing in literature databases."
- "The six methodological filters developed from the first study reported sensitivity of 99.5% to 100% and precision of 16.7% to 21.1%."
- "Both studies were published 10 years ago and labelling and indexing of observational studies has changed since then."

Connections: Li 2019 completes the triad of Cochrane filter reviews with Beynon (2013): Search strategies to identify diagnostic accuracy studies (DTA) and Escobar Liquitay (2023): Search strategies (filters) to identify systematic reviews in MEDLI… (systematic reviews), sharing their ISSG appraisal checklist, relative-recall/gold-standard framing and cautious conclusions about filter maturity; Escobar's discussion cross-cites it. Its observational-filter evidence is exactly the gap that interviewees in Lefebvre (2017): Assessing the performance of methodological search filters asked to be filled ("Tested filters for observational studies"). Its message that indexing (not just terms) is the bottleneck parallels Booth (2016): Searching for qualitative research for inclusion in systematic reviews's qualitative-indexing critique, and its emphasis on testing filters against known items relates to the seed-set-based tools evaluated in Dullea (2026): Search strategies built with the PubReMiner text-mining tool vs cur… (PubReMiner PMIDs) and Paynter (2021): Evidence synthesis search strategies with and without text-mining t… (seed citations).

Lefebvre (2017): Assessing the performance of methodological search filters — HTA monograph: five literature reviews and a qualitative study on search-filter performance (Health Technol Assess 2017;21(69))

What the paper is and why it matters: This NIHR HTA monograph (Lefebvre, Glanville et al., commissioned to inform NICE methods development) is the most comprehensive methodological investigation into how search filters are built, tested, compared, chosen and reported. It combines five literature reviews (filter performance reporting; filter comparisons; DTA-methods analogies; how searchers choose filters; how clinicians choose tests) with interviews and a questionnaire of expert searchers, culminating in a proposed standard approach and reporting pro forma for filter performance measurement. It converts scattered filter research into actionable

guidance and reveals how information specialists actually work — including their reliance on "gut feeling."

Methods in detail: Five reviews were conducted in 2010/11: Review A examined 23 single-filter studies across four designs (economic, diagnostic, systematic review, RCT), extracting reference-standard construction, term identification and performance measures; Review B examined 18 studies comparing multiple filters (2–38 each), coding methods, translation, testing and display; Review C reviewed 47 reports of DTA methods to transfer diagnostic-accuracy principles (representative sampling, spectrum bias, sample-size tables, ROC/forest/ROC-space plots, precision-vs-specificity in skewed data) to filter testing; Review D searched for studies on how searchers choose filters (finding none); Review E reviewed 7 studies on clinician test choice. Additionally, 12 semi-structured interviews (16 information professionals, NICE-family organisations, Jan–Mar 2009) and a web questionnaire via seven email lists plus Twitter/Facebook (July–Aug 2011) gathered searcher practices. Findings fed a suggested measurement approach: hand-searched or relative-recall reference sets; sensitivity, precision/NNR, and "reduction in number needed to read"; a reporting pro forma for the ISSG website.

Key results and numbers: Review A: 20 of 23 studies used internal reference standards (17 derived by hand-searching, 2–161 journals; standards of 58–1,587 records); only 9 of 23 studies externally validated filters. Sensitivity was reported by all 19 internally validated studies, precision by 16, specificity by 14, accuracy by 7, NNR by 4; results almost always in tables. Review B: 16 of 18 studies reported sensitivity/recall, 13 precision, 7 specificity; gold standards ranged 33–1,955 records; 14 studies tested in MEDLINE; no numerical synthesis methods were identified; only 2 studies presented results graphically; in every study reporting new-filter development with comparison, the new filter "performed better" — a bias red flag. Review C: precision rather than specificity better reflects false-positive changes in skewed retrieval (filters A/B/C at 83% sensitivity with precision 25%/12.5%/4.6% but specificity 92%/82%/46%); sample-size tables suggest 70 relevant records plus 1,400 records at 5% prevalence (7,000 at 1%). Review D: 2,266 records identified, 837 deduplicated — zero eligible studies on filter choice. Review E: 7 studies; test performance the main cited factor (Sox et al.: 73%/71% of clinicians given test-characteristics information chose the best test vs 21% of controls). Interviews: ISSG Search Filters Resource most-mentioned source; Cochrane RCT and McMaster Hedges filters most cited. Questionnaire (n=90): 94.4% had used filters; 76.7% for extensive searches; 81.1% choose different filters for sensitivity vs precision; 55.6% sometimes amend filters (83.3% of amenders test, 75.9% document); 86.7% describe filters used, but 57.8% do not justify their choice.

Conclusions and takeaways: Filter studies should use representative, adequately sized reference standards, explicit methods and external validation; performance measures should be reported systematically (sensitivity + precision/NNR as core; "reduction in number needed to read" as a workload metric; precision over specificity for skewed retrieval) and displayed graphically (paired forest plots, ROC-space plots) rather than only in dense tables. Searchers value filters mainly for reducing record volume and focusing single-study-type searches, but choose them informally — often via provenance or interface defaults — and want synopses, ratings and validated filters for more designs and databases. The monograph predicts semantic/textual-analysis approaches superseding Boolean filters.

Limitations and caveats: A serious delay between the work (searches 2010/11) and publication (2017), acknowledged as regrettable due to the principal investigator's serious illness — the evidence base predates PRISMA-era reporting and all modern automation. Review D found no literature on filter choice by design. The questionnaire had no calculable response rate, ambiguous resource naming, and self-selection toward expert searchers. Interviewees were NICE-family staff,

possibly not generalisable, and senior staff no longer searching daily. Reviews A/B covered only four study designs; review B was not a full systematic review. Exhibit tables in the pdftotext extraction are partially garbled/column-scrambled.

Practical implications for a non-academic researcher using PubMed and AI: The monograph's toolkit is directly usable: build a quick test set of known-relevant records (from a prior review or your own scoping), run candidate filters, and compute sensitivity, precision and reduction-in-records-to-screen yourself. Use PubMed's built-in Clinical Queries/Hedges buttons knowingly — they were the most-used source — but check "what's under the hood" for reporting. Ask an AI to generate two or three filter variants and evaluate them against your test set as the pro forma prescribes; this converts gut-feeling filter choice into a documented mini-experiment.

Notable quotes:

- "Search filter studies should use a representative reference standard and explicitly report methods and results."
- "Search testing is pragmatic/intuitive rather than being a formal scientific process (there is not enough time to do this)."
- "Provenance – I judge according to who developed the filter."
- "please accept search filters are out-of-date methodology. Please don't develop more filters!"

Connections: This monograph is the methodological hub for the batch: its Reviews A/B include the very studies assessed by Beynon (2013): Search strategies to identify diagnostic accuracy studies, Li (2019): Search strategies to identify observational studies in MEDLINE and … and Escobar Liquitay (2023): Search strategies (filters) to identify systematic reviews in MEDLI…, and it supplies the rationale for their shared ISSG appraisal checklist and threshold logic. Its recommendation to assess subject searches with PRESS is realised in McGowan (2016): PRESS Peer Review of Electronic Search Strategies: 2015 Guideline S…, and its DTA-style testing framework underlies the sensitivity/precision/NNR outcomes in Paynter (2021): Evidence synthesis search strategies with and without text-mining t… and Dullea (2026): Search strategies built with the PubReMiner text-mining tool vs cur…. Its foresight about semantic analysis and "sliding scale" sensitivity foreshadows LLM-based search generation.

Dullea (2026): Search strategies built with the PubReMiner text-mining tool vs cur… — SWAR: PubReMiner text-mining strategies vs conventional searching (Cochrane Evidence Synthesis and Methods 2026)

What the paper is and why it matters: This Study Within A Review (SWAR) prospectively compares MEDLINE search strategies built solely from PubReMiner word-frequency tables against conventional librarian-built, PRESS-peer-reviewed strategies across twelve consecutive reviews at Ireland's HIQA HTA directorate. It matters because it is a rare multi-review, real-world evaluation of an AI/text-mining tool for search-strategy development — directly addressing the Priority III rapid-review research priorities — and its conclusion (inconsistent, generally inferior sensitivity; not recommended as a replacement) is a sober counterweight to hype about automating search building.

Methods in detail: Twelve reviews (of 15 conducted September 2023–January 2025; three excluded: non-in-house search, Ovid-only need, grey-literature-only obesity search) were recruited consecutively and registered in the SWAR repository (SWAR 26). The conventional comparator was each review's in-house comprehensive MEDLINE (EBSCOhost) strategy built by an experi-

enced librarian (>20 years) and externally peer-reviewed with the PRESS checklist. Two external librarians (also >20 years' experience) independently built PubReMiner strategies from word-frequency tables generated from PMIDs supplied by review teams, following a pre-specified work instruction refined through two pilot reviews, conducted within one week of the conventional search. The reference standard was each review's complete multi-source search and its final included-study set. Primary outcomes: sensitivity and precision; secondary: NNR, unique references, construction time. Records matched by DOI with title-based and manual fallbacks (mean missingness 4.80%). Analysis in R: Wilcoxon signed-rank tests (Holm correction), Friedman and Nemenyi tests; no meta-analysis due to heterogeneity; McNemar/Gwet's AC1 not applicable as no 2×2 table could be constructed.

Key results and numbers: Sensitivity of PubReMiner strategies was generally lower than conventional; in Review 5, PubReMiner 1 achieved 83.87% vs conventional 58.06% (PubReMiner 2: 38.71%), i.e., the automated search found included records the conventional MEDLINE search missed. The Friedman test showed a statistically significant difference across the three searches for sensitivity (p=0.0065 in the abstract; p=0.007 in the results text); Wilcoxon pairwise (Holm-corrected) showed conventional vs PubReMiner 2 significant (p=0.0428) but conventional vs PubReMiner 1 not (p=0.194); Nemenyi post-hoc found conventional vs PubReMiner 1 significant (p=0.028). Precision was low (<5% in all but Review 1) with no significant differences (Wilcoxon p=0.379; Friedman p=0.183). NNR showed no significant differences (p=0.307, 0.161, 1.000; Friedman p=0.183); Review 3 yielded no relevant records at all (sensitivity/precision 0%, NNR infinite). Unique references between the two PubReMiner librarians were not statistically different (p=0.339). PubReMiner searches were substantially faster to construct, but time was self-reported and compared only against the full comprehensive search. Inter-operator variability was marked despite identical PMIDs, protocols and work instructions.

Conclusions and takeaways: PubReMiner offers efficiency advantages but inconsistent, generally inferior retrieval performance; the authors explicitly "do not recommend that our approach of using PubReMiner for search strategy building be used." They situate the finding against the broader literature: a large LLM (Literature Search Sandbox) achieved 85% mean sensitivity but with IQRs of 40–100% and NNR 205–5,810, judged unusable without scrutiny; Spillias et al. found human–AI collaboration outperforms either alone. Key mechanistic insight: tools like PubReMiner are capped by the fields they mine (title/abstract/MeSH) — key terms may exist only in full text, which paywalls block. The paper advocates transparent evaluation of any statistical or AI tool before routine adoption and showcases multi-review SWARs as a design for methodological research.

Limitations and caveats: Only 12 reviews (3 excluded/edge-case), limiting power for precision and secondary outcomes; one review had zero relevant records, complicating pairwise statistics. Unique references found by PubReMiner were not screened for possible additional relevant studies, so sensitivity may be slightly understated. Efficiency comparisons are confounded (self-reported; comprehensive vs MEDLINE-only). PubReMiner depends on supplied PMIDs (seed quality unassessed). Results are specific to one centre's high-rigour, PRESS-reviewed conventional practice, so the gap may be smaller against weaker comparators. Statistical significance was borderline and test-dependent (Wilcoxon vs Nemenyi).

Practical implications for a non-academic researcher using PubMed and AI: PubReMiner-style frequency mining is fast (minutes vs hours) and, in one review here, surfaced studies the expert search missed — so use it as a supplementary term-discovery aid, never as your only search. If you have no librarian, the workflow is: gather seed PMIDs on your topic, mine frequent title/abstract/MeSH terms, build a strategy, then self-audit with PRESS and a

known-item test set. The paper's warning generalises to ChatGPT-style search building: expect wide, unpredictable variability between runs and users; verify sensitivity against a reference set before trusting an AI-built PubMed query.

Notable quotes:

- "While PubReMiner offers efficiency advantages, its inconsistent performance in retrieving relevant studies suggests that it should not replace conventional search strategies."
- "important key words and MeSH terms may not be found in the fields (e.g., title, abstract and MeSH terms) that tools such as PubReMiner use to generate their frequency tables."
- "researchers should look for transparent evaluations of similar statistical or AI tools used in evidence synthesis before employing them in their routine research practices."

Connections: Dullea 2026 is the multi-review scale-up of Paynter (2021): Evidence synthesis search strategies with and without text-mining t…, which it repeatedly cites as a "previously reported smaller study" of text-mining tools with similar conclusions. It uses McGowan (2016): PRESS Peer Review of Electronic Search Strategies: 2015 Guideline S…'s PRESS checklist to quality-assure its comparator, and its outcome definitions (sensitivity, precision, NNR) follow the framework formalised by Lefebvre (2017): Assessing the performance of methodological search filters. Its discussion of LLM search generation links to the filter-validation tradition of Beynon (2013): Search strategies to identify diagnostic accuracy studies, Li (2019): Search strategies to identify observational studies in MEDLINE and … and Escobar Liquitay (2023): Search strategies (filters) to identify systematic reviews in MEDLI…, whose standard (>90% sensitivity, >10% precision) the PubReMiner strategies clearly fail; and its human-expertise message echoes Booth (2016): Searching for qualitative research for inclusion in systematic reviews's finding that searcher judgement remains central.

Paynter (2021): Evidence synthesis search strategies with and without text-mining t... — Prospective comparison of search strategies with and without text-mining tools (J Clin Epidemiol 2021)

What the paper is and why it matters: This prospective comparative study (AHRQ EPC Scientific Resource Center) is a head-to-head, real-world test of search strategies developed with freely available text-mining tools (TMTs) versus librarians' usual practice (UP), embedded in seven actual evidence syntheses. It quantifies the trade-off driving modern search-methods debates: TMT searches cut strategy-development time by more than half (5 vs 12 hours) at a modest sensitivity cost (84.9% vs 92%), with topic complexity determining who wins. It seeded — and is confirmed by — the later multi-review SWAR evidence on PubReMiner.

Methods in detail: Investigators (librarians/methodologists in the AHRQ EPC program) recruited review teams from AHRQ, the VA Evidence Synthesis Program and a Cochrane Review Group; nine projects were recruited (of 12 approached), seven analysed (five de novo SRs, one SR update, one evidence map; two excluded for protocol violation and timing). For each review, the team's librarian built a MEDLINE strategy by usual practice while a randomly selected investigator-librarian built one using TMTs, both simultaneously and independently on the same platform (legacy PubMed or Ovid MEDLINE) and date range. TMTs were chosen freely per category from a pre-agreed list: keyword/phrase tools (AntConc, PubReMiner) and subject-term tools (MeSH on Demand, PubReMiner, Yale MeSH Analyzer). Both searches were PRESS-peer-reviewed. The SRC research associates merged results, deduplicated by PMID, and the review

team screened; citations included in the draft report formed the reference standard. Librarians logged time in 15-minute increments; simple vs complex topics were prospectively classified; analysis was descriptive (means, SDs, 95% CIs).

Key results and numbers: Across all seven reviews, UP searches were more sensitive: 92% (95% CI 85–99) vs 84.9% (95% CI 74.4–95.4) for TMT; UP identified 5–19 additional relevant citations per project, but TMT searches found 1–4 unique relevant citations the UP searches missed. NNR was similar: UP 83 (SD 34) vs TMT 90 (SD 68). Development time: UP averaged 12 hours (SD 8) vs TMT 5 hours (SD 2). By subgroup — simple reviews (n=4): sensitivity UP 96% vs TMT 92.3%; on average one study found by UP was missed by TMT, and TMT found no unique articles; TMT's NNR was ~22 higher per relevant article (11–44 minutes extra double-screening). Complex reviews (n=3): sensitivity UP 87% vs TMT 75%; TMT's NNR was 14 lower per relevant article (7–28 minutes saved screening). Time savings were larger for complex reviews (8 hours) than simple (5 hours). Neither approach was perfectly sensitive.

Conclusions and takeaways: TMT-based searching "is not ready to be used as the sole process for developing SR searches," but the time savings and relatively high sensitivity make it "useful in reviews that do not require maximum sensitivity, such as rapid or scoping reviews," and useful adjunctively with UP to catch citations the usual process misses — though combining approaches raises the screening burden. The authors note their results were less favourable to TMTs than Hausner et al.'s "objective approach" (noninferior sensitivity using WordStat), possibly because they used free tools with little guidance, or because their UP searches were unusually rigorous. Future work: comparative tool functionality studies, larger samples, and best practices.

Limitations and caveats: Small sample (seven reviews) precluding inference; participation bias possible; TMT librarians were experienced searchers but relatively inexperienced with TMTs; comparisons were non-standardised (free tool choice, platform choice, inconsistent time-keeping); one update review reused an existing search, potentially skewing results. Sensitivity rests on the reviews' reference standards (draft-report included citations), which are imperfect. Descriptive statistics only; no adjustment for topic, review type or librarian skill. Generalisability is limited to EPC-style searches with trained librarians.

Practical implications for a non-academic researcher using PubMed and AI: If you must build your own PubMed search quickly, Paynter's data support a pragmatic protocol: feed 20–50 known-relevant PMIDs into free tools (Yale MeSH Analyzer for MeSH coverage; PubReMiner/AntConc for keywords), draft the strategy, then PRESS-audit it. Budget roughly 5 hours instead of 12 and expect ~8 percentage points lower sensitivity than an expert would achieve — acceptable for scoping, not for a definitive review. Always run a supplementary pass (citation chasing, related articles), because both approaches here missed relevant papers.

Notable quotes:

- "UP searches were more sensitive (UP 92% (95% CI, 85-99); TMT 84.9% (95% CI, 74.4-95.4)), with lower NNR (UP 83 (SD 34); TMT 90 (SD 68))."
- "For complex SR topics, TMT searches were faster and less sensitive than UP searches but identified unique eligible citations not found by the UP searches."
- "TMT is not ready to be used as the sole process for developing SR searches… this technology may be useful in reviews that do not require maximum sensitivity, such as rapid or scoping reviews."

Connections: Paynter 2021 is the direct precursor of Dullea (2026): Search strategies built with the PubReMiner text-mining tool vs cur…, which replicates its design across 12 reviews and reaches the same verdict (citing Paynter's follow-up on pros/cons of text-mining). Both evaluate exactly the tool family named in Lefebvre (2017): Assessing the performance of methodological

search filters's forecast of textual-analysis methods, and both measure sensitivity/precision/NNR in the framework Lefebvre formalises. Both PRESS-peer-review their searches per McGowan (2016): PRESS Peer Review of Electronic Search Strategies: 2015 Guideline S…. The finding that neither UP nor TMT is fully sensitive reinforces the supplementary-searching doctrine of Booth (2016): Searching for qualitative research for inclusion in systematic reviews, and weak performance on complex topics parallels the filter under-performance shown by Beynon (2013): Search strategies to identify diagnostic accuracy studies, Li (2019): Search strategies to identify observational studies in MEDLINE and … and Escobar Liquitay (2023): Search strategies (filters) to identify systematic reviews in MEDLI….

Note on source texts: all analyses are based strictly on the pdftotext extractions. Extractions of Beynon (2013): Search strategies to identify diagnostic accuracy studies *are severely truncated (abstract and plain-language summary only);* Lefebvre (2017): Assessing the performance of methodological search filters*'s exhibit tables and* Escobar Liquitay (2023): Search strategies (filters) to identify systematic reviews in MEDLI…*'s study-characteristics tables are partially garbled by column-scrambling;* Booth (2016): Searching for qualitative research for inclusion in systematic reviews *contains an internal inconsistency on the year of MEDLINE's Qualitative Research MeSH term (2002 vs 2003).*

14.3 Supplementary searching, reporting, and knowledge translation

Briscoe (2020): Conduct and reporting of citation searching in Cochrane reviews — Cross-sectional study of how backward and forward citation searching is conducted and reported in Cochrane systematic reviews (Res Synth Methods 2020)

What the paper is and why it matters: This is a cross-sectional audit of 215 Cochrane reviews published over a 3-month window (November 2016–January 2017), examining how two supplementary search techniques — backward citation searching (checking reference lists) and forward citation searching (using citation indexes to find studies citing a source study) — were actually conducted and described. It matters because citation searching is nearly ubiquitous in systematic reviews (a prior study found 81% of reviews used backward and 12% forward citation searching), yet it is thinly covered by formal guidance. The paper is one of the few to move beyond reporting mere prevalence, analyzing five concrete methodological decisions: which set of studies to start from, which citation index to use, manual versus index-assisted backward searching, use of multiple indexes, and non-standard approaches like snowballing. It also proposes a small reporting checklist extending the MECIR standards.

Methods in detail: The authors searched the Cochrane Database of Systematic Reviews using a wildcard in the Search All Text field, date-limited to November 2016–January 2017, exporting results to EndNote. Reviews were eligible if they explicitly described conducting citation searching (merely listing a citation index among databases was insufficient, since the index might have been keyword-searched). Reviews finding no studies at all, and withdrawn reviews, were excluded. Manual inspection of abstracts, methods and appendices was supplemented with Control-F keyword searches (backward, forward, citation, reference list, Web of Science, Scopus, Google). A data-extraction form was developed from the MECIR standards; one author extracted data from all reviews with a 10% subset cross-checked by a second author, discrepancies resolved

by discussion. Reviews not stating whether backward searching was manual or index-assisted were coded as manual, based on the team's institutional experience — an explicit, acknowledged assumption tracked separately. Categories for the "set of studies" used were key studies, included studies, identified studies, plus a non-exclusive systematic-reviews category.

Key results and numbers: Of 215 reviews, 17 were excluded (7 withdrawn, 10 finding no studies), leaving 198. Backward citation searching was reported in 172 (87%); forward in 18 (9%), spread across 14 Cochrane review groups; no review reported forward without backward. The set of studies used was reported by 92% of backward-searchers and 83% of forward-searchers. For backward searching, included studies dominated (94 of 159, 59%), with identified studies 39%, key studies only 2%; systematic reviews were used in 41%. For forward searching, included studies were used in 53% and key studies in 33% — a striking difference the authors attribute to the absence of a MECIR conduct standard for forward searching. No review used a citation index for backward searching; only 3 explicitly said the approach was manual. Among the 18 forward-searching reviews, Web of Science dominated (13 of 15 naming an index, including 7 using the Science Citation Index), Scopus 2, Google Scholar 1; one review used two indexes (Scopus + Web of Science). Only two reviews reported all the authors' suggested details (index name, set of studies, date); under MECIR alone, compliance was 98% (backward) and 89% (forward). Snowballing appeared in zero reviews.

Conclusions and takeaways: Backward citation searching was far more uniform than forward searching, plausibly because MECIR mandates backward searching of included studies (C30) but says nothing about forward searching. Reporting was largely MECIR-compliant, but compliance does not equal transparency: searchers rarely stated the manual approach, the set of source studies, or the date of forward searching. The authors recommend reporting the specific studies used when the set is narrower or broader than included studies, stating explicitly that a citation index was used for citation searching, dating forward searches, and reporting in the methods section rather than only in an abstract or PRISMA flowchart. They also suggest weighing manual versus index-assisted backward searching for its screening and de-duplication benefits, and note Google Scholar's value despite export problems.

Limitations and caveats: The central limitation is the assumption that silence on the approach means backward searching was manual — some authors may have used a citation index without reporting it (the authors found one such case in an earlier iteration of a review). The "identified studies" versus "included studies" distinction may be semantic rather than procedural, since many reviews use "identified" loosely; the true use of included studies may be higher. The 3-month sample was a convenience sample from an earlier web-searching study, limiting generality and currency. Data extraction of a 10% subset, and single-primary-extractor design, leave some room for coding error. The text also contains minor pdftotext jumbling (tables slightly scrambled), but the narrative is clean.

Practical implications for a non-academic researcher using PubMed and AI: PubMed has no true citation-chasing function (the paper notes a "Similar Articles" search is algorithmic similarity, not citation linking), so a solo researcher replicating backward citation searching should work manually from the reference lists of included studies or reviews — exactly what 87% of Cochrane reviews do. AI assistants can help triage long reference lists by extracting candidate references, but the paper's core lesson is transparency: record which studies you started from, which tool you used, and (for forward chasing via free tools like Google Scholar) the search date. Citation searching is especially valuable when core concepts are hard to keyword, and it complements rather than replaces database searching.

Notable quotes: "Citation searching uses the citation network surrounding a source study to identify similar studies." — "Reporting was generally compliant with MECIR, but this is not always sufficient to ensure full transparency." — "Google Scholar should not be dismissed as a useful tool, particularly for identifying grey literature, which has been estimated to comprise around half (48%-65%) of Google Scholar's content."

Connections: Directly operationalizes the supplementary-search theme shared with Godin (2015) — both are about systematic searching beyond databases — and complements Laynor (2024), whose protocol uses forward and backward citation chasing ("snowballing") as a deliberate, documented search component, precisely the non-standard approach Briscoe found absent in Cochrane reviews. Pollock (2022) shares the same JBI-style three-step search logic and reference-list screening. Briscoe's finding that compliance ≠ transparency echoes the transparency preoccupations of Luo (2025, GAMER) and Haby (2016) on rapid-review reporting. Godino (2023) is the practical record-management counterpart: exported citation results need deduplication, which Excel handles. Grimshaw (2012) supplies the rationale — systematic reviews, well searched, are the unit of knowledge translation.

Luo (2025): GAMER Statement: reporting guideline for generative AI tools in med... — GAMER Statement: international consensus reporting guideline for generative AI tools in medical research (BMJ EBM 2025)

What the paper is and why it matters: The GAMER (Generative Artificial intelligence tools in MEdical Research) Statement is the first universal reporting guideline specifically for generative AI (GAI) use across all phases and types of medical research. Existing AI guidelines (CONSORT-AI, SPIRIT-AI, DECIDE-AI, STARD-AI, TRIPOD-AI, TRIPOD-LLM, CLAIM, CANGARU) target either AI in general, specific study designs, or specific LLM applications; journal and institutional policies on GAI are inconsistent and often developed without rigorous process. GAMER fills that gap with a nine-item checklist produced via a scoping review, a two-round international Delphi, and online consensus meetings under EQUATOR Network methodology. It matters for anyone using ChatGPT-class tools — including for literature work — because it defines what transparent disclosure looks like.

Methods in detail: The team followed EQUATOR-recommended methodology. Four groups were formed (Advisory Committee, Core Team, Delphi Expert Group, Coordination Team), with the protocol published and the project registered on the EQUATOR website on 3 November 2023. Delphi experts were recruited two ways: a PubMed search on 3 October 2023 identifying 200 experts who had published on GAI in medicine, plus snowball sampling through collaborators, with attention to disciplinary and geographic diversity. Candidate items were generated from (1) existing AI reporting guidelines, (2) journal/publisher GAI policies, (3) a scoping review of how GAI tools are reported in the literature, and (4) literature recommended by the Advisory Committee. Panellists rated items 1–7 for inclusion; rules were: median 1–3 excluded, 4–5 (or 6+ with substantial comments) carried forward, median 6–7 with no substantial comments included. Two online Zoom meetings (30–31 May 2024; 14 and 12 participants) resolved disputed items; non-attendees gave feedback via an online form. An explanations-and-elaborations (E&E) document and glossary accompanied the checklist.

Key results and numbers: 51 experts from 26 countries participated overall; 44 (from 26 countries/regions) took part in the Delphi — 43 in round 1, 33 in round 2. Round 1 presented 7 items; all reached the inclusion threshold (median ≥6), but one (declaring who was responsible for

GAI use) was dropped as redundant given authors' collective responsibility, and another (date of use) was folded into the tool-specification item. Round 1 yielded 135 comments including 4 new items; 3 survived to the final checklist after round 2 (which produced 130 further comments). The final checklist has nine items: (1) general declaration of any GAI use (with N/A option), (2) tool specifications — name, version/release date, dates used, interface vs API, temperature/token settings, (3) prompting technique and unedited responses, (4) declaration of new/fine-tuned models and their base model, (5) the GAI tool's role in all study phases, (6) specific manuscript sections touched, (7) how generated content was verified and modified, (8) data privacy and confidentiality safeguards, (9) impact on interpretation of results and conclusions. Notably, the authors themselves disclose using ChatGPT 4o (V.gpt-4o-2024-05-13) on 1 September 2024 for language and style, verified by four named authors.

Conclusions and takeaways: GAMER gives a standardized, brief (nine-item) minimum reporting standard covering tool identity, role, prompts, verification, privacy, and interpretive impact, applicable to every study type and every research phase — not just manuscript writing. The panel deliberately chose the broad term "GAI tools" (over "LLMs") to future-proof the guideline for visual and multimodal models. Reporting location (mostly methods) is recommended, not mandated. The authors plan dissemination via translations, journal editor outreach, a dedicated website, and a standing coordination group reviewing the guideline annually. The checklist doubles as an appraisal tool: reviewers and readers can use it to judge whether GAI use was disclosed and verified.

Limitations and caveats: Women made up only 15% of the expert panel, an imbalance the authors attribute partly to recruiting first/last authors of GAI-related papers. No patient or public representatives were involved, on the stated rationale that their GAI familiarity might be limited — a decision that will draw criticism. Consensus thresholds were pre-set but the Delphi was modified rather than classic; response attrition (43→33) between rounds is modest but real. As with any reporting guideline, adoption by journals is voluntary and unproven; the checklist's brevity means some edge cases (e.g., translation-only tools, which are excluded) require judgment. The PDF extraction is generally clean, with minor hyphenation artefacts.

Practical implications for a non-academic researcher using PubMed and AI: If you use an LLM to summarize search results, draft screening text, or polish a report, GAMER is a ready-made disclosure template: name the tool and version, give the date, state its role, keep the prompts and unedited outputs, describe how you verified anything it produced (especially citations — the E&E shows an author fact-checking each GPT-generated citation), and state whether it could have changed your conclusions. Item 8 warns against pasting identifiable personal data into chatbots. For literature work specifically, the verification item is the most important: GAI-generated references must be checked against PubMed, not trusted.

Notable quotes: "GAMER provides universal and standardised guideline for GAI use in medical research, ensuring transparency, integrity and quality." — "Content directly generated by GAI tools may contain false or exaggerated information, so it is recommended to manually proofread, verify and, if necessary, revise the generated content." — "We used ChatGPT 4o … to revise the original draft written by XL and YC for language and style on 1 September 2024."

Connections: GAMER is the AI-transparency analogue of the reporting-transparency concerns in Briscoe (2020) and Haby (2016): all three argue minimum standards exist but fuller disclosure (approach used, shortcuts taken, prompts/verification) is needed for reproducibility. Laynor (2024) and Pollock (2022) are the protocol-writing mirror image — GAMER shows how to report AI use in the research itself, and its scoping-review-plus-Delphi pipeline resembles Pollock's planned tool-development pathway. Godin (2015) and Godino (2023) document the

manual labor GAMER-style tools may eventually assist, provided use is disclosed as GAMER requires. Grimshaw (2012) predates GAI but its "knowledge management" barriers (volume, access, appraisal skill) are exactly what GAI tools are marketed to solve — and exactly what GAMER's verification item polices.

Pollock (2022): Methodological quality, guidance, and tools in scoping reviews (pro... — Scoping review protocol on methodological quality, guidance, and appraisal tools for scoping reviews (JBI Evid Synth 2022)

What the paper is and why it matters: This is a published protocol (JBI Evidence Synthesis 2022;20(4):1098–1105, DOI 10.11124/JBIES-20-00570) for a scoping review that will identify and map evidence and tools for assessing the methodological quality or risk of bias *of scoping reviews themselves*. Scoping reviews have proliferated and repeatedly been shown to be conducted inconsistently, and although they are not required to appraise their included sources, the reviews themselves increasingly inform policy and future research — so their trustworthiness matters. Well-established appraisal tools exist for systematic reviews (AMSTAR/AMSTAR 2, ROBIS, CASP, JBI checklists), but the preliminary searches found only reporting checklists (PRISMA-ScR; Cooper et al.) and no appraisal tool for scoping reviews. This protocol is the groundwork for building one.

Methods in detail: The protocol follows JBI methodology for scoping reviews, reported per PRISMA-P (since scoping-review-protocol guidance did not yet exist), with PRISMA-ScR planned for the final review. A three-step search strategy was designed with a research librarian: an initial limited MEDLINE (PubMed) search on 20 May 2020 (yielding 5,105 results) to harvest keywords and MeSH terms; then full searches of MEDLINE (Ovid), Embase (Ovid), and CINAHL (EBSCO); then Google and Google Scholar searched in incognito mode, screening the first 200 results per Haddaway et al.'s advice. Eight target websites will be hand-screened (EQUATOR Network, CEBM, CRD York, CASP, McMaster, Evidence Synthesis International, JBI, Campbell Collaboration). Reference lists of included sources will be screened, the team will contribute sources from their own files, and stakeholders will be solicited via newsletters (JBI Scoping Review Network, SPOR Evidence Alliance, KT Canada) and a Twitter/LinkedIn call-out. Two reviewers will screen and extract independently, pilot-tested, in JBI SUMARI; extraction fields include tool name, audience, items, presentation format, biases assessed, development process, and validity/reliability evidence. No date or language limits beyond 2005 (Arksey & O'Malley) apply.

Key results and numbers: As a protocol, it reports planned rather than final results. Concretely reported figures: the initial PubMed search retrieved 5,105 results; a preliminary search identified exactly two checklists addressing reporting standards for scoping reviews (PRISMA-ScR and Cooper et al. 2021) and confirmed no existing or in-progress systematic or scoping review of appraisal tools for scoping reviews; the closest prior synthesis (Zeng et al.) covered preclinical/clinical studies, systematic reviews/meta-analyses, and guidelines but omitted scoping reviews. Five review questions are posed, covering existing tools/guidance, their domains, the bias types assessed, the validity/reliability testing undertaken, and the individual items included. Planned analyses include computing medians and interquartile ranges of item counts across tools, tabulating the frequency of bias types, and grouping similar items into a preliminary list that could seed a future tool.

Conclusions and takeaways: The protocol establishes that a gap exists: no validated tool appraises the methodological quality or risk of bias of scoping reviews, despite their growing policy use. It distinguishes methodological quality (design features like blinding) from risk of bias (whether findings should be believed) and notes the terms are used interchangeably in the literature, so all will be captured. The anticipated outputs are a comprehensive inventory of tools and guidance, a catalogued item pool, and — flagged explicitly — a future modified Delph
i to vote items toward a dedicated appraisal instrument. Dissemination is planned through JBI channels, YouTube, social media, and a research brief.

Limitations and caveats: Being a protocol, it contains no results, and protocols often deviate in execution. The scoping-review format means breadth over depth: no formal quality appraisal of the guidance documents themselves, and analysis details (agreement procedures beyond discussion) are left open. Restricting the search start to 2005 is justified by Arksey & O'Malley but could miss earlier analogs. The PubMed search strategy shown in Appendix I looks partly garbled in the pdftotext extraction (mismatched quotes, an odd "OR 'Or'[All Fields]" fragment), so the exact string should be verified against the published appendix. Google Scholar screening limited to the first 200 results follows cited precedent but is an acknowledged pragmatic cap.

Practical implications for a non-academic researcher using PubMed and AI: The paper is a template for rigorous, transparent searching: start narrow in PubMed to learn the vocabulary, adapt the string across databases, add Google/Google Scholar and targeted websites for grey material, and document everything. For a non-academic reviewing literature, the key insight is that "scoping review" outputs are heterogeneous and largely unappraised — when you cite a scoping review (including any AI-assisted evidence map), treat it as descriptive mapping, not verified evidence, since no AMSTAR-equivalent existed at time of writing. The described newsletter/social-media solicitation is a practical grey-literature trick anyone can copy.

Notable quotes: "The conduct of scoping reviews has repeatedly been found to be inconsistent — that is, there is no one standard approach followed by all authors, which may increase potential bias." — "stakeholders still need the validity and reliability of scoping reviews." — "Future research will involve a modified Delphi method to vote on the items identified within the scoping review."

Connections: Shares the Arksey & O'Malley/JBI framework and PRISMA-ScR reporting with Laynor (2024), whose protocol cites this very JBI guidance (Peters et al. 2022) — the two are sibling scoping-review protocols. Both use Google/Google Scholar grey-literature searching like Godin (2015) and reference-list screening like Briscoe (2020). Its plan to develop an appraisal tool via Delphi mirrors Luo (2025)'s Delphi-based guideline construction. Its concern with review quality links to Haby (2016)'s AMSTAR-based quality assessment of reviews and to Grimshaw (2012)'s premise that syntheses must be trustworthy before translation. Godino (2023)'s Kmet quality-scoring step shows the appraisal layer applied to primary studies, the analogue of what Pollock proposes for reviews.

Godin (2015): Applying systematic review search methods to the grey literature: a... — Case study applying systematic review search methods to the grey literature: Canadian school breakfast program guidelines (Systematic Reviews 2015)

What the paper is and why it matters: This methods paper demonstrates, through a live case study, how to search grey literature with the same rigor expected of academic database searches. The subject — guidelines for school-based breakfast programs in Canada — was chosen

deliberately because such guidelines live on government and NGO websites, not in journals, making database searching structurally inadequate. Grey literature matters for synthesis because excluding it can inflate treatment-effect estimates via publication bias, and because much policy-relevant work is never formally published. Yet, as the authors note, "there is no 'gold standard' for rigorous systematic grey literature search methods," and Cochrane Handbook guidance is thin. The paper provides one of the most-cited practical templates: a four-strategy, pre-specified, PRISMA-reportable search plan.

Methods in detail: Before searching, the team wrote a grey-literature search plan defining resources, terms, websites, and limits, plus eligibility criteria (e.g., published by Canadian federal or provincial/territorial government or NGO; English; current version; intended for school administrators/coordinators; containing breakfast-program standards in at least two of four areas). Four complementary strategies were used: (1) grey-literature databases — searched 20 March 2015 in three Canadian collections (Canadian Research Index; Canadian Public Policy Collection; Canadian Health Research Collection), with results exported to Excel and de-duplicated via the remove-duplicates function; (2) custom Google search engines — the Ontario Public Health Libraries Association engine and Carleton/Waterloo government-documents engine, each hit with ten unique search strings, first ten pages (100 results) screened, relevant pages bookmarked in Chrome and filed in strategy-labelled subfolders; (3) targeted website browsing — a Google search identified 77 relevant organizations, each site hand-searched via its own search bar (13 April 2015), yielding 126 records; (4) content-expert consultation — 35 experts emailed 30 March 2015 with reminders after 7 days and a 3-week window. Screening used abstracts, executive summaries, or tables of contents (since grey documents often lack abstracts), with numbered exclusion codes; one reviewer screened, erring toward inclusion.

Key results and numbers: The four strategies produced 302 potentially relevant items; 15 publications met all criteria (11 of 15, 73.3%, were duplicates across strategies). Targeted web searches were the most productive strategy, identifying all but one included publication; grey-literature databases performed worst, contributing only one publication (mostly returning position papers, evaluation reports, and primary studies). Content experts identified 9 of the 15 included publications; four publications were uniquely identified by a single strategy, meaning reliance on any one source would have produced a less thorough review. Twenty-nine of 35 (82.9%) experts responded. Eight of 15 (53.3%) documents were from NGOs, seven (46.7%) from provincial governments; nothing came from federal or territorial governments (Saskatchewan, Quebec, New Brunswick, PEI also unrepresented). Publication dates ranged 2006–2015 (or 2009–2015 as stated in the characteristics discussion). The whole search was executed in 2 months by one lead author, with phase-level hours approximated retrospectively.

Conclusions and takeaways: Systematic search methods — a documented plan, multiple complementary strategies, preset screening page-limits, numeric exclusion codes, PRISMA-style flow reporting — are feasible for grey literature and were judged "manageable, comprehensive, and intuitive when applied in practice." The method is explicitly offered as a template adaptable to other disciplines (the authors suggest environmental databases, discipline-specific custom Google engines, or building your own). Excel proved serviceable for record management (direct database export, de-duplication, sortable screening columns) though it could not import from Google and demanded browser-to-spreadsheet switching. Substantively, the case study surfaced a policy gap invisible in academic sources: no federal or territorial breakfast-program guidelines exist.

Limitations and caveats: The authors concede the search's inherent biases: personalized "filter bubble" effects in web search engines, custom-engine retrieval quirks (thousands of claimed hits

but few displayed pages), and the transience of web content, which undermines later reproduction or updating. Only one reviewer screened — a deviation from the two-reviewer norm — and time data were approximated retrospectively rather than logged prospectively. Grey documents have misleading "catchy" titles and often lack dates or contributor information, complicating screening and extraction. Standard strength-of-evidence criteria may not fit grey sources. The findings themselves are Canada-specific and topic-specific.

Practical implications for a non-academic researcher using PubMed and AI: This is arguably the most directly actionable paper in the batch for a non-academic: if your question is answered by reports, guidelines, or organizational websites rather than journal articles, PubMed alone will fail you. Copy the four-strategy structure — subject databases, disciplined Google searching with a fixed results-screening limit, hand-searching key organization sites, and asking human experts (with a reminder protocol). AI chatbots can accelerate the expert-consultation analogue (suggesting which organizations publish on a topic) but the paper's discipline — pre-specified strings, page limits, bookmark filing, exclusion codes — is what keeps the search defensible, and current LLMs cannot browse with reproducibility guarantees.

Notable quotes: "there is no 'gold standard' for rigorous systematic grey literature search methods and there are few resources on how to conduct this type of search." — "relying on a single database as an information source … would have omitted nearly half of all eligible documents." — "grey literature often contains titles that do not accurately describe the document and are 'catchy' but misleading."

Connections: Complements Briscoe (2020) as the other supplementary-search methods paper: Briscoe covers citation chasing within academia, Godin the web/grey periphery — and both cite the same logic that text-based database searching misses hard-to-capture content. Godin's Excel workflow is the grey-literature sibling of Godino (2023)'s Excel method, including the same praise and the same complaint (no direct export from Google). Its expert-consultation strategy parallels the snowball sampling Luo (2025) used to recruit experts and the snowballing citation chasing in Laynor (2024). Its "custom Google engine, first N pages" cap is methodologically akin to Pollock (2022)'s Haddaway-based first-200-results rule. Its warning about publication bias supports Grimshaw (2012)'s argument that syntheses must be comprehensive to be trustworthy.

Haby (2016): Best methodologies for rapid reviews for evidence-informed decision... — Rapid review of the best methodologies for rapid reviews in evidence-informed health decision making (Health Res Policy Syst 2016)

What the paper is and why it matters: This is a rapid review *about* rapid reviews, commissioned by Brazilian health policy decision-makers, asking: what are the best methodologies for producing evidence syntheses quickly enough for policy and practice use? It matters because the "lack of timely and relevant research" is among the most frequently reported barriers to evidence-informed decision making, and rapid-review products have proliferated with little methodological grounding. The paper is a self-demonstrating artifact: it applies systematic-review methodology (registered protocol, PROSPERO CRD42015015998, PRISMA) within a compressed timeline, then transparently documents its own shortcuts and AMSTAR self-assessment — modeling exactly the transparency it recommends.

Methods in detail: Systematic review methods were used per a pre-defined, registered protocol. Inclusion covered systematic reviews (priority) and strong primary designs (RCTs, quasi-RCTs, controlled before-after, interrupted time series, economic evaluations), with a fallback category

for weaker evaluations; 'products' (papers containing rapid reviews) were admitted as units too, a mid-review decision alongside two added outcomes (comparison of synthesis methods; cost-effectiveness). Publications from 2004 onward (the Mexico Ministerial Summit year) in English, French, Portuguese, or Spanish were eligible. Eleven databases (CINAHL, Cochrane Library including DARE/HTA/NHS EED/methods studies, EconLit, EMBASE, Health Systems Evidence, LILACS, Medline) plus Google and Google Scholar were searched 15 January–3 February 2015; supplementary searches (reference lists, contact with 9 key authors, 5 responding) ran in May 2015. One reviewer screened initially (erring toward inclusion), two reviewed full texts; extraction by one, checked by a second; quality assessed independently by two reviewers using AMSTAR for reviews (0–3 low, 4–7 medium, 8–11 high) and Cochrane Risk of Bias for the RCT. Narrative synthesis only.

Key results and numbers: Five systematic reviews (from seven articles) plus one RCT met the criteria; 75 papers were excluded at full text and 12 weak-design evaluations were listed but not extracted. AMSTAR scores were poor: four reviews scored 2 (low quality), one scored 4 (medium); none high quality, so findings are "indicative only." Consistent across reviews: no agreed definition and no agreed methodology for rapid reviews; one quoted review notes the term "is framed in the literature as utilizing various stipulated time frames between 1 and 6 months." Completeness of methods reporting varied — about 43% of rapid reviews examined by Abrami et al. did not describe their methodology comprehensively. Time-to-complete ranged widely: several days to 1 year (Abrami), 3 weeks to 6 months (Ganann), and 7–12 months (mean 10.42, SD 7.1) (Harker & Kleijnen). Documented shortcuts include fewer databases, limited grey literature, English-only or last-5-years restrictions, single-reviewer screening/extraction, limited or no risk-of-bias assessment, and minimal synthesis. Crucially, little empirical evidence exists on whether shortcuts change conclusions: of two known comparisons, one (four surgical topics) found no opposed essential conclusions between rapid and full reviews, the other (potato peels for burns) found differing conclusions, confounded by the full review's poor search. The RCT (low risk of bias; Opiyo et al.) found no difference in correct responses across evidence-pack formats, but the rapid-review-style "pack C" improved the composite clarity/accessibility score versus systematic review alone (adjusted mean difference 0.52, 95% CI 0.06–0.99), and interviewees preferred it.

Conclusions and takeaways: Rapid reviews are legitimate — "there is no evidence available to suggest that rapid reviews should not be done or that they are misleading in any way" — but the field lacks definitions, agreed methods, and evidence on shortcut impacts. The authors propose a working definition (components of the systematic review process simplified, omitted, or made more efficient to produce information faster, ideally with minimal quality impact, with a close relationship to the end-user) and note a high-quality rapid review may outperform a low-quality systematic review. Priority shortcuts are those least likely to affect quality: narrowing scope, limiting data extraction to key results, restricting study types. Users should demand an AMSTAR rating and an explicit list of shortcuts; producers should improve write-up transparency, put detail in appendices, and maintain skilled standby staff.

Limitations and caveats: The evidence base itself was low quality (four of five reviews AMSTAR 2), limiting any firm conclusion. Reviews were heterogeneous and often non-random samples of rapid-review products, so shortcut frequencies cannot be quantified. Only two primary rapid-vs-systematic comparisons existed, both below the review's own design bar. Language restriction to four languages and the 2004-onwards limit were pragmatic; few studies came from developing countries. The review itself took 5–7 months, straining the "rapid" label, and meta-analysis was impossible. Table 1′s layout is somewhat garbled in the pdftotext extraction (review names detached from their rows), though the narrative is intact.

Practical implications for a non-academic researcher using PubMed and AI: If you need evidence fast, this paper tells you which shortcuts are safe: narrow the question, restrict study types and date range, limit data extraction — and which are risky to do silently: skipping quality appraisal, grey literature, or dual screening. Publish/transparency rule of thumb: state your timeframe, your shortcuts, and your limitations; consider a self-appraisal against AMSTAR. AI now offers a tempting new "shortcut" (automated screening/synthesis) that this literature has not yet evaluated — the paper's central warning, that shortcuts' effects on conclusions are empirically unknown, applies with even greater force to LLM-assisted reviews.

Notable quotes: "None of the systematic reviews were of sufficient quality to allow firm conclusions to be made." — "the term 'rapid review' does not appear to have one single definition but is framed in the literature as utilizing various stipulated time frames between 1 and 6 months." — "a high quality rapid review is likely to provide an answer closer to the 'truth' than a systematic review of low quality."

Connections: Extends Grimshaw (2012)'s barrier analysis (untimely evidence) into a methods question, and its AMSTAR use links directly to Pollock (2022), whose protocol targets appraisal tools for another review type. Its transparency demands parallel Briscoe (2020)'s reporting findings and prefigure Luo (2025)'s GAMER logic — all three argue that what you did must be fully describable. Its "little evidence shortcuts change conclusions" caveat is the caution a non-academic needs when tempted by Godino (2023)'s lean Excel pipeline or AI-accelerated workflows. Its search mirrors Godin (2015)'s multi-source approach (databases + Google + experts + reference lists). Laynor (2024) is the counterpoint: where Haby shows the field of rapid reviews is methodologically immature, Laynor pre-registers everything.

Grimshaw (2012): Knowledge translation of research findings — Knowledge translation of research findings: five-question framework and evidence on strategies to close the evidence-practice gap (Implementation Science 2012)

What the paper is and why it matters: A highly cited debate/discussion paper (with Eccles, Lavis, Hill, Squires) synthesizing concepts and evidence on knowledge translation (KT), defined as "ensuring that stakeholders are aware of and use research evidence to inform their health and healthcare decision-making" — i.e., T2 translation of clinical knowledge into improved health. It matters because it frames the entire downstream rationale for literature searching and evidence synthesis: synthesis is worthless if it never reaches decision-makers, and the paper documents just how large that failure is — US patients received only 55% of recommended care, and 20–30% of patients may receive unnecessary or potentially harmful care. It organizes KT around five questions (what, to whom, by whom, how, with what effect) and reviews the effectiveness evidence for each audience.

Methods in detail: This is not an empirical study but a structured narrative synthesis based on a previously published monograph chapter, organized by Lavis et al.'s five questions. For "what": the authors argue from Ioannidis' meta-research (Proteus phenomenon, unstable early effect sizes, highly cited findings later contraindicated) that up-to-date systematic reviews should be the basic unit of KT, tailored into products per audience (decision aids, guidelines, actionable messages, policy briefs). For "to whom": tables map stakeholders (consumers, professionals, administrators, policy makers, regulators, industry, funders, researchers) to research types, and clinical-research scenarios (stop use vs. promote) to primary audiences. For "by whom": credibility matters (endorsed messengers influence guideline use); knowledge infrastructures (Rx

for Change, Health Systems Evidence databases; knowledge brokers; training) are proposed instead of ad hoc researcher messengers. For "how": barrier assessment should inform strategy choice; effectiveness evidence is summarized from Cochrane EPOC reviews for professionals and Cochrane Consumers and Communication reviews for consumers, plus policy-maker studies. "With what effect" distinguishes evidence-informed (policy) from evidence-based (professional) endpoints.

Key results and numbers: For professionals (EPOC reviews, to August 2011; >7,000 trials identified, 300+ systematic reviews of behavior-change strategies): printed educational materials — median absolute improvement 4.3% (range −8.0% to +9.6%, 12 trials); educational meetings — 6.0% (IQR +1.8% to 15.3%, 81 trials, >11,000 professionals; interactive and mixed formats larger); educational outreach — prescribing 4.8% (IQR +3.0% to +6.5%) and other behaviors 6.0% (IQR +3.6% to +16.0%, 69 trials, >15,000 professionals); local opinion leaders — median 12.0% (IQR +6.0% to +14.5%, 18 trials); audit and feedback — 5.0% (IQR +3% to +11%, 118 trials, larger when baseline compliance low); computerized reminders — OR 1.52 (95% CI 1.27–1.82, 28 trials plus meta-regression); tailored interventions — 4.2% (IQR +0.8% to +18.8%, 26 trials). Multifaceted interventions showed no dose-response with component number. For consumers: decision aids (86 trials, >20,209 participants) improved knowledge, risk-perception accuracy, and value-consistent choices; written medicines information improved knowledge in 6 of 12 trials but little behavior change; self-management programs gave small short-term gains. For policy makers: one study found 23% (3 months) and 63% (2 years) of respondents used at least one provided systematic review; a randomized trial found tailored messages plus registry access significantly improved public health policies/programs; Lavis' review of 16 studies found researcher–policy-maker interaction and value-matched research most consistently associated with research use.

Conclusions and takeaways: Single studies are the wrong unit for translation; systematic reviews of global evidence should anchor it, "to increase the 'signal to noise' of knowledge translation activities." Strategy choice should follow assessed barriers, since barriers operate at structural, organizational, peer, professional, and interaction levels beyond individual knowledge deficits. The evidence base for professional- and consumer-targeted KT is substantial if incomplete; for policy makers and senior managers it is much weaker despite a profusion of innovative approaches (SUPPORT Tools, graded-entry formats, knowledge brokers, deliberative dialogues) that need evaluation. Grol's challenge — "evidence-based medicine should be complemented by evidence-based implementation" — is the closing call.

Limitations and caveats: As a narrative synthesis of selected (mostly Cochrane) reviews, it is not a systematic overview; review selection favors EPOC quality, and effect sizes are medians across heterogeneous interventions and contexts, so transferability is uncertain. The policy-maker evidence rests on very few studies with inconsistent findings. Terminology proliferation is acknowledged as itself a barrier. Tables 1–3 are severely scrambled in the pdftotext extraction (rows split across columns), so table-derived numbers should be verified against the published version. Being 2012, it predates rapid reviews as a named field (though Haby 2016 cites its framework), predates GAI entirely, and its Cochrane-centric view reflects that era.

Practical implications for a non-academic researcher using PubMed and AI: For a non-academic producing evidence summaries, the paper's core lessons are: (1) build on systematic reviews, not single studies — search PubMed for reviews and meta-analyses first; (2) know your audience and tailor format (graded-entry: one page of messages, short summary, then full report); (3) identify barriers before choosing how to communicate; (4) credibility of the messenger

matters. AI tools map onto the "knowledge management" barriers Grimshaw lists — volume, access, time, appraisal skill — but the paper warns that b

etter knowledge management alone is "unlikely by itself to be sufficient," a caution against assuming an AI-drafted summary will actually change anyone's behavior.

Notable quotes: "the basic unit of knowledge translation should be up-to-date systematic reviews or other syntheses of the global evidence." — "patients in the USA received 55% of recommended care." — "'evidence-based medicine should be complemented by evidence-based implementation.'"

Connections: The conceptual anchor of the batch: Haby (2016) directly builds on its barrier framework and cites Lavis; Pollock (2022), Briscoe (2020), Godin (2015), and Godino (2023) all describe the synthesis machinery Grimshaw argues must precede translation; Laynor (2024) studies the researchers who produce that machinery. Its Cochrane/EPOC reliance connects to Briscoe's Cochrane sample and Haby's AMSTAR. Its "knowledge brokers" and Health Systems Evidence database anticipate the intermediary role that both Luo (2025)'s reporting standards and the AI tools GAMER regulates now play; GAMER's verification requirement operationalizes Grimshaw's demand that translated content be trustworthy.

Godino (2023): How to structure Microsoft Excel documents for systematic reviews — Practical guide to structuring Microsoft Excel documents for systematic review management and PRISMA reporting (Nurse Researcher 2023)

Note on text quality: this file is an OCR transcription of a scanned PDF with frequent artefacts — e.g., "tree"/"Fxcel" for "free"/"Excel", "ot"/"l" for "of"/"1", "nurscresearcher.com" for "nursingresearcher.com", "van dar Mierden" for "van der Mierden", and heavily garbled figure captions and flowchart labels (e.g., "17=" for "n="). Figure-level numbers are unreliable; body text is recoverable.

What the paper is and why it matters: A short practice paper by a genetic/research nurse at the University of Bologna explaining a free, six-step Microsoft Excel method for managing a systematic review's reference flow — from importing records to producing the numbers for a PRISMA flowchart. It matters because dedicated review-management tools are unevenly accessible: the paper cites van der Mierden et al.'s evaluation of 16 tools, in which none supported the complete process, and the best performers (DistillerSR, EPPI-Reviewer, Covidence, SWIFT Active Screener) are paid. For students, nurses, and independent researchers without institutional software, Excel is effectively universal. The method is presented as refined over years of use in published reviews by the author's group (Godino & Skirton 2012; Godino et al 2013, 2016; Bertonazzi et al 2022).

Methods in detail: The method has four phases and six steps. Step 1: download all references from each database as Excel/CSV exports, saved in one folder and renamed with database and date (e.g., 'Pubmed_22.05.22.xlsx'). Step 2: build a workbook with named worksheets — 'All references', 'All references with duplicates', 'All references without duplicates', 'Full-text screening', 'Articles included' — plus a per-database sheet; standard headings (Database, Title, Authors, Abstract, Year, Language, Journal, Document Type, DOI) are pasted into the master sheets. Step 3: de-duplicate by sorting on Title/Authors/Database, then applying Conditional Formatting → Highlight Cells Rules → Duplicate Values; duplicates are flagged with a 1 in a 'Duplicates' column, totalled with =SUM(…), and removed in the no-duplicates sheet. Step 4: title/abstract screening in the clean sheet — one column per inclusion/exclusion criterion plus

a 'To read full-text' column, coding 1s, stopping at the first failing criterion, then =SUM totals for exclusions and for full-text retrieval. Step 5: full-text screening in the dedicated sheet with a 'To include' column (also 1 for unretrievable articles, to document them). Step 6: produce the final 'Articles included' sheet, duplicable for quality assessment and synthesis. Quality appraisal uses Kmet et al. (2004) criteria (14 items quantitative, 10 qualitative; 0/1/2 scoring; percentage of maximum; ~60% suggested cut-off), with two independent assessors in dedicated columns.

Key results and numbers: The paper reports no empirical evaluation of the method itself — it is a how-to, with two worked PRISMA flowchart examples drawn from the author's unpublished work and Bertonazzi et al (2022). From the (OCR-garbled) figures: the 2020-style example lists records from PubMed, CINAHL, Psychology Database, ASSIA (n=342), and Cochrane Library, with roughly 1,358 records screened, 297 assessed for eligibility, and detailed reason-coded exclusions (e.g., not on topic, chromosomal/genetic techniques, ethical/legal, books/thesis, commentary/letter, abstract only, review, no primary research), plus a Google Scholar arm (58 records identified, 57 sought and assessed); the 2009-style example shows 82 database records plus 1,507 from other sources, 1,529 after duplicate removal, 24 full-text assessments, and 11 studies in the qualitative synthesis. These numbers are illustrative of scale, and several digits are OCR-corrupted, so they should not be quoted as findings.

Conclusions and takeaways: The Excel method "is efficient and free and can produce transparent and complete reports of systematic reviews," constituting "a valid alternative" to advanced software. Its strengths: whole-process auditability ("Why did I not include this article?" is answerable from the workbook), accurate counts, simple navigation, low cost, and innovative data-extraction features. Weaknesses, candidly listed: it cannot search databases, cannot screen in real time, and presupposes sound knowledge of the systematic review process and its information-management pitfalls — "not considering its limitations could lead to an increased workload rather than a simpler one." The resulting documents are positioned as a direct source for writing the review's methods and results.

Limitations and caveats: No comparative evaluation against reference managers or screening software; single-author experiential basis; the worked examples are the author's own projects, so generalizability beyond health/genetics topics is asserted rather than tested. Excel's de-duplication is literal-string based, so variants of the same record (punctuation, subtitle differences) can survive — a risk the paper does not discuss. Dual screening is accommodated for quality assessment but the screening steps as described are essentially single-reviewer plus criteria columns, which falls short of standard dual independent screening. As flagged, the OCR extraction garbles figures, DOIs (e.g., "10.774B/nr,2023.e1866"), and some reference details.

Practical implications for a non-academic researcher using PubMed and AI: This is the batch's most immediately usable paper for a solo, non-funded reviewer: PubMed exports can go straight into this workbook structure, and the PRISMA numbers fall out of =SUM formulas. Pair it with an AI assistant to draft the inclusion/exclusion criteria columns or to pre-triage obviously irrelevant titles — but keep the human-coded 1/0 decisions and totals, since the method's value is its audit trail. Be aware of its blind spots: no real-time collaboration, brittle de-duplication, and no support for the dual-screening norm that quality assessment standards (and the rapid-review literature) emphasize.

Notable quotes: "It is 'a replicable, scientific and transparent process… that aims to minimise bias through exhaustive literature searches… and by providing an audit trail of the reviewer's decisions, procedures and conclusions' (Tranfield et al 2003)." — "The Excel method is efficient and tree [free] and can produce transparent and complete reports of systematic reviews." (OCR

artefact noted) — "not considering its limitations could lead to an increased workload rather than a simpler one."

Connections: The practical record-management layer beneath the batch's methods papers: Godin (2015) already used Excel for grey-literature de-duplication and screening and lists the same pros/cons (including no Google export). Godino's Kmet-based quality step is the primary-study analogue of what Pollock (2022) seeks for reviews and what Haby (2016) performed with AMSTAR on reviews. Its PRISMA-flow output is exactly the reporting transparency Briscoe (2020) found wanting for citation searching and Haby found wanting for rapid reviews. Luo (2025) is the natural next step: GAMER-style disclosure of any AI assistance layered onto Godino's audit-trail workbook. Grimshaw (2012) supplies the downstream purpose — a transparently selected evidence base is what deserves translation.

Laynor (2024): Identifying meta-research with researchers as study subjects (proto... — Registered Report protocol for a scoping review of meta-research with researchers as study subjects, with a novel PubMed search strategy (PLoS ONE 2024)

What the paper is and why it matters: A Registered Report protocol (PLoS ONE 19(5):e0303905) by a health-sciences librarian and a population-health researcher for a scoping review of "research-on-researchers" — studies in which researchers themselves are the study population (perceptions, behaviors, incentives, engagement). It matters on two fronts. Substantively, meta-research with researchers as subjects can reveal how to support researchers and build research capacity, but the literature is invisible to normal searching. Methodologically, the paper is a masterclass in search-strategy development for a topic with no stable terminology: "researcher" appears in tens of millions of articles, the field was named only recently, databases lack MeSH indexing for it, and "meta-research" is itself confused with meta-analysis. As a PLoS Registered Report, methods were peer-reviewed before results exist, inverting the usual incentive structure.

Methods in detail: The protocol follows JBI scoping-review methodology and Arksey & O'Malley, reported per PRISMA-ScR. Stage 1 fixes the question: what research has been performed where researchers are the study population, with objectives to describe such research, identify its data sources and methods, and summarize barriers to conducting it. Stage 2 searches three core databases — PubMed, CINAHL, PsycInfo (the latter two also indexing theses, dissertations, and conference proceedings) — scoped to health-related disciplines for feasibility, with any-language abstracts excluded (English abstract required), researchers-as-subjects required, and CBPR- or QI-only studies excluded (physicians doing QI research on themselves count). The search strategy was built by "harvesting" keywords and MeSH terms from a known-item seed set (S2 Table), then iteratively tested for retrieval of those known items. It combines two approaches OR-ed together: (A) Component A researcher terms (Research Personnel MeSH; clinician-scientist; investigator; physician-scientist; professor; research assistant; research staff; scientist; trialist; etc.) each proximity-searched (within two words) with every Component B action term (engaged, involvement, participate, perceptions, responded, …) in title/abstract; plus (B) Component A terms AND meta-research terms (meta-research, meta-science, research-on-research) with Research Personnel subheadings ("psychology", "trends", "statistics and numerical data"). Post-screening, forward and backward citation chasing (snowballing) supplements the searches. Screening runs in Covidence with dual independent review at title/abstract and full-text stages,

pilot-tested (100 abstracts; 10 full texts), with kappa ≥0.8 targeted, consensus resolution, and a third reviewer for impasses.

Key results and numbers: As a protocol, no review results are reported, but the search-strategy development narrative contains concrete numbers: the seed-set harvesting and iterative testing (with yields reported in S3 Table) showed that restricting "researcher" and synonyms to title/abstract "still yielded a large number of results" that would mostly be screened out, motivating the proximity-phrase approach; the paper also quantifies the terminology problem — "researcher" appears in tens of millions of articles versus "smoker" in "less than one-fourth" that number. Five protocol stages are specified; the data-extraction form covers author(s), year, country, funding, aim, population and sample size, discipline, design, data-collection methods, response rates, and barriers/facilitators to research-on-researchers, to be piloted on five studies and iteratively updated. Analysis will include a bibliometric mapping (context, geography, year), tabular summary, and inductive thematic analysis of qualitative barrier/facilitator data.

Conclusions and takeaways: The protocol's methodological thesis is generalizable: for topics whose key terms are ubiquitous in unrelated literature, pairing iterated Component-A×Component-B phrases under a proximity operator balances sensitivity (every term combination) with specificity (the two-word window), providing "a model for developing search strategies on topics difficult to search due to their interdisciplinarity, broad scope, or lack of well-defined terminology." Snowballing after database screening adds recall without exploding the screening burden. The review itself is positioned to define research-on-researchers as a sub-field, map its landscape, and extract transferable lessons for engaging researchers as participants. Limitations acknowledged: health-domain-only scope, three databases only, and residual sensitivity/specificity trade-offs.

Limitations and caveats: It is a protocol — no findings yet, and Registered Reports do not guarantee completion or immutability in practice. The health-research scope deliberately excludes other fields, so the bibliometric picture will be partial. The proximity strategy's assumption — that studies about researchers will name them prominently in title/abstract — is plausible but unproven at scale; some eligible studies will use neither term pattern. English-abstract requirement biases against non-English literature. Kappa ≥0.8 is ambitious for conceptually fuzzy inclusion decisions. No inter-rater data from the pilot is reported in the protocol text itself (details deferred to supplementary tables).

Practical implications for a non-academic researcher using PubMed and AI: The paper demonstrates PubMed-native techniques a non-academic can copy today: use MeSH (Research Personnel) plus a controlled synonym list; exploit PubMed's proximity operator (added 2022) to bind terms like ("research assistant" AND participation) within two words; test your string against a hand-list of known relevant papers before trusting it — a cheap sensitivity check. For AI: the paper is also a caution — LLM-suggested synonyms for an ill-defined field will inflate an already explosive search; the disciplined iterative seed-set testing shown here is a better guide. Finally, the snowballing step pairs naturally with free citation tools (Google Scholar), echoing Briscoe (2020).

Notable quotes: "the term meta-research is also occasionally used to refer to meta-analysis, which is a research process used to systematically synthesize the findings of independent studies and distinct from meta-research the discipline." — "'researcher,' is used in tens of millions of published articles and therefore a less useful term for identifying a study population than the term 'smoker'." — "iteratively phrases and proximity searching were used" [sic; the text reads "iterated phrases and proximity searching were used"] "after a process of testing different approaches to the search strategy on a sample set of known studies."

Connections: Methodologically the bridge between Briscoe (2020) and the scoping-review protocols: it *uses* the forward/backward citation chasing Briscoe audited, within the Arksey & O'Malley/JBI/PRISMA-ScR framework that Pollock (2022) also follows (and Pollock's team's 2022 JBI guidance is Laynor's reference 19 — the two protocols are direct siblings). Its snowball-sampling and expert-seed-set logic parallels Godin (2015)'s expert consultation and Luo (2025)'s expert recruitment. Its core problem — searching for studies about studies — is the meta-level of Grimshaw (2012)'s synthesis enterprise and of Haby (2016)'s review-of-reviews; both Haby and Laynor depend on searches that must find reviews themselves, and Haby's term-proliferation finding (20–31 names for rapid reviews) is the same disease Laynor diagnoses for meta-research. Godino (2023) offers the low-cost management tool for the screening volumes Laynor's strategy will generate.

14.4 The AI core — automation of screening and extraction

Source: pdftotext extractions in C:\Users\kukac\Desktop\Research\fulltexts. All statements below are based strictly on the extracted text; minor extraction artefacts are noted where relevant.

Lund (2024): ChatGPT in medical libraries: possibilities and future directions — Integrative review: ChatGPT's possibilities in medical libraries (Health Information & Libraries Journal 2024)

What the paper is and why it matters: This is one of the first integrative reviews specifically on ChatGPT in *medical* libraries, published in a library-and-information-science venue (Health Information & Libraries Journal, 2024;41:4–15) by authors from the University of North Texas, Aligarh Muslim University and Central University of South Bihar. It matters because medical librarians sit at the interface between clinicians and the evidence base, and the paper frames ChatGPT as a potential adjunct to reference, search-support and data-organisation services — while insisting it remain "supplementary." It is an early, methodologically explicit snapshot (search cut-off 30 July 2023) of how the library profession was thinking about generative AI, useful as a baseline against which later automation studies (Li 2025, Khan 2024) can be judged.

Methods in detail: The authors used Cooper's integrative review method (five stages: problem formulation, search methods, data evaluation, data analysis, presentation). They searched eight databases — EBSCOhost, Dimensions.ai, Science Direct, Google Scholar, Medline, Scopus, PubMed and Web of Science — on 30 July 2023 with the terms "ChatGPT and libraries" and "AI and libraries", in English only. A PRISMA-style flow is given, but the numbers are internally inconsistent: the abstract reports 166 papers with 30 excluded for irrelevance and 136 selected; the text says 166 identified, 20 excluded for irrelevance (→146), 5 excluded at abstract review, 5 more for methodological concerns (prompt-only studies "lacked both discussion and analysis"), giving 136; the flow diagram shows 196 combined results, 166 screened, 30 duplicates removed, 136 full-text assessed. Quality appraisal used the CASP Qualitative Research Checklist (9 items A–I); only papers scoring 6–9 were kept, which reduced 136 articles to 29 for analysis. Data analysis was directed/thematic content analysis, producing two groups: "ChatGPT and medical libraries" (20 tabulated articles) and "AI and medical libraries" (9 tabulated articles).

Key results and numbers: No performance metrics are reported; the outputs are qualitative. The 29 included articles support that ChatGPT can "address reference inquiries, facilitate website navigation, and aid in research, cataloguing, classification and collection development" (Adetayo 2023a, as summarised). Adetayo's survey of 54 Adeleke University students found ChatGPT "a valuable tool for reference services," but with limits: "difficulties

in answering complex questions, the possibility of providing incorrect information, the risk of outdated data and the inability to comprehend emotions." Aithal & Aithal position ChatGPT as complementary to traditional/digital libraries; Chen finds value in answering reference queries vs traditional library chatbots; Panda & Kaur suggest LLM-based chatbots can replace knowledge-base chatbots. On the AI side, Abayomi et al. report librarians aware of AI but deterred by job-loss fears; Wheatley & Hervieux find a "notable lack of awareness" of AI trends at top US/Canadian research universities. The paper also offers a 10-step evaluation protocol (define criteria, curate a test dataset, realistic scenarios, staff testing, accuracy/context/response-time checks, feedback, reporting) and 6 usage guidelines for librarians.

Conclusions and takeaways: ChatGPT could enhance reference services, research support, cataloguing, customised information support, medical-writing collaboration and chatbot services in medical libraries — but only as a complement to human librarians who supply "critical thinking, emotional intelligence and domain expertise." Continuous monitoring, "validation against trusted sources," quality-control processes, ethical guidelines, privacy safeguards and user education are prerequisites. Medical libraries are urged to evaluate ChatGPT locally (the 10-step protocol), to stay informed, and to participate in research on its effectiveness and ethics.

Limitations and caveats: The authors list them: possible publication bias (English-only, no grey literature); reliance on a single methodology; search terms that may have missed studies; methodological exclusions that may have discarded "valuable insights"; subjectivity in CASP scoring; and rapid obsolescence of findings in a fast-moving field. Beyond these, the internally inconsistent PRISMA counts (166/196/146/136) suggest loose reporting; applying a *qualitative* CASP checklist to quantitative or mixed studies is questionable; and most of the 29 included papers are commentaries, short surveys or proof-of-concept pieces rather than evaluations — so the review's synthesis is correspondingly speculative.

Practical implications for a non-academic researcher using PubMed and AI: Treat the review as a checklist culture rather than a results source. Its concrete guidance transfers directly: define your objective before prompting; craft specific, medically contextualised prompts; cross-verify anything clinical or bibliographic against PubMed/reputable sources; and remember ChatGPT's answers about the literature are unverifiable until checked in a database. The 10-step evaluation idea scaled down — build a small test set of your own questions, run them through your AI tool, judge accuracy and response time before trusting it — is a workable personal QA habit.

Notable quotes:

- "Positioned as a complementary tool, ChatGPT excels in delivering customised information support to medical library users, distinct from traditional library systems."
- "It is essential to emphasise that ChatGPT is designed as a supplementary technology, complementing the expertise of human librarians rather than replacing them."
- "continuous monitoring, validation against trusted sources, and the implementation of quality control processes are necessary to minimise the risk of misleading or unreliable information."

Connections: Shares an author (Brady Lund) and themes with the wider batch: it supplies the librarian's-eye "verify everything" stance that Yarar (2026): Concerns of using LLMs in health care research: umbrella review's umbrella review later substantiates (hallucinations, bias, transparency). Its "supplementary tool" framing recurs verbatim-like in Yarar (2026): Concerns of using LLMs in health care research: umbrella review. It predates and contrasts with the automation pipelines of Li (2025): Enhancing systematic literature reviews with generative AI, Khan (2025): Collaborative LLMs for automated data extraction in living systemat… and Pilz

(2024): Semi-automated title-abstract screening with NLP and machine learning — which show what disciplined, evaluated AI use looks like — and with [[Mahboubi
(2025): SciScribe: automating and contextualizing literature reviews]], which operationalises grounded reference-like Q&A. Gallo (2025): Best practices in LLM research: repeat prompting provides the statistical caution Lund's generation of commentary lacked. Wang (2020): Error rates of human reviewers during abstract screening gives the human-error baseline that justifies libraries accepting imperfect automation.

Li (2025): Enhancing systematic literature reviews with generative AI — Five-module GPT-4 system for systematic literature reviews in HTA (JAMIA 2025)

What the paper is and why it matters: Li et al. (Regeneron Pharmaceuticals and IMO Health) present an end-to-end, human-in-the-loop "AI-SLR" system using GPT-4 in a zero-shot setting to run the core stages of a clinical systematic literature review (SLR) for health technology assessment (HTA) submissions. It matters because HTA agencies demand rigorous, PICOs-structured SLRs on tight timelines (e.g., EU Joint Clinical Assessment deadlines of 100 days for new molecular entities, 60 days for line extensions), and the authors claim this is the first LLM system instructed with PICOs criteria — embedding a real clinical-review methodology into an LLM workflow rather than generic prompting.

Methods in detail: The system has five modules: (1) PubMed/Embase searching with Boolean medical terms (yielding 3,713 abstracts for relapsed/refractory multiple myeloma, RRMM, and 1,720 for advanced melanoma); (2) setting the study protocol with PICOs criteria authored by reviewers with SLR expertise and clinical background; (3) GPT-4 abstract screening against those criteria, with a human-in-the-loop loop — users review at least 10% of randomly selected abstracts (or 30 if <300 results), record agreement/disagreement, tag which PICOs criterion they dispute or add free text, and see live precision/recall/accuracy to refine criteria iteratively; (4) targeted GPT-4 data extraction with user-defined field descriptions (study details, patient characteristics, outcomes with seven elements including group, n, %, median, hazard ratio); (5) structured summary reports. To protect sensitivity, the model was "instructed to include abstracts in cases where the model is not certain or confident." Four evaluation sets were built by reviewers with 5–15 years' HTA-SLR experience: Set 1 (49 RRMM/50 melanoma abstracts, 18/21 relevant), Set 2 (PICOs exclusion reasons), Set 3 (extracted fields), Set 4 (two vendor-conducted SLRs: 3,665 RRMM abstracts with 2,071 inclusions (56.5%); 2,753 melanoma abstracts with 145 inclusions (5.3%)). Metrics: precision, recall/sensitivity, accuracy, specificity, NPV, F1, Cohen's $\varkappa$ and PABAK.

Key results and numbers: Abstract screening across the four sets: average sensitivity 90% (89.64%), F1 82, accuracy 89%, $\varkappa$ 0.71; per-set ranges in the Discussion were sensitivity 82–97%, F1 69–90, accuracy 80–96%, $\varkappa$ 0.57–0.84, PABAK 0.61–0.92. Agreement was stronger against Set 1 ($\varkappa$ 0.74 RRMM, 0.84 melanoma) than Set 4 (0.58 and 0.67) — the authors attribute this to vendors using extra information and an active-learning AI tool; notably, on a random 56-case subset the authors sided with GPT-4 over the vendor in 23 disagreed melanoma cases. Exclusion-reason classification: accuracy 97% (RRMM) and 84% (melanoma), F1 98 and 89, with many criteria at 100%. Data extraction: overall F1 93; extraction sensitivity 94%; itemised precision/recall/F1 80–100/83–100/84–100; hazard-ratio extraction was weakest (precision 71%, recall 63%, F1 67; n=6 instances). Compared with prior GPT screening studies (Guo 81%, Alshami 80%, Khraisha 42% sensitivity), the confidence-inclusive prompt strategy raised sensitivity; $\varkappa$

(0.71) beat Guo (0.23) and Khraisha (0.34). Error analysis found the system often *more* correct than vendor labels, but it struggled to separate original research from reviews and to catch non-randomised designs.

Conclusions and takeaways: A zero-shot, PICOs-driven, human-in-the-loop LLM system can plausibly handle screening, exclusion-reason labelling and abstract-level extraction for clinical SLRs, with sensitivity aligned to the inclusive screening culture of HTA ("irrelevant articles can be excluded during full-text review"). It removes the need for annotated training data, gives subject-matter experts control via prompt/criteria adjustment, could cut the 6–16-month SLR timeline, and may enable living SLRs. Reported human inter-reviewer $\varkappa$ (~0.82, Hanegraaf) brackets the system's Set 1 performance within the publishable range (0.6–0.9).

Limitations and caveats: Evaluation covers only two oncology indications, so generalizability beyond oncology is untested. Set 4′s RRMM relevance rate (56.5%) is atypical (usual 3–20%), depressing specificity (59%) — a known prevalence effect, only partly corrected by PABAK. Human reference standards are imperfect (cf. Wang (2020): Error rates of human reviewers during abstract screening); the melanoma vendor itself used AI-assisted screening. Extraction was abstract-only, on small eval sets (18–21 relevant abstracts), with all-seven-field correctness being a strict bar; Table 2/4 numbers are partly garbled in the extraction, and several cells read "NA." Authors are Regeneron/IMO Health employees (declared as playing no role). The human-in-the-loop mechanism's effectiveness and real time savings are future work, not demonstrated results.

Practical implications for a non-academic researcher using PubMed and AI: The architecture is a template: write PICOs-style inclusion/exclusion criteria as the prompt, force the model to include anything uncertain (optimising sensitivity), have a human audit a random ≥10% sample, compute agreement, and refine criteria until performance stabilises. Even manually with ChatGPT-class tools, auditing a sample and tracking agreement converts a black box into a monitored assistant. Note also the stated reality that ~85% of citations are typically excluded at title/abstract — most of a search's bulk is rejectable, which is where AI leverage is greatest.

Notable quotes:

- "This humanin-the-loop AI-assisted SLR system demonstrates the potential of GPT-4′s in context learning capabilities by eliminating the need for manually annotated training data."
- "To maintain high sensitivity, we instructed the model to include abstracts in cases where the model is not certain or confident."
- "To our knowledge, this is the first time that PICOs criteria is used to instruct an LLM."

Connections: Directly cites Wang (2020): Error rates of human reviewers during abstract screening (ref 10) as motivation — human screening errors and disputes — and benchmarks against earlier GPT screening studies discussed in Yarar (2026): Concerns of using LLMs in health care research: umbrella review-style concern terms (hallucination risk is implicit). Complements Pilz (2024): Semi-automated title-abstract screening with NLP and machine learning: Pilz uses supervised ML/fine-tuned BERT-family models needing labelled training data; Li shows the zero-shot LLM alternative. Shares the two-reviewer/inter-rater logic with Khan (2025): Collaborative LLMs for automated data extraction in living systemat…, though Khan's concordance-between-LLMs trick is different. Its $\varkappa$-vs-human benchmark (Hanegraaf) pairs with Gallo (2025): Best practices in LLM research: repeat prompting's warning that LLM study statistics need correlation-aware analysis. Lund (2024): ChatGPT in medical libraries: possibilities and future directions supplies the professional context; Mahboubi (2025): SciScribe: automating and contextualizing literature reviews is the curation-side counterpart.

Khan (2025): Collaborative LLMs for automated data extraction in living systemat... — Collaborative LLMs (GPT-4-turbo + Claude-3-Opus) for data extraction in living systematic reviews (medRxiv preprint, posted September 23, 2024)

What the paper is and why it matters: A large multidisciplinary team (Mayo Clinic, Arizona State, Yale, Brown, Dana-Farber, ASCO and others) tests whether two different frontier LLMs run in parallel can mimic the classic two-reviewer process of systematic reviews for *data extraction* — the most laborious step of living systematic reviews (LSRs). The hypothesis is elegant: where two independent models agree ("concordance"), the answer is probably right; where they disagree, have each model critique the other's answer ("cross-critique"). It matters because it treats inter-annotator agreement not as a nuisance but as a quality signal, and it is one of the few extraction studies with a genuine held-out test set. Caveat: it is a non-peer-reviewed preprint and states it "should not be used to guide clinical practice."

Methods in detail: Gold-standard data came from a published LSR/network meta-analysis on first-line treatments of metastatic castration-sensitive prostate cancer (mCSPC): 10 clinical trials, 22 full-text publications, 23 variables across trial characteristics, population characteristics and meta-analysis outcomes (overall and subgroup). Publications were randomly split into a prompt-development set (n=5 publications, 115 responses) and a held-out test set (n=17, 391 responses). Pipeline: rule-based preprocessing in Python (PyMuPDF text extraction; removal of references, acknowledgments, appendices, author lists etc.; token-aware chunking with 1,000-token overlap; table/figure pages captured as 500-DPI images, 169 images total). GPT-4-turbo (gpt-4-0125-preview text, gpt-4-vision-preview images) and Claude-3-Opus (claude-3-opus-20240229) extracted independently at temperature 0. Post-processing consolidated duplicate/conflicting chunk-level answers via GPT-4T. Concordance required exact numerical matches or semantic equivalence (expert-judged) for text. Discordant pairs were cross-critiqued: each model received the source plus the other's response and instructions to verify. Evaluation against the gold standard used accuracy, precision, recall, F1 and hallucination rate with 95% CIs; Wilcoxon signed-rank tests compared approaches; a sensitivity analysis excluded trials overlapping the dev set.

Key results and numbers: Prompt-development set: 110/115 (96%) concordant, concordant accuracy 0.99 (95% CI 0.98–1.00), 0% hallucinations; single-LLM accuracy 0.95 for both models; 3/5 discordant pairs became concordant after cross-critique (accuracy 1.00). Test set: 342/391 (87%) concordant, accuracy 0.94 (0.93–0.95), precision 1.00, recall 0.92, F1 0.96, hallucinations 0.25%. Concordant responses significantly outperformed single LLMs (GPT-4T accuracy 0.89, F1 0.92; Claude-3O accuracy 0.90, F1 0.92; p-values 0.003–0.035). Discordant responses were poor: GPT-4T accuracy 0.41, Claude-3O 0.50, with hallucination rates of 26.9% and 41.0% respectively in that subset. Cross-critique converted 25/49 (51%) discordant responses to concordant (accuracy 0.76, hallucinations 8.3%); 24/49 stayed discordant (GPT-4T accuracy 0.18; Claude-3O 0.45; hallucinations 56.3%/47.9%). Sensitivity analyses (unique trials only) were consistent (concordant accuracy 0.97; post-critique 0.85). Cost: $3.40 per publication for extraction plus $2.06 for cross-critique, "in almost real-time." Trial-characteristic variables approached perfect accuracy; variables needing implicit judgement lagged.

Conclusions and takeaways: Concordance between two independent LLMs is a cheap, effective proxy for correctness in data extraction, and cross-critique — the LLM analogue of reviewer discussion — rescues about half of disagreements. The collaborative approach beats either model alone on accuracy and hallucinations, is generalisable via templated prompts to new variables/articles, and could unblock "truly 'living'" reviews whose bottleneck is extraction

(manual extraction error rates up to 50% are cited). Residual disagreement should route to a third reviewer, human or LLM. External validation on other datasets is explicitly required before widespread use.

Limitations and caveats: Only prostate-cancer oncology trials; only 23 variables; small sample (22 publications, 391 test responses); possible test contamination (different publications of the same trial in dev and test — addressed but not eliminated by sensitivity analysis); possible LLM training-data exposure to the publications themselves (overestimating performance); concordance for text variables judged by expert reviewers, adding subjectivity; and the preprint is not peer-reviewed. Per-category discordance confidence intervals are wide (e.g., Claude dev-set hallucination CI 0–100%). The two-reviewer simulation still lacks the diversity of real reviewer disagreement.

Practical implications for a non-academic researcher using PubMed and AI: You can replicate the core trick cheaply: run the same extraction or summarisation question through two different models (e.g., two chatbots), and treat agreement as a weak quality check and disagreement as a flag for manual verification. Expect ~87–96% raw agreement on factual extraction and remember that disagreement correlates with a high hallucination rate — precisely the cases where you must read the source. Temperature 0 (or equivalent "deterministic" settings) helps reproducibility.

Notable quotes:

- "In instances of discordant responses, cross-critique can further increase the accuracy."
- "The average cost for data extraction was $3.40 for one publication, with an additional cost of $2.06 for crosscritique in almost real-time."
- "Large language models, when simulated in a collaborative, two-reviewer workflow, can extract data with reasonable performance, enabling truly 'living' systematic reviews."

Connections: Shares the Mayo Clinic evidence-synthesis lineage and a co-author (M. Hassan Murad) with Wang (2020): Error rates of human reviewers during abstract screening — Khan cites Mathes on extraction-error frequency and the same cost literature; Wang's human error rates are the yardstick the collaborative approach implicitly beats. Its two-reviewer framing parallels Li (2025): Enhancing systematic literature reviews with generative AI's human-in-the-loop SLR and Hanegraaf's ϰ benchmark; both are oncology, GPT-4-based, temperature-controlled. Its dependence on repeat/cross prompting sits squarely in Gallo (2025): Best practices in LLM research: repeat prompting's territory — correlated outputs and statistical caution. Its hallucination quantification (0.25–50%) gives concrete numbers for the concerns catalogued in Yarar (2026): Concerns of using LLMs in health care research: umbrella review and echoes Lund (2024): ChatGPT in medical libraries: possibilities and future directions's "verify against trusted sources."

Pilz (2024): Semi-automated title-abstract screening with NLP and machine learning — Semi-automated title-abstract screening with NLP and machine learning (Systematic Reviews 2024)

What the paper is and why it matters: A methods paper from German biometry/surgery groups (Heidelberg, Fraunhofer ITWM, Halle-Wittenberg) demonstrating, in full practical detail, how to semi-automate title-abstract (TIAB) screening — the bottleneck of systematic reviews — using classical NLP + ML plus fine-tuned language models, validated on two *real* systematic reviews with meta-analysis. Its distinctive value is practical guidance: it explicitly computes

the needed training-set size, derives a defensible inclusion cutoff, and walks through an applied workflow, whereas prior work was either proof-of-concept or inflexible closed tools.

Methods in detail: Two real review datasets: INTRISSI (intravenous/oral antibiotic prophylaxis and mechanical bowel preparation to prevent surgical site infection in elective colorectal resection; 4,958 citations, 4,460 screened after deduplication/language/abstract filters; 225 sent to full text; 128 finally included; 10,164 word stems appearing ≥2×) and PPU (surgical vs alternative treatment of perforated peptic ulcer; 1,343 citations screened; 79 to full text; 4,370 stems). Text processing: lowercasing, lemmatization *plus* Porter stemming (to combat the p>n problem), removal of special characters/numbers, stopword and junk-word deletion against Grady Ward's English word list, building a document-term count matrix. Four classifiers — support vector machine, logistic regression, random forest, LightGBM — were trained with elastic-net variable selection (applied only on training data) and grid-search/five-fold-CV hyperparameter tuning. Additionally, two language models were fine-tuned: SciBERT (BERT trained on scientific text) and GPT-2, *without* hyperparameter tuning for runtime reasons. Training size was grown iteratively from 10% to 90% of data in 10% steps; AUC (bootstrap CIs) was the primary metric given class imbalance. A novel cutoff rule: in each CV fold, the cutoff is the smallest predicted inclusion probability among truly included publications; the K fold cutoffs are aggregated as minimum, median or mean.

Key results and numbers: INTRISSI: random forest AUC exceeded 87% with only 447 training publications and stabilised from 1,341 publications (≈69 inclusions) onward; tree-based methods and fine-tuned LMs reached AUCs above 0.95 (up to 97–98% at the largest training sizes); SciBERT/GPT-2 were weak with small training data, then equalled but did not outperform the tree ensembles. Logistic regression and SVM consistently lagged. PPU (smaller): larger samples needed for AUC >80%; from ~1,000 training publications the language models were best, and *only SciBERT* exceeded 90% AUC; logistic regression/SVM again underperformed. Worked illustration (INTRISSI, 40% training = 1,788 publications, test 2,672): random forest with the minimum CV cutoff (0.219) achieved a true-positive rate of 1.0 (no relevant paper missed) but a true-negative rate of only 0.465 (1,491 forwarded to full text); a cutoff near 0.3 would yield only 2 false negatives. Time model: traditional dual screening reads 4×4,460=17,840? — no: 8,920 abstract-readings (two reviewers × 4,460) vs 7,606 in the ML workflow, saving 1,314 minutes (>21 hours) at 1 minute/abstract.

Conclusions and takeaways: NLP+ML pipelines can credibly semi-automate TIAB screening: with ≥1,000 labelled (double-screened) publications, tree ensembles or fine-tuned SciBERT reached AUC >0.95 in one project and ~0.85–0.92 in another. The authors recommend a per-project method comparison (invoking the no-free-lunch theorem), conservative cutoff choice (minimum of CV cutoffs) when missing a relevant study is unacceptable, and hybrid workflows — e.g., machine as second reviewer with human arbitration on disagreements, or full ML triage with human checks near the cutoff. Full-text automation is flagged as a harder future problem (noise, figures, tables).

Limitations and caveats: Both case studies are surgical RCT reviews with relatively small inclusion counts; generalizability to other review types is asserted, not shown. The cutoff evaluation uses test-set knowledge unavailable in practice, and the procedure provides "no estimate of the procedure's recall" — a candid, important limitation. Language models were denied hyperparameter tuning and cutoffs for runtime reasons, disadvantaging them unfairly; the newest models (e.g., GPT-4) were unavailable. The true-negative rates were non-monotonic with training size (random batch effects). Only two datasets, one domain; the 1-minute-per-abstract assumption is borrowed.

Practical implications for a non-academic researcher using PubMed and AI: The key transferable lessons: (1) ML screening needs *labelled* examples — expect to hand-screen several hundred to ~1,000+ abstracts before automating the rest; (2) pick the cutoff to protect recall, accepting many false positives; (3) beware imbalance (~2–5% inclusion rates are normal) which inflates apparent accuracy; (4) if you cannot code, the logic still applies in tools like Rayyan/ASReview that the paper situates. The paper's honest framing — machine as second reviewer, human arbitrating disagreements — is directly adoptable.

Notable quotes:

- "mistakenly excluding relevant publications is worse than including too many publications since the first error cannot be corrected during full-text screening."
- "the training sample should contain at least 69 inclusions"
- "tree-based machine learning methods show a comparable performance to freely available fine-tuned language models."

Connections: The pre-LLM counterpoint to Li (2025): Enhancing systematic literature reviews with generative AI (which argues LLMs remove the annotation burden Pilz's pipeline requires) — Li even cites the ML screening-savings literature (61–79%) that frames Pilz's field. Its error-averse cutoff philosophy aligns with Wang (2020): Error rates of human reviewers during abstract screening's finding that ~1 in 9 human screening decisions errs. Its AUC/kappa-era evaluation contrasts with Gallo (2025): Best practices in LLM research: repeat prompting's warnings about LLM-specific statistical pitfalls. Khan (2025): Collaborative LLMs for automated data extraction in living systemat… and Mahboubi (2025): SciScribe: automating and contextualizing literature reviews handle later pipeline stages (extraction, curation) that Pilz explicitly defers to future work. Lund (2024): ChatGPT in medical libraries: possibilities and future directions's librarian audience is precisely who Pilz says lacks programming-friendly guidance.

Mahboubi (2025): SciScribe: automating and contextualizing literature reviews — SciScribe: literature-review curation platform in cardiac surgery (J Thorac Cardiovasc Surg 2025)

What the paper is and why it matters: A Cleveland Clinic–IBM collaboration presents SciScribe, a web platform built on IBM's Deep Search document-processing engine, designed not to replace PubMed-style search but to automate the *curation* steps after search: extraction, organisation, contextualisation and grounding of full-text publications — demonstrated via a cardiac-surgery use case on aortic valve replacement (AVR) and structural valve deterioration (SVD). It matters as a concrete "curation not search" archetype, notable for its emphasis on traceability (grounding) and for linking literature to official records (ClinicalTrials.gov NCT numbers, CMS OpenPayments industry payments) — a conflict-of-interest context flag that few tools offer.

Methods in detail: SciScribe ingests PubMed Central (PMC) full-content publications (5,084,756 documents, coverage to February 2023), plus PDFs converted via computer-vision pipelines (table structure recognition included), and structured records from ClinicalTrials.gov (to August 2022) and OpenPayments (2016–2022). Everything is normalised to a single structured format and indexed by content and semantic similarity; authors, institutions and identifiers (NCT numbers) are detected for record linking. Keyword search operates across millions of documents (BM25-style hit ranking); personal collections of hundreds-to-thousands of documents support natural-language question answering via Retrieval Augmented Generation (RAG): retrieve semantically relevant passages, pass to an LLM, compose an answer with the supporting

passage highlighted in the source. The demonstrated workflow: search with key phrases → contextualised results with industry-payment/trial flags → per-document conversational Q&A to judge relevance → build a private library (101 relevant publications in the use case) → ask "component questions" (e.g., "what anatomic factors improve freedom from SVD?") across the whole library → auto-generate a review table (rows=documents, columns=questions), each cell clickable to its grounding passage → synthesize the final review.

**Key resul

ts and numbers:** This is a demonstration/position paper, not an evaluation study, so no accuracy, $\varkappa$, F1 or error-rate metrics are reported. The concrete quantities are corpus and workflow figures: >5,084,756 PMC documents indexed; coverage dates (PMC to Feb 2023, ClinicalTrials.gov to Aug 2022, OpenPayments 2016–2022); 101 relevant publications collected for the AVR/SVD question. Sample LLM answers are reproduced verbatim — including one exemplary refusal where the model correctly states the abstract "does not provide any details about anatomic and procedural factors that are associated with these outcomes." A planned pilot study is described (n=30–50 reviewers in three arms — SciScribe, Manual, Other AI — answering the same question within 2 weeks, with quality/fidelity/veracity and time compared, plus pre/post surveys), but no results exist yet in this paper.

Conclusions and takeaways: Generative-language techniques can automate the repetitive steps of literature review — accumulation, Q&A-based relevance triage, structured extraction into review tables — while grounding every answer in source passages and adding external context (trial registrations, industry payments) that shapes interpretation. The authors are careful about limits of trust: SciScribe "cannot systematize the review, substitute for the reviewer's necessary scrutiny, or interpret statistics"; context flags are informational, not judgments; and convenience can breed bad habits (not checking grounding, over-narrow questions). They propose safeguards like requiring the reader to confirm the grounding of every table cell, "akin to a surgical checklist," and call for empirical evaluation.

Limitations and caveats: No quantitative evaluation whatsoever — all claims of acceleration and error mitigation are asserted, not measured (the pilot study is future work). Coverage is closed-access-bounded: 5 million PMC documents "represent a fraction of the 35 million covered by MEDLINE," and paywalled/open-access licensing restricts ingestion. Hallucination risk is acknowledged (LLMs "fall back on information that was baked in" during training); the authors mitigated it by decomposing questions, but did not quantify residual error. Contextual links can be wrong (name-matching to OpenPayments can create "false flags"). The primary literature corpus was frozen months before publication, so recency is limited. As a journal "Perspective"-style piece with vendor involvement (IBM-CCF Discovery Accelerator), it is promotional in genre, though conflicts are declared as none.

Practical implications for a non-academic researcher using PubMed and AI: The workflow is imitable in spirit with free tools: (1) keep a personal collection of full texts (PMC is open); (2) ask each paper targeted questions and demand the model point to the passage it used; (3) tabulate answers into a comparison spreadsheet; (4) enrich with ClinicalTrials.gov lookups via NCT numbers. The grounding discipline — never accept an answer you cannot click back to a passage — is the single best practice this paper teaches, and it matches the "verify against trusted sources" mantra of the library literature.

Notable quotes:

- "there is a need to go beyond search tools and toward curation tools that support the entire review process with rich grounding and context."

- "SciScribe ranks search results, and subsequent distilled answers, according to their similarity to the search criteria. This ranking does not reflect a judgment of the papers."
- "Don't just review a subject, SciScribe it!"

Connections: The curation-stage complement to the screening/extraction automation of Pilz (2024): Semi-automated title-abstract screening with NLP and machine learning, Li (2025): Enhancing systematic literature reviews with generative AI and Khan (2025): Collaborative LLMs for automated data extraction in living systemat… — all four map onto the SLR pipeline's different stages. Its PMC grounding and explicit hallucination mitigation address the concerns later systematised by Yarar (2026): Concerns of using LLMs in health care research: umbrella review. Its "reader remains in charge" stance is the operational version of Lund (2024): ChatGPT in medical libraries: possibilities and future directions's librarian-as-verifier role. Its reliance on Q&A repetition invites Gallo (2025): Best practices in LLM research: repeat prompting's correlation caution if such tools are ever formally evaluated, and its contextual conflict-of-interest flags speak to the bias/ethics themes across the batch.

Wang (2020): Error rates of human reviewers during abstract screening — Human error rates during abstract screening in systematic reviews (PLoS ONE 2020)

What the paper is and why it matters: This is the reference-standard quantification of how often trained human reviewers err during abstract screening — the empirical foundation for the whole automation field. The argument is simple but consequential: automation tools are validated against human "gold standards," yet if humans themselves err on roughly 1 in 9 abstract-level decisions, then demanding perfection from machines is incoherent, and human-level error rates may suffice — especially for a tool acting as a second reviewer. It remains the most-cited justification for sensitivity-tolerant, semi-automated screening (cited directly by Li (2025): Enhancing systematic literature reviews with generative AI and, institutionally, connected to Khan (2025): Collaborative LLMs for automated data extraction in living systemat…'s Mayo team).

Methods in detail: Retrospective analysis of all systematic reviews completed at a US AHRQ-funded Evidence-based Practice Center (one of 12) between June 1, 2010 and December 31, 2017. Eligibility: standard dual independent screening of titles/abstracts and full texts, with the rule that *inclusion by either reviewer at abstract stage automatically promoted the citation to full-text screening*; full-text disagreements resolved by consensus or third-reviewer arbitration; use of DistillerSR software to log every inclusion/exclusion decision; no automation used in the reviews themselves. An "error" was defined as any single-reviewer abstract-stage decision inconsistent with the final included study list (false inclusion or false exclusion). Error rate = errors ÷ total screened decisions (citations × 2). Analyses computed per-review outcomes, then inverse-variance-weighted means across reviews (Stata 15.1), stratified by clinical area (9 areas) and question type (methodology, diagnostic/screening/prognostic, treatment).

Key results and numbers: 25 systematic reviews; 139,467 citations; 329,332 screening decisions; 85 unique reviewers (28 core investigators, 57 external collaborators; the abstract says "86," an internal inconsistency). Final inclusion rate: 5.48% of citations (95% CI 2.38–8.58%); abstract-stage inclusion rate: 18.07% (12.65–23.48%); total abstract-screening error rate (false inclusions + false exclusions): 10.76% (7.43–14.09%) — "an error about 1 in 9 abstracts." Variation was large: by clinical area, 5.76% (hematology) to 21.11% (primary care… per Table 2 mental health/primary care figures reach 17.73–21.11%); by question type, methodology 8.73%, diagnostic/

screening/prognostic 14.97%, treatment 10.57%. Context figures cited: ~0.9 min/reference for abstract screening, 7 min full text, 53 min data extraction; 1,000 retrieved studies ≈ 952 hours; estimated ML workload reductions of 30–70%; 39 automation tools available by July 2019.

Conclusions and takeaways: Human reviewers — even in a rigorous dual-screening EPC environment — make substantial, variable errors at abstract screening. Therefore, "when developing and refining an automation tool, achieving error rates similar to humans may be adequate," and such a tool could legitimately "serve as a single reviewer that gets paired with a second human reviewer." The paper reframes validation of automation: the comparison standard is imperfect, so evaluation should target human-comparable error rather than zero error.

Limitations and caveats: Stated by the authors: small sample, especially per clinical-area strata (some n=1 reviews); findings may not generalize to other topics or teams with different expertise mixes. Crucially, the automatic-inclusion-on-conflict rule means *dual erroneous exclusions are invisible* — when both reviewers wrongly exclude, the citation vanishes and cannot be counted, so the measured 10.76% is a lower-bound-flavoured composite. Full-text screening error was not measurable (conflicts untracked). Some "errors" reflect poor or insufficient abstract reporting rather than reviewer fault — "they are not the fault of human reviewers." Single-institution, single-software setting.

Practical implications for a non-academic researcher using PubMed and AI: Three practical anchors: (1) expect and budget for error — roughly 10% of abstract decisions will be debatable or wrong even with careful dual screening; (2) structure your own (AI-assisted) screening so inclusion by *either* you or the tool triggers full-text review — erring on the side of inclusion is cheap at abstract stage; (3) don't dismiss an AI screener for being imperfect; benchmark it against the ~10.76% human baseline instead of an imaginary 0%. Also note the efficiency arithmetic (0.9 min/abstract) when estimating how much screening time automation could return to you.

Notable quotes:

- "Yet, human reviewers make errors in inclusion and exclusion of references."
- "an error about 1 in 9 abstracts"
- "it is important to keep in mind that the gold standard is not perfect and that achieving error rates similar to humans is likely adequate and can save resources and time."

Connections: The empirical pivot of this batch: Li (2025): Enhancing systematic literature reviews with generative AI cites it as ref 10 and builds the human-in-the-loop system it motivates; Pilz (2024): Semi-automated title-abstract screening with NLP and machine learning's recall-first cutoff logic operationalises its inclusion-bias lesson; Khan (2025): Collaborative LLMs for automated data extraction in living systemat… shares the Mayo evidence-synthesis milieu and co-author (M. Hassan Murad) and extends the error lens to data extraction; Gallo (2025): Best practices in LLM research: repeat prompting flips the perspective — the *evaluation studies* built on Wang's framing can themselves be statistically flawed; Yarar (2026): Concerns of using LLMs in health care research: umbrella review and Lund (2024): ChatGPT in medical libraries: possibilities and future directions supply the qualitative concerns (fatigue, bias) that make Wang's error rate psychologically plausible.

Gallo (2025): Best practices in LLM research: repeat prompting — Best practices in LLM research: repeat prompting and correlated outputs (JAMIA 2025)

What the paper is and why it matters: A brief communication (Stanford/VA Palo Alto) that delivers a statistical warning shot to the entire medical-LLM evaluation literature: when researchers prompt an LLM many times to inflate sample size, the repeated outputs are *highly correlated*, and analysing them as independent observations produces artificially narrow confidence intervals and tiny P-values — enough to fabricate "findings." Using real data from a published JAMA study claiming affiliation bias in LLM peer review, the authors show that the correct mixed-effects analysis erases the original result entirely. It matters because repeat prompting is ubiquitous in medical LLM studies (they cite several in JAMA, Nat Med, Lancet Digital Health, NEJM AI).

Methods in detail: Secondary analysis of the von Wedel et al. dataset: GPT-3.5 was prompted to accept/reject 30 medical abstracts, each duplicated with 30 university affiliations across three tiers (tier labels themselves elicited from ChatGPT: 10 "top-tier," 10 "mid-tier," 10 "low-tier"), prompted 250 times per abstract-affiliation combination = 225,000 observations. The original analysis used a difference-in-proportions test (equivalently simple logistic regression assuming independence). Gallo et al. refit the data with a mixed-effects logistic regression including random effects for both abstract and affiliation (either alone fails to capture correlation at the abstract-affiliation level), compared model fit via AIC/BIC/likelihood-ratio tests, computed the intraclass correlation coefficient (ICC) from the random-effects model, derived the effective sample size per grouping, and ran cluster-trial-style power calculations (300 groupings per arm, 5-percentage-point detectable difference, 35% comparator acceptance rate). A supplementary example used their prior diagnostic-reasoning study with 5 prompt *variations* to test whether prompt variety reduces correlation.

Key results and numbers: Original independent-analysis odds ratios were small but "highly significant": top-tier vs mid-tier OR 1.04 (95% CI 1.02–1.06), P<.001; low-tier vs mid-tier OR 0.97 (0.95–0.99), P=.002. Under mixed-effects modelling both became non-significant: OR 1.07 (0.95–1.22), P=.27 and OR 0.94 (0.83–1.07), P=.36. The random-effects model fit better (AIC 181,072 vs 181,579; BIC 181,123 vs 181,621; likelihood ratio P<.001). ICC = 0.69, implying an effective sample size of just 1.45 of the 250 repeats per abstract-affiliation group — a >100-fold inflation of effective N. Statistical power fell from 100% to 33% when clustering was accounted for. The supplementary prompting-variation example showed ICC 0.73 — variation in prompts did *not* dissolve correlation. Acceptance rates by tier: 38.4% (top), 37.5% (mid), 36.7% (low).

Conclusions and takeaways: Repeat-prompted LLM outputs are strongly correlated (contradicting suggestions that they behave like independent samples); ignoring this violates independence assumptions and can "essentially nullify" a study's conclusions. Repeat prompting itself remains valuable for measuring stability and uncertainty — but analysis must use methods from clustered data (mixed-effects models, cluster-randomised-trial power formulas). The field needs best practices, reporting guidelines (e.g., TRIPOD-LLM, in progress) and evaluation frameworks (e.g., UK AI Safety Institute's "Inspect"); ICCs may vary by task, and lower temperature settings likely raise correlation (though GPT models are not fully deterministic even at temperature 0). Even modest ICCs (0.1–0.3) sharply reduce effective sample size and power, so correction is always necessary.

Limitations and caveats: It is a single reanalysis of one dataset with one model (GPT-3.5) and one task (abstract peer review); the observed ICC (0.69) "is relatively high, and other studies

may find lower correlation." No new primary data; the supplementary variation example is small and also high-ICC. The authors appropriately note that model choice should follow the data-generating process rather than fit tests, and that typical ICCs across task types (diagnosis, summarisation, QA, peer review) remain uncharacterised. As a brief communication, details of the mixed-model specification live in supplementary code.

Practical implications for a non-academic researcher using PubMed and AI: If you read (or plan to write) any study claiming LLM superiority or bias based on asking a model the same question many times, check how repeats were handled — a claim built on 10,000 repeats of 20 distinct prompts may rest on ~20 effective observations. In your own use, don't treat "I asked 5 times and 4 answers agreed" as 80% confidence; correlated repeats mean agreement is weaker evidence than it looks — which also tempers naive versions of the two-model agreement trick in Khan (2025): Collaborative LLMs for automated data extraction in living systemat…. Ask the same question once per model, then compare across models.

Notable quotes:

- "This is analogous to prompting a single individual to answer a survey question 250 times in a manner that they do not remember their previous responses, but then analyzing the data using methods that assume the responses came from 250 different individuals."
- "Properly accounting for repeat prompting effectively decreased sample size by over 100-fold and decreased power by a factor of 3 in this case."
- "accounting for correlation in repeated prompting of LLMs is critical for valid study design and even reversed study conclusions in this case."

Connections: The methodological watchdog for the batch: Li (2025): Enhancing systematic literature reviews with generative AI and Khan (2025): Collaborative LLMs for automated data extraction in living systemat… both rely on repeated or dual prompting — Khan's parallel-LLM concordance is a *between-model* design that partially sidesteps Gallo's within-model concern (though temperature-0 determinism itself raises correlation questions), and Li's iterative screening rounds deserve the same scrutiny. Its bias-study subject matter feeds directly into Yarar (2026): Concerns of using LLMs in health care research: umbrella review's theme that bias claims need rigorous, reproducible evaluation. It implicitly validates the conservative stance of Lund (2024): ChatGPT in medical libraries: possibilities and future directions and the human-error realism of Wang (2020): Error rates of human reviewers during abstract screening (claims of machine superiority need statistics that survive clustering) and underscores why Mahboubi (2025): SciScribe: automating and contextualizing literature reviews's planned comparative pilot needs a proper design.

Yarar (2026): Concerns of using LLMs in health care research: umbrella review — Umbrella review: concerns of using LLMs in health care research and practice (JMIR 2026)

What the paper is and why it matters: From Newcastle University's NIHR Innovation Observatory, this is claimed as the first umbrella review of *concerns* about LLMs across health care research and practice, synthesising 42 systematic reviews covering 12 distinct populations (researchers, general clinicians, plastic surgeons, psychiatrists, neurosurgeons, and 7 more specialties). It matters because it triangulates, in one place, the worry-list that individual primary studies (including several in this batch) address piecemeal — hallucinations, bias, transparency, privacy, cost — and because its scathing quality appraisal of the underlying review literature is itself a finding about the field's maturity.

Methods in detail: Protocol registered on PROSPERO (CRD420250640997). SPIDER framework; inclusion: systematic reviews published from January 2017, in English, with LLM/GenAI-use concerns as an outcome, in health/social care research contexts; excluded: preprints, letters, editorials, conference abstracts, commentaries, primary studies. Seven databases searched (Ovid MEDLINE, Ovid Embase, Scopus, Web of Science, JBI, Cochrane DSR, Epistemonikos) on 26 February 2025, updated 25 February 2026; de novo MEDLINE strategy peer-reviewed by an information specialist; no language limits on searching but English publication required for inclusion. Rayyan-based duplicate independent title/abstract and full-text screening by 2 reviewers (disagreements to a third). Data extraction: 50% double-extracted. Risk of bias via AMSTAR-2 (independently, 2 reviewers); certainty via GRADE for qualitative syntheses; narrative thematic synthesis (29 inductive/deductive codes → 7 themes → 3 overarching themes) following SWiM and reported per PRIOR/PRISMA/PRISMA-S. No meta-analysis (qualitative data).

Key results and numbers: 449 records identified (abstract says 448 retrieved — minor inconsistency); 63 duplicates removed; 386 title/abstract screened; 317 excluded; 69 full-text assessed; 27 excluded (most often: concerns not addressed); 42 reviews included. Populations: researchers most common (n=17), then general clinicians (5), plastic surgeons (5), psychiatrists (4), neurosurgeons (3), plus 8 single-review specialties. LLM focus: 15 ChatGPT-only, 15 multiple LLMs, 12 generic AI/NLP/etc. Median 32 primary studies per review (IQR 19–83; range 5–315). Quality: AMSTAR-2 rated 2 reviews (4.8%) low and 40 (95.2%) *critically* low; only 14.2% had risk-of-bias assessment, 23.8% a search strategy, 19% complete double data extraction; GRADE certainty: 21.4% low, 78.6% very low. Reviewer $\varkappa$: 0.92 (AMSTAR-2), 0.75 (GRADE). Findings include: ~a quarter of LLM studies show bias (~15% on sensitive topics); image references anatomically incorrect/fabricated in 81% of examined cases; ~40% of AI-written discharge summaries contain hallucinations; ChatGPT-4 scored 85% on neurology licensing questions vs 73.8% human; fracture-detection 83%; only 26% of user-evaluation studies used RCT-type designs; female ophthalmologists recommended less than one-third as often as males. Interrater-$\varkappa$ testing was done in only 2 of 42 constituent reviews (5.1%).

Conclusions and takeaways: Three overarching concern themes, in frequency order: (1) technical capability (dominated by data quality/reliability, transparency/"black box", performance limits, hallucinations and repeatability problems); (2) ethical, legal and societal (accountability, academic integrity, privacy/GDPR, consent, bias and entrenched inequalities, doctor–patient relationship, deskilling, infodemics); (3) costs (environmental/computational, financial — least developed). The authors argue data quality "is at the heart of these concerns," call for research integrity policies, reporting guidelines and audits, propose replacing the stigmatising term "hallucination" with "AI misinformation"/"fact fabrication"/"stochastic parroting," and warn against labelling any LLM "completely transparent" since implicit bias persists in even "explicitly unbiased" models. Adoption should proceed "with equity, diversity, inclusion, and safety at the core."

Limitations and caveats: Self-declared: the constituent reviews are of "very poor quality" (95.2% critically low AMSTAR-2), certainty is mostly very low, and primary-study overlap could not be computed (corrected covered area impossible due to poor reporting), so findings "may be overinflated." Thematic analysis was performed by a single reviewer (mitigated by group discussion; blinded dual coding recommended for future work). Heterogeneity and sensitivity analyses were not assessed; search-date limits in a fast-moving field; English-only inclusion despite no language limits at search. Table 1′s header confusingly reads "39 included studies" while 42 were included — an editorial inconsistency. All quantitative claims (e.g., 81% fabricated images,

40% hallucinated summaries) are second-hand from low-quality sources and should be treated as indicative, not established.

Practical implications for a non-academic researcher using PubMed and AI: Use it as a due-diligence checklist: before relying on an LLM for literature work, assume (a) outdated knowledge (ChatGPT-3.5 pretrained to Sept 2021; GPT-4 to Apr 2023 — so always verify recency in PubMed), (b) fabricated references and confidently delivered errors, (c) demographic and topical bias, (d) unclear provenance. Disclose where/how AI was used; never cite a reference you haven't opened; treat the tool as a "supplementary tool," a phrase the review found recurring across specialties. The cost theme is a reminder that heavy LLM use has financial and environmental overheads worth limiting.

Notable quotes:

- "Data quality is at the heart of these concerns, and combative action must ensure health care professionals and researchers have the resources required to overcome these apprehensions if AI is to be used routinely."
- "the overall rating of reviews using AMSTAR-2 was either lo

w (n=2, 4.8% reviews) or critically low (n=40, 95.2% reviews)"

- "while we should strive for more transparency, particularly through legislation, guidance, and reporting, we should avoid labeling any LLM as completely 'transparent.'"

Connections: The umbrella that contextualises everything else in the batch: its hallucination/fabrication concerns are quantified at the study level by Khan (2025): Collaborative LLMs for automated data extraction in living systemat… (0.25–50% hallucination rates) and flagged as a limitation by Mahboubi (2025): SciScribe: automating and contextualizing literature reviews; its "supplementary tool" consensus echoes Lund (2024): ChatGPT in medical libraries: possibilities and future directions; its criticism of low-quality LLM evaluation research is exactly what Gallo (2025): Best practices in LLM research: repeat prompting demonstrates concretely via ICC analysis; its evidence-synthesis-specific worries (automation pushing "low-quality automated reviews," infodemics) frame the methodological rigour attempted by Li (2025): Enhancing systematic literature reviews with generative AI, Pilz (2024): Semi-automated title-abstract screening with NLP and machine learning and Wang (2020): Error rates of human reviewers during abstract screening, whose error-rate data show why human-in-the-loop verification must remain central.

14.5 Field maps and AI literacy

Analyses based strictly on pdftotext extractions of the eight source files. Minor extraction artefacts are noted where relevant.

Su (2025): LLMs in medical diagnostics: scoping review with bibliometric analysis — Scoping review with bibliometric analysis of LLMs in medical diagnostics (J Med Internet Res 2025;27:e72062)

What the paper is and why it matters: This is a hybrid scoping review and bibliometric mapping of research on large language models (LLMs) applied to medical diagnosis, published in JMIR by a large team from Xiangya Hospital, Central South University (China). It is the first quantitative, comprehensive charting of how LLMs are being integrated into diagnostic medicine, combining macroscopic trend detection (VOSviewer keyword clustering) with qualitative appraisal of study designs and validation metrics. It matters for the synthesis report because it demonstrates a methodological innovation — using bibliometric clustering *before* data

extraction to derive research subquestions — and because it documents the field's explosive growth, model dominance (GPT-4), performance benchmarks, and ethical/regulatory barriers, all directly relevant to how a non-academic researcher should understand what AI can and cannot do in clinical literature contexts.

Methods in detail: The review followed the Joanna Briggs Institute Manual for Evidence Synthesis and the PRISMA-ScR checklist. Five databases were searched — Web of Science, PubMed, Embase, IEEE Xplore, and ACM Digital Library — from 2022 to January 2025 (first search 27 September 2024; final search 12 January 2025). Records were imported into Rayyan; four authors independently screened in blind mode, with conflicts resolved by a fifth author. Inclusion required reporting outcomes of LLMs in medical-diagnosis-related fields; conference abstracts, comments, and retracted articles were excluded. In a distinctive step, the team exported Web of Science data into Microsoft Excel and ran a bibliometric analysis in VOSviewer *before* full-text extraction, identifying 65 keywords in 3 thematic clusters ("assessment," "application," "impacts") that then structured the extraction template. Two authors extracted data independently with consensus resolution, preceded by a pilot on 10% of articles. A "nested" bibliometric analysis within clusters 1 and 2 categorized performance evaluation into disease classification, medical question answering (QA), and quality of generated diagnostic content. Additionally, ClinicalTrials.gov was searched (17 March 2025) for real-world LLM trials.

Key results and numbers: From 17,757 records, 2,445 duplicates were removed, 6,708 records assessed, and 95 articles included. Publications surged after 2023, with a marked rise in 2024; the United States and China were the most productive countries, with the US showing the densest collaboration network. GPT-4 and variants dominated (70/95, 74%), followed by GPT-3.5 (34/95, 36%), specifically developed GPTs (9, 9%), Claude (9, 9%), Gemini (8, 8%), and Llama (8, 8%). Three development paths for task-specific GPTs were described: prompt tuning, fine-tuning (e.g., Med-PaLM 2), and pretraining from scratch (e.g., BioGPT on PubMed abstracts). Among diagnostic domains, "research has focused on radiology, which accounts for 16 (17%) studies," with psychiatry and neurology also prominent — the authors suggest this is because those specialties have "fewer objective tools." The ClinicalTrials.gov analysis found 23 trials across 11 specialties (oncology 6/23, 26%; ophthalmology 5/23, 22%); 16 (70%) provided diagnosis support, 6 (26%) direct diagnosis. On bias: mitigation via RLHF reduced GPT-3.5′s racial bias in cost prediction from 18% to 5% in GPT-4, but the "uncertainty rate" increased from 16.25% to 29.46%.

Conclusions and takeaways: LLMs "hold transformative potential for medical diagnostics" but require rigorous validation, bias mitigation, and multimodal integration. The authors propose a four-part road map: (1) multimodal AI integration (text, imaging, lab, genomic data with cross-modal alignment); (2) explainable AI, including "chain-of-reasoning frameworks" and adversarial testing, plus prospective LLM-assisted vs conventional workflow trials; (3) specialty-specific optimization benchmarks; (4) ethical and regulatory harmonization (bias auditing standards, federated learning, adaptive regulation with clinician oversight). They conclude the field can transition "from demonstrating diagnostic potential to delivering measurable improvements in patient outcomes" only through collaboration between AI researchers, clinicians, and policy makers.

Limitations and caveats: The authors list several: reliance on selected databases (e.g., excluding arXiv) may have omitted pertinent work despite citation screening; lack of longitudinal studies of long-term clinical impact; a temporal limitation because LLM technology evolves faster than the review cycle (studies after January 2025 may outpace findings); heterogeneity precluded meta-analysis, so only descriptive statistics were given; and some clinical trials did not report

which LLMs they used or did so without proper disclosure of model type. The LLM definition used ("more than one million parameters") is idiosyncratic and very low by modern standards.

Practical implications for a non-academic researcher using PubMed and AI: The paper shows PubMed-derived datasets (MedQA, PubMedQA) are the standard benchmarks for judging AI diagnostic claims — a useful filter when reading vendor or media hype: ask which model, which dataset, which comparator. Its evidence of race/gender/cost bias in GPT-3.5/4 outputs cautions against uncritically using general chatbots for clinically sensitive summaries. The observation that fine-tuning and prompt engineering raise real-world usability suggests that, for literature work, structured prompting beats casual queries. The regulatory fragmentation discussion (FDA software-as-medical-device, EU AI Act/GDPR) explains why published claims outpace approved clinical tools.

Notable quotes:

- "LLMs hold transformative potential for medical diagnostics but require rigorous validation, bias mitigation, and multimodal integration to address real-world complexities."
- "Such methods only provide a shield to prevent the occurrence of bias in diagnosis without changing its core."
- "It is suggested for other countries' researchers to strengthen international cooperation and communication for this research topic."

Connections:: Shares the scoping-review method with, but uniquely adds a bibliometric layer that informs extraction — a contrast useful for methods discussion.. Its finding of pervasive racial/gender bias connects to, which shows outputs also vary systematically across LLM tools.. The heavy dominance of GPT-4 (74%) contrasts with's multi-tool comparison and with, which found ChatGPT-4o outperforming Gemini Pro.. Its clinical-trial scan complements, who reviews how such evidence-synthesis claims should be appraised.. ; Johnson (2022)'s SR Toolbox would classify Su's search/screening tools (Rayyan); ; Verma (2024) is a parallel example of a small-scale, single-maintainer evidence-curation infrastructure.

Ma (2026): AI in implementation research: applications and recommendations — Scoping review: AI in implementation research (Transl Behav Med 2026;16:ibag070)

What the paper is and why it matters: This scoping review from Deakin University (Australia) is, by the authors' account, the first completed synthesis of how AI has been applied *within* implementation research — not as a clinical innovation to be implemented, but as a tool supporting implementation processes and research methodology. It matters because it defines a clean three-role taxonomy for AI in this space (AI as the innovation being implemented; AI as a support tool; AI as a methodological tool), deliberately excluding the first. Its findings — only seven eligible studies, mostly from high-income countries, none addressing causality — document an extremely immature evidence base and thereby function as a sobering counterweight to AI optimism in the evidence-synthesis literature.

Methods in detail: The review followed Arksey and O'Malley's framework and is reported per PRISMA-ScR. A protocol was developed a priori but not registered (available on request). Searches on 31 July 2025 covered MEDLINE Complete (EBSCOhost), Scopus, IEEE Xplore, and ACM Digital Library, 1 January 2018–31 July 2025, combining "artificial intelligence" terms (ML, NLP, generative AI, LLMs) with "implementation research" terms (implementation science, dissemination science, knowledge translation, translational research); a health sciences librarian helped develop the strategy. Grey literature was not searched. Selection ran in four

Covidence-managed stages: dual title/abstract screening, full-text review by two reviewers, a calibration "pilot" on a subsample of 12 full texts that triggered formal criteria refinement, and full re-assessment against the updated criteria. A piloted extraction form captured study characteristics, AI applications, implementation frameworks and outcomes, and mapping to the EPIS framework (Exploration, Preparation, Implementation, Sustainment) and Trinkley et al.'s six challenges (speed, sustainability, equity, generalizability, context–outcome relationships, causality). Synthesis was narrative and framework-guided.

Key results and numbers: The search identified 2,995 records; after 1,306 duplicates, 1,689 were screened, 187 went to full text, and 7 studies were included (published 2019–2025; six involving high-income countries). Applications fell into four categories: (i) implementation monitoring and feedback (Berkel et al.'s ML/NLP on session audio for fidelity; Bishara et al.'s "Opal" perioperative ML platform linked to EHRs); (ii) data-driven strategy selection and tailoring (Moussa et al.'s random-forest model predicting effective strategy–barrier matches "with high accuracy (≈97%)" from CFIR/TDF-coded barriers; Shin et al.'s LLM-customized translational design cards); (iii) literature and qualitative data synthesis (Scaccia and Scott's LDA topic modeling of 15 years of implementation literature; Rinne et al.'s ChatGPT summaries of an expert panel — fast but "lacked nuance and contextual depth"); and (iv) knowledge dissemination (Desmedt et al.'s NotebookLM-generated podcasts, judged engaging, "with many listeners unable to identify the content as AI-generated"). Only two studies used implementation frameworks (CFIR/TDF); only one measured implementation outcomes (fidelity, acceptability). No study addressed causality or mechanism testing.

Conclusions and takeaways: AI applications are at an early stage and are currently used "to optimize existing workflows rather than to advance explanatory aims of the field." Applications cluster in the implementation phase of EPIS, leaving exploration, preparation, and sustainment phases underexplored. Four future directions are proposed: deliberately embed AI within implementation frameworks (CFIR, TDF) to clarify mechanisms; examine how AI tools interact with established strategies; assess impacts on core implementation outcomes; and treat governance/transparency and equity as core components, including bias audits and participatory co-design.

Limitations and caveats: The authors acknowledge that the tiny eligible corpus (n=7) "constrain[s] the breadth of conclusions"; strict eligibility requiring a direct connection with implementation research may have missed studies using different terminology; exclusion of non-English publications may have omitted relevant research; the protocol was not registered; ethical considerations were acknowledged rather than directly examined; and the small number of studies precluded subgroup analyses. The equity category list also contains an internal numbering artefact ("(i)…(ii)…(iv)") in the source text.

Practical implications for a non-academic researcher using PubMed and AI: The review legitimates cautious use of LLMs for time-saving preliminary synthesis (ChatGPT summarizing expert panels) while documenting concrete failure modes: hallucination risk, loss of nuance, missed dissenting perspectives, prompt-dependent output quality, and the need for human editing. Its three-role taxonomy is a useful checklist for classifying any "AI in health" paper before citing it. The finding that AI rarely connects to theory warns the independent researcher that AI-assisted workflows lack validated evaluation standards — treat AI outputs as drafts requiring verification.

Notable quotes:

- "No completed prior review has systematically examined the literature on AI applications in implementation research."

- "AI applications are rarely grounded in established implementation frameworks or evaluated against recognized implementation outcomes, representing a missed opportunity to advance the science."
- "AI-generated podcasts were perceived as engaging and understandable; many authors assumed hosts were human."

Connections:: Methodologically a sibling of (PRISMA-ScR scoping review, similar databases including IEEE Xplore/ACM DL) but finds an almost empty field (7 vs 95 studies) — a sharp contrast for the synthesis… Its transparency/disclosure themes (it discloses using "Deakin Gem (powered by Claude Haiku 4.5)" for grammar/formatting) echo the AI-disclosure checklist items in.. Its concern that AI outputs vary by prompt links to's tool-variation findings and, which engineered prompts to a checklist.. Johnson (2022) would supply the scoping-review stage tools it implicitly used; ; Verma (2024) mirrors its librarian-supported search development.

Garry (2026): Exploring variation in research priorities generated by AI tools — Variation in research priorities generated by five AI tools (J Glob Health 2026;16:04037)

What the paper is and why it matters: This methodological paper from Queen's University Belfast and Stellenbosch University asks a deceptively simple question: if researchers use LLM chatbots to generate health research priorities, does the choice of tool matter? Comparing ChatGPT-4, Claude Sonnet 4, DeepSeek-V3, Gemini 2.5, and Perplexity AI 4 across four diseases (cancer, COVID-19, HIV, Alzheimer's), it shows outputs differ systematically and stably between tools. It matters because it is one of the first empirical demonstrations that "which AI tool" is a substantive methodological decision — directly informing any synthesis report on AI-assisted research workflows, where tool selection is rarely justified.

Methods in detail: The authors typed an identical prompt into each of five tools in default settings — "Please define, and list in order of importance, the 20 greatest research priorities that would lead to improvements in the treatment of cancer/HIV/COVID-19/Alzheimer's disease" — making 20 distinct requests on 9 July 2025. Similarity was quantified with SBERT-BioBERT sentence embeddings (Python via Google Colab): for each of tool A's top 10 priorities, its strongest match in tool B's top 20 was identified by highest cosine similarity score, and the procedure reversed (all 20 pairings, 200 strongest matches per disease, 800 in total). Question 1 (most/least similar tool) compared mean scores of matches involving vs not involving each tool. Question 2 rank-ordered pairings; priorities unique to a tool were those failing a 0.6 cosine-similarity cut-off. Reliability was tested by re-running the identical prompt at a second time point (31 July 2025) and with a "tweaked" role-play prompt for the two most dissimilar tools. Data are on the Open Science Framework.

Key results and numbers: Gemini was the most similar to the other tools: matches involving Gemini averaged 0.741 vs 0.714 otherwise (MD = 0.027; P = 0.001), significant for COVID-19 and cancer individually; Claude showed a weaker similar effect (MD = 0.016; P = 0.042). DeepSeek was the most dissimilar (MD = −0.032; P = 0.001, significant in all four diseases), with Perplexity also below average (overall MD = −0.024; P = 0.002). The most dissimilar pair was DeepSeek–Perplexity: the two lowest of the 20 pairings for HIV and among the lowest for the other diseases. Substantively, "DeepSeek tended to emphasise technical medical issues, while Perplexity generally emphasised issues relating to patient care and health system efficiency" (e.g., DeepSeek: APOE4-targeted gene therapy, tau aggregation inhibitors; Perplexity: caregiver support, care-model optimization, health-system preparedness). Notably, "none of the

cancer priorities fell below this threshold" — cancer priorities converged most (mean 0.771–0.804). The DeepSeek–Perplexity distinction persisted across re-runs and the tweaked prompt, with Perplexity consistently stressing public health, equity, and low-resource contexts.

Conclusions and takeaways: If a single AI tool must be used, Gemini offers the least idiosyncratic outputs; researchers wanting multiple perspectives should combine DeepSeek and Perplexity with a representative tool, since the pair maximizes coverage of the two divergence poles (technical vs public health priorities). The recommendations are explicitly provisional. The authors frame tool variation as one component of a larger program on AI in priority-setting, flagging unresolved questions about validity (do AI priorities replicate human-generated ones?), transparency, and geographic/disciplinary/socioeconomic biases in training data — and speculate about industry convergence vs divergence (e.g., culturally aligned models).

Limitations and caveats: The study is explicitly exploratory; it "is not evaluating how 'good', 'true', 'accurate', or 'valuable' the priorities are." Only one date per condition (plus one re-run), default settings only, a single prompt template (plus one tweaked version), top-10/top-20 comparison windows, and a 0.6 similarity cut-off are all design choices that could affect results. Rapid LLM version turnover means patterns may not persist; the authors themselves call for more time points, more tools, and more nuanced prompts.

Practical implications for a non-academic researcher using PubMed and AI: The core lesson is methodological hygiene: an AI-generated list (of priorities, gaps, or key papers) is a property of the specific model and date, not a neutral view of a field. A non-academic researcher doing desk research should either state which tool/version was used, or triangulate across two deliberately different tools (e.g., a technical-leaning and a public-health-leaning one) to avoid systematic blind spots. The SBERT/cosine-similarity approach also shows a replicable, low-cost way to compare AI outputs rather than eyeballing them.

Notable quotes:

- "Our exploratory analysis suggests that Gemini performs well for researchers who prefer to generate health-related research priorities using a single AI model."
- "Opening the 'black box' of how AI produces research priorities must be a priority in itself."
- "We are not evaluating how 'good', 'true', 'accurate', or 'valuable' the priorities are… Rather, we are simply identifying whether the tool would produce relatively unusual priorities or would instead produce ones that are typical."

Connections:: Directly tests the cross-tool generalizability assumed (or ignored) in and, where ChatGPT dominates; ; its finding that DeepSeek/Perplexity diverge substantively cautions against benchmark comparisons that treat models as interchangeable.. Its concern with output stability across re-runs and prompts parallels's one-month stability check (ICC 0.92–1.00) — a notable difference in rigour vocabulary (cosine similarity vs ICC).. Its priority-setting framing connects to's reviewer checklist criterion A1 (is there a justified rationale?) — both scrutinize how research agendas get set..'s emphasis on AI transparency and equity mirrors Garry's "black box" argument; ; ; Johnson (2022) and Verma (2024) offer the human-built tool catalogues against which AI-generated priority lists must be verified.

Demir (2025): AI-driven abstract generating under the PRISMA-A framework — LLM-generated PRISMA-A abstracts: ChatGPT-4o vs Gemini Pro (BMC Oral Health 2025)

What the paper is and why it matters: This Turkish orthodontics study is the first in dentistry to systematically evaluate whether LLMs, guided by a tailored prompt, can generate

structured abstracts of systematic reviews and meta-analyses that comply with the PRISMA-A (Abstract) checklist. It matters because abstracts are the primary currency of literature searching — anyone using PubMed reads abstracts, and this paper quantifies how faithfully two flagship models can produce them from full texts, and where they fail (synthesis of results, risk of bias). It also demonstr

ates a reproducible prompt-engineering workflow that an independent researcher could adapt.

Methods in detail: A comprehensive search across PubMed (MEDLINE), Web of Science, and Scopus (24 January 2025) identified 162 full-text SRs/MAs published since January 2019 in Q1 orthodontic journals, selected only if they contained all components of PRISMA 2020 and the PRISMA-A 2021 checklist. Original abstracts were deliberately not used for comparison. Full texts were converted to Word documents with all metadata (including the original abstract) removed, but with tables preserved as editable objects and figures retained, exploiting the models' multimodal parsing. A structured, PRISMA-A-aligned prompt was developed via iterative chain-of-thought pilot rounds (10 random abstracts per round), with checklist-item omission rates quantified and two senior orthodontists reviewing outputs; in the final iteration both models produced abstracts "without any omissions." Each document was uploaded in a fresh chat session to ChatGPT-4o and Gemini Pro (10–11 February 2025), with a one-month stability re-run (10–11 March 2025). An 11-item Overall Quality Score (OQS, 0–2 per item, max 22; title excluded) was scored by two independent raters (third senior researcher for adjudication) against source full texts, requiring both structural presence and factual accuracy.

Key results and numbers: ChatGPT-4o achieved a mean OQS of 21.67 (98.5% of maximum) vs Gemini Pro's 21.00 (95.5%), a highly significant difference ($p < 0.001$). Note a minor internal inconsistency: the abstract reports SDs of 0.58 and 0.71, while the tables report 0.53 and 1.00. Item-level significant differences favoured ChatGPT-4o on Risk of Bias (1.98 vs 1.93, $p = 0.019$), Included Studies (1.96 vs 1.87, $p = 0.002$), Synthesis of Results (1.96 vs 1.82, $p < 0.001$), and Funding (2.00 vs 1.79, $p < 0.001$); section-level differences appeared in Methods (7.77 vs 7.62, $p = 0.007$), Results (3.93 vs 3.70, $p < 0.001$), and Other (4.00 vs 3.75, $p < 0.001$). Background items showed no difference ($p = 0.411$). Test–retest and inter-rater ICCs both ranged 0.92–1.00. Post-hoc power was ≈0.91 (Cohen's $d = 0.36$). Journal-level descriptive accuracies favoured ChatGPT-4o throughout (e.g., EJO 97% vs 91%; Progress in Orthodontics Synthesis of Results 91% vs 69%), though no inferential journal comparisons were performed. The lowest-scoring item for both models was "specify the methods used to present and synthesise results."

Conclusions and takeaways: Both LLMs, guided by the tailored prompt, can generate PRISMA-A-compliant abstracts, but ChatGPT-4o consistently outperformed Gemini Pro, plausibly due to stronger table-based reasoning (the authors cite its TabFact/WikiTableQuestions benchmark performance) — critical because SRs concentrate quantitative detail in tables. The 0.67-point overall gap "reflects substantial gains in the Methods and Results domains—areas with the greatest manual editing burden." Prompt specificity appears decisive: prior studies with generic prompts (CONSORT-A) reported lower quality. The models are framed as assistants: "these models are designed to augment—not replace—expert judgment," and all outputs require critical human review.

Limitations and caveats: The authors note that performance on purely narrative, text-only inputs may differ (tables/figures were intentionally preserved); readability, linguistic fluency, and clinical validity were not assessed; rapid LLM updates threaten reproducibility despite the one-month stability check; interpretive scoring variability remains possible despite high ICCs; and the orthodontics-only, Q1-journal sample limits generalizability. The two-model version pairing (GPT-4o vs Gemini 2.0-era Pro) is already dated. The extraction of this PDF also garbles

several table percentages (e.g., "94. %"), and the null-hypothesis framing ("equivalent OQS") sits awkwardly with the small, likely trivially meaningful 3% absolute difference.

Practical implications for a non-academic researcher using PubMed and AI: For anyone summarizing papers with ChatGPT or Gemini: preserve tables when feeding PDFs to the model, use checklist-aligned structured prompts rather than "summarize this," and expect the weakest performance precisely where it matters most — synthesis of results and risk-of-bias statements. ChatGPT-4o was more reliable for table-borne numbers. Crucially, the paper shows even near-ceiling checklist scores do not guarantee accuracy: scoring required checking every value against the source, which is exactly the verification habit a non-academic user must adopt.

Notable quotes:

- "The structured prompt developed in this study may assist researchers in generating PRISMA-A-compliant outputs more efficiently, helping to accelerate the completion and standardisation of high-level clinical evidence reporting."
- "These models are designed to augment—not replace—expert judgment: all AI-generated content must undergo critical human review."
- "A checklist item was scored as present only if it appeared in the abstract and its reported value matched the source full text."

Connections::. Its head-to-head ChatGPT-vs-Gemini design is the natural continuation of's five-tool comparison, though it tests task performance rather than output similarity.. Its finding that models struggle with synthesis/risk-of-bias reporting dovetails with's reviewer checklist items B6 and C2–C5 — exactly the items a human reviewer must still verify..'s caution about hallucination and human oversight is empirically supported here; ; Johnson (2022)'s Toolbox indexes PRISMA-A among its 112 guidance documents; ; Verma (2024) is a reminder that even curated abstract databases (like LERRN) rely on human selection where these models would automate.

Limitations and caveats: As a narrative review it is selective and citation-curated rather than exhaustive; its field moves so fast that 2024 content (GPT-4, Gemini 1.5, Claude 3) is now superseded. Many claims rest on preprints and industry sources; the authors themselves flag that AI-drug-discovery claims come largely from companies with commercial stakes. Clinical adoption of AI pathology is "hampered by the lack of digital workflows" in North America and Europe, and explainability methods "are still limited." No methods section means no appraisal of evidence quality per claim.

Practical implications for a non-academic researcher using PubMed and AI: This is the batch's most directly actionable paper for a non-specialist: start with zero-shot LLM use and prompt engineering; use in-context learning or RAG (feeding documents to the model) before considering fine-tuning; prefer local/open models for confidential data; and match tool complexity to task (GUI tools for routine image work; Python only when needed). Its warning that handcrafted hypotheses underperform data-driven learning with sufficient data — and the reverse with little data — is a transferable heuristic for judging any AI-assisted analysis, including literature mining.

Notable quotes:

- "Artificial intelligence (AI) has been commoditized. It has evolved from a specialty resource to a readily accessible tool for cancer researchers."
- "We postulate that any cancer researcher nowadays needs to acquire a certain level of AI literacy."

- "The most substantial contribution of LLMs to cancer research might not be the human-like text processing capabilities per se, but the complete upending of traditional machine learning workflows from rigid train–test–evaluate cycles to zero-shot applications."

Connections: Provides the theoretical substrate for Su (2025): LLMs in medical diagnostics: scoping review with bibliometric analysis (whose "prompt tuning/fine-tuning/pretrained from scratch" taxonomy mirrors this paper's framework) and for Demir (2025): AI-driven abstract generating under the PRISMA-A framework's prompt-engineering workflow (an applied instance of zero-shot/checklist-guided use). Its multimodal vision-LLM discussion (PathChat, LLaVA-Med) underpins Su's radiology/dermatology findings. Garry (2026): Exploring variation in research priorities generated by AI tools treats the LLM ecosystem whose architecture this paper explains. Ma (2026): AI in implementation research: applications and recommendations's NotebookLM podcast and RAG-style synthesis are concrete cases of its "AI for language" chapter. Its industry-bias caution aligns with Begagic (2025): The role of reviewers in the era of systematic reviews: a practical…'s integrity concerns; Johnson (2022) and Verma (2024) are, in its terms, the off-the-shelf human-curated "tool" layer that non-computational researchers can adopt without programming.

Begagic (2025): The role of reviewers in the era of systematic reviews: a practical... — Practical guide: the role of reviewers of SRs and meta-analyses, with a 23-item checklist (Biomol Biomed 2026;26(1):40–50)

What the paper is and why it matters: This narrative review distils PRISMA-2020, MOOSE, and PRIOR recommendations into a practical guide for peer reviewers of systematic reviews and meta-analyses (SRMAs), centred on a streamlined 23-item Yes/No/NA checklist across four sections (initial evaluation, methodology, results, emerging issues/ethics). It matters because SRMAs sit atop the evidence hierarchy yet are produced at scale (38,449 PubMed-indexed SRMA publications by 2022) amid paper mills and AI-generated content; this paper is the batch's integrity lens, showing what a competent appraisal of any evidence synthesis — including AI-assisted ones — must check.

Methods in detail: A narrative (non-systematic) review organized by the lifecycle of an SRMA review: initial evaluation (justification, PRISMA/MOOSE/PRIOR checklist compliance), methodology (prospective registration in PROSPERO/INPLASY/OSF with verification against the protocol; PICOS-framed eligibility; reproducible search strategy — "searching at least two reputable databases is a minimal requirement," with Embase + MEDLINE/PubMed + Google Scholar or Web of Science as a "foundational trio"; PRISMA flow diagram scrutiny; dual data extraction; RoB 2/ROBINS-I; pre-specified fixed- vs random-effects and subgroup plans), results (appropriate effect measures; I^2 interpretation with 0–25%/~50%/>75% conventions; Cochran's Q caveats; forest-plot inspection; publication-bias assessment via funnel plots/Egger's test for ≥10 studies; trim-and-fill; sensitivity analyses; GRADE certainty rating), and emerging issues (paper mills, misconduct, AI misuse). Recommendations are synthesized from the named frameworks plus COPE and ICMJE guidance, with case examples (Scottish Medical Journal retractions, Wiley mass retractions).

Key results and numbers: The centrepiece is Table 1, a 23-item checklist: A1–A2 (rationale; attached checklist), B1–B7 (registration number; PICOS criteria; search reproducibility; flow diagram; dual extraction; RoB tool; pre-specified analysis), C1–C7 (effect measures; I^2; subgroup/meta-regression justification; forest plots; publication bias; sensitivity analyses; GRADE), D1–

D7 (author transparency; AI-use disclosure; paper-mill screening; plagiarism; verification of "hallucinated" references; conflicts of interest; journal-policy compliance). Contextual figures: SRMAs grew from the first SR in PubMed/MEDLINE in 1957 to 38,449 publications by 2022; a 2022 COPE/STM report estimated 2%–46% of submissions to certain journals (2019–2021) traced to paper mills; Wiley retracted 511 papers in one announcement with ~1,200 more under review; a JAMA Network Open study found 299 of 200,000 life-science SRMAs cited retracted paper-mill articles (0.15% contamination, rising over time, with ~one-third of citations occurring post-retraction); conference abstracts showed roughly a two-fold increase in AI-content likelihood between 2021 and 2023.

Conclusions and takeaways: Reviewers are the last line of defence: they must verify registration, search adequacy, heterogeneity handling, and evidence certainty, while also screening for paper-mill signatures (clusters of improbably positive results from one region/author cluster; identical means and SDs; fabricated references), undisclosed AI use, and overstated conclusions. AI is "a dual-edged sword": legitimate for screening and plain-language drafting when disclosed and human-verified, unethical "when AI is used to perform the authors' critical thinking." Suspected misconduct should go confidentially to editors, never as direct accusation. The GAMER checklist is endorsed for AI transparency.

Limitations and caveats: As a narrative review it cites illustrative rather than exhaustively searched evidence, and several supporting studies are themselves narrative or single-country reports. The checklist's 23 items compress complex judgments (e.g., I^2 thresholds "are not absolute") into binary answers, risking false reassurance when everything is marked "Yes." It is written from a clinical-medicine perspective; the statistics guidance (e.g., Egger's test) is simplified for non-statisticians. The PDF extraction shows mild layout artefacts (columns interleaved, e.g., "e2,000" for €2,000), and the I^2 thresholds attributed generally are convention-dependent.

Practical implications for a non-academic researcher using PubMed and AI: The checklist doubles as a consumer's appraisal tool: before trusting any systematic review found on PubMed, run the A/B items yourself (Is it registered in PROSPERO? Is the search strategy reproducible? ≥2 databases? Flow diagram present?). Its AI chapter is a practical scam-detector: verify a few cited references actually exist and support the claim (fabricated references are a hallmark of AI-generated or paper-mill text); use Retraction Watch and simple title searches; distrust "too good to be true" result consistency. For one's own AI-assisted writing, mirror item D2: disclose the tool and verify everything.

Notable quotes:

- "A meta-analysis is only as reliable as the studies it includes; if paper mill products… enter the evidence pool, the meta-analysis may be compromised."
- "Reviewers should act almost like a co-pilot, verifying every instrument reading."
- "Reviewer's vigilance can prevent these issues from polluting the scientific literature."

Connections:: Supplies the appraisal standards that measures against — its item B6/C7 (RoB, GRADE) are precisely where Demir's LLM abstracts scored worst..'s bias/hallucination findings give empirical content to Begagic's item D5 (verify hallucinated references).. Johnson (2022)'s Toolbox indexes the very frameworks (PRISMA, quality-assessment guidance, N = 70 documents) that Begagic operationalizes — the two papers are complementary catalogue-and-checklist resources..'s AI-disclosure statement and transparency concerns instantiate Begagic's D2/item D7 expectations.. ; Verma (2024) offers the librarian-curated model of quality control that peer review here struggles to scale.

Johnson (2022) — The Systematic Review Toolbox 2022 relaunch (Syst Rev 2022; commentary)

What the paper is and why it matters: This commentary describes the 2022 relaunch of the Systematic Review (SR) Toolbox — a free, open, searchable web catalogue of software tools and guidance for evidence synthesis, originally built in 2014 by Christopher Marshall. It matters because it is the field's canonical meta-resource: it both documents how the evidence-synthesis landscape expanded (citing taxonomies of 9, 10, 22, and finally 48 review types) and empirically maps which review families and production stages are well- or under-served by existing tools. For a synthesis on literature-searching methodology, it is the reference infrastructure behind tool discovery.

Methods in detail: In February 2022 the team conducted a mapping exercise of all records in the SR Toolbox (n = 352) using an Excel coding framework adapted from Sutton et al.'s (2019) review-family taxonomy (48 types in seven families), with domains for tool type (guidance vs software), review family, review stage (protocol, search, screening, extraction, quality assessment, synthesis, reporting, reference management, stakeholder engagement), and cost. Traditional and purpose-specific reviews were excluded from coding scope (insufficiently systematic; too broad), though scoping and mapping reviews were separated out from the purpose-specific family. A single reviewer (EEJ) coded all records and checked hyperlink validity (5 tools with dead links were excluded); a second reviewer (HOK) checked "a small percentage" of records; data were imported into a relational Microsoft Access database, exported as CSVs, rebuilt online in MySQL via phpMyAdmin, and fronted by a rebuilt PHP search interface allowing filtering by review family and stage. A basic descriptive analysis tabulated tool/guidance counts per family and stage.

Key results and numbers: The relaunched toolbox went live on 13 May 2022 with 235 software tools and 112 guidance documents. Most tools and guidance (223 and 78 respectively) serve systematic reviews; reviews of reviews are comparatively underserved (66 tools, 28.1%; 22 guidance, 19.6%); qualitative reviews had 108 tools (46%) but only 19 guidance documents (17%) — "the most under-served review family in terms of guidance." By stage, guidance is dominated by quality assessment (70; 62.5%) then reporting guidelines (26; 23.2%); software clusters on searching (84) and synthesis (82); "there appears to be a paucity of software tools (N = 2; 0.9%) and guidance (N = 3; 2.7%) that relates to stakeholder engagement." 215 tools (91.5%) and 61 guidance documents (54.5%) apply to multiple families; 75 tools (32%) span multiple stages. 181/235 tools (77%) are free; 96/112 guidance documents (85.7%) open access. Popularity evidence: in January 2022 the site logged ~28,500 hits, 6,100 visits, ~4,500 unique visitors. Currency: 71 tools (30.2%) had no identifiable update date; most updated tools were refreshed within the previous four years (N = 115, peaking at 51 in 2021); 63.9% of guidance postdates 2015 (earliest 1998).

Conclusions and takeaways: The updated interface makes the Toolbox navigable for users unfamiliar with underlying computational concepts (e.g., machine learning), responding to adoption barriers. The gap analysis identifies future research needs: tools/guidance for reviews of reviews and qualitative reviews; refinement of the "Other" category (network meta-analyses, prognostic reviews); possible inclusion of living systematic reviews as a category (currently absent despite ML supporting their production); and stakeholder engagement as the clearest unmet need. The trajectory of guidance publications has risen steadily, "particularly high since 2015."

**Limitations and ca

veats:** The mapping was performed by a single reviewer with only partial second-checking — the authors concede "it is possible that there may [be] some minor inaccuracies in coding." The cost and update-date fields were often unverifiable (30.2% of tools undated), so currency claims are uncertain. The analysis is purely descriptive (counts), with no assessment of tool quality or effectiveness. In the pdftotext extraction, Tables 2 and 3 are garbled — the row values appear detached from their labels — so family- and stage-level numbers must be read from the prose. Scope decisions (excluding traditional and purpose-specific reviews) affect comparability with other taxonomies.

Practical implications for a non-academic researcher using PubMed and AI: The Toolbox (systematicreviewtools.com) is the single best free entry point for an independent researcher choosing screening, search, or synthesis software: filter by review stage (e.g., "screening" or "searching") and cost ("completely free") without needing to understand machine learning. Its finding that 77% of tools are free and most guidance is open access means a robust workflow is achievable at zero cost. The gap analysis is also a warning: for qualitative synthesis or stakeholder engagement, expect to rely on generic guidance rather than purpose-built tools — and on AI tools, which postdate this catalogue.

Notable quotes:

- "The SR Toolbox provides a platform for those undertaking evidence syntheses to locate guidance and software tools to support different aspects of the review process across multiple review types."
- "A potential barrier to adoption includes inexperience of some of the underlying principles of tools, such as machine learning."
- "There appears to be a paucity of tools and guidance relating to stakeholder engagement within the review process."

Connections:: Is the infrastructure layer beneath the whole batch::'s Rayyan screening,'s Covidence workflow, and's PRISMA-A/PRISMA-2020 checklists are exactly the categories of tools and guidance the Toolbox indexes (and its analysis shows quality-assessment guidance is the best-covered area, matching Demir's scoring framework)… Its note that ML already supports living systematic reviews foreshadows the AI-in-synthesis papers („); ;'s checklist items correspond to the Toolbox's best-stocked guidance category (quality assessment); ; Verma (2024) is its closest structural analogue — a single-maintainer, curated tool/resource catalogue, but for librarian resource reviews rather than review methods.

Verma (2024) — LERRN: free citation database for resource reviews (JMLA 2024;112(2), Virtual Project)

What the paper is and why it matters: This short "Virtual Project" piece by Louisa Verma (Portland Cement Association library manager) describes LERRN — the Librarians' Electronic Resource Reviews Network — a free, self-maintained citation database collocating librarian-written reviews of electronic resources (databases, platforms, tools). It matters as a proof-of-concept that a single librarian can build a functioning, niche evidence-curation service from free components (Zotero, RSS feeds, a low-cost ILS), addressing a real gap: resource reviews are scattered across journals that small or solo-staffed libraries cannot afford to subscribe to. It is the batch's smallest paper but a telling model of low-cost scholarly infrastructure.

Methods in detail: The method is infrastructural rather than empirical. A journal list was compiled from known titles, library-association publications, the Directory of Open Access Journals (DOAJ), and the JournalTOCs (UK) website for LIS-related journals. Citations and

abstracts from journals publishing content relevant to electronic resources are harvested into Zotero via imported RSS feeds. The author reviews these feeds every other month, selecting articles of value for "electronic resource purchasing, training, marketing, or use," and tags each with four tag types (year, resource name, topic, article type) to support browsing by Tags, Categories, Series, and Authors. Selected records are imported into Librarika, a low-cost cloud-based integrated library system (ILS), using a vendor-supplied CSV template. The scope covers electronic resources in science, technology, engineering, mathematics, and medicine; article types include reviews, overviews, comparisons, projects, search tips, book reviews, and general electronic-resource-management (ERM) pieces. Updates occur six times a year (January, March, May, July, September, November), announced via a What's New blog on a Google Sites website that also lists the indexed journals.

Key results and numbers: The database was implemented in 2022 and, at the time of writing (accepted February 2024), "includes 920 articles (a majority published from 2019 to the present)" and has recorded "over 8,000 page views." Citations reach back to 2019, chosen because most resources lack older reviews worth indexing. The service is sustained by one person: "The LERRN database and website was created and is currently maintained, updated, and subsidized solely by the author." No usage analytics beyond page views, no user evaluation, and no formal quality assessment of selected articles are reported — the numbers given are throughput and reach, not impact measures.

Conclusions and takeaways: LERRN demonstrates that collocating scattered resource reviews into one browsable, free database is feasible with commodity tools and a modest ongoing time commitment (RSS monitoring + tagging, bimonthly). Its stated trajectory is expansion: adding social-sciences and humanities coverage and extending the backfile pre-2019 for resources without recent reviews. The author's expectation is that, as content grows, "it will become a useful tool for librarians" — a modest but realistic claim, with the underlying lesson that curation value comes from selection criteria and tagging consistency rather than sophisticated technology.

Limitations and caveats: The author identifies the main one: Librarika is designed as an online catalogue, "intended as an online catalog and, as such, it provides less flexibility for tailoring searching, browsing, or highlighting content," limiting the discovery experience even though it handles importing, browsing, and full-text linking. Other unstated limitations are evident: single-author curation introduces individual selection bias and creates a key-person sustainability risk; inclusion criteria ("value in assisting with… purchasing, training, marketing, or use") are subjective; coverage is restricted to STEM/medicine and to journals the author could identify; there is no evaluation of whether users actually find it useful beyond page-view counts; and the paper predates any assessment of AI-based alternatives to manual RSS screening.

Practical implications for a non-academic researcher using PubMed and AI: LERRN is a template for a DIY current-awareness system: subscribe to journal RSS feeds (or journal TOC alerts), triage them in a free reference manager (Zotero), apply a consistent tagging scheme, and publish the result through a cheap catalogue or static site. For a non-academic researcher tracking an evidence base, the same pipeline — substituting PubMed's own RSS/API or saved-search alerts for journal feeds — yields a personal, transparent, AI-free (or AI-assisted) literature database. Its bimonthly update rhythm shows realistic maintenance effort, and its single-volunteer fragility argues for documenting criteria so the system can be handed over or audited.

Notable quotes:

- "Electronic resource reviews written by librarians are a valuable way to identify potential content platforms and stay current on new resources."

- "Small or solo-staffed libraries may not subscribe to library and information sciences databases or journals that contain reviews of electronic resources."
- "The LERRN database and website was created and is currently maintained, updated, and subsidized solely by the author."

Connections:: Is the batch's micro-scale counterpart to Johnson (2022):: both are curated tool/resource catalogues built by small teams (one person vs four editors), and Johnson's Toolbox is the natural place a resource-review service like LERRN would be indexed… Its librarian-mediated search development parallels's credited health-sciences librarian and reflects the information-specialist culture underlying's search-strategy checklist item (B3) and's five-database search design; ; and supply the AI layer that could either support or undermine such hand-curated services.

14.6 The librarian craft — roles, competencies, and training

Lizarondo (2025): Data extraction and transformation in mixed methods systematic reviews — JBI guidance on data extraction and qualitization in convergent integrated mixed-methods systematic reviews (JBI Evid Synth 2025;23(3):429–440)

What the paper is and why it matters: This is an official methodology guidance paper from the JBI Mixed Methods Systematic Review (MMSR) Methodology Group, authored by an international panel of 13 evidence-synthesis experts (co-first authors Lizarondo and Stern, JBI Adelaide, plus JBI centres of excellence in the US, UK, Canada, Portugal, and Brazil). It addresses a poorly understood step in the JBI "convergent integrated" approach to mixed-methods reviews: converting quantitative data into qualitative form ("qualitizing") so that quantitative and qualitative evidence can be merged into integrated findings. The paper matters because it clarifies the boundary between data extraction and data transformation, a distinction that JBI's 2020 methodological guidance left vague and that reviewers frequently confused in practice. It claims to be "the most comprehensive guidance currently available" on this process, and it was formally ratified by the JBI International Scientific Committee in June 2024, giving it normative weight for anyone conducting JBI mixed-methods reviews.

Methods in detail: The group used a multipronged, consensus-driven development process. First, a structured literature search was run in PubMed, CINAHL (EBSCOhost), and Google Scholar in May 2023, using broad index headings and title/abstract keywords (mixed methods, data transformation, qualitization, evidence synthesis, systematic review), with no limits on date, language, or publication type, yielding 208 papers mentioning data transformation or qualitization. Second, to understand real-world practice, two members (LL and CS) used Google Scholar to identify the first 30 papers citing the JBI MMSR methodology, ranked by Google's relevance algorithm rather than recency; non-systematic reviews and non-convergent-integrated reviews were excluded, leaving 11 reviews for analysis. For these, the group extracted qualitization methods, retrieved the underlying primary studies to verify how qualitization actually occurred, and compared approaches to identify themes and gaps. Third, a narrative summary informed draft guidance, which was circulated to the wider methodology group and refined through online meetings and email between July and October 2023. Fourth, external validation took place at the JBI iGNITE international conference in November 2023. Finally, the guidance went to the JBI International Scientific Committee in January 2024, was presented for debate on March 19, 2024, and was ratified on June 25, 2024 after minor revisions.

Key results and numbers: The literature search returned 208 papers, but most concerned primary mixed-methods studies, not systematic reviews, and none offered comprehensive procedural guidance for reviews. Among the 11 examined reviews that claimed the JBI convergent integrated approach, three distinct transformation approaches emerged: Approach 1 (1 review) extracted verbatim numeric results plus the primary authors' narrative and treated these as already-qualitized statements; Approach 2 (3 reviews) extracted non-verbatim data, tweaking or rephrasing for clarity/context; Approach 3 (7 reviews, the most prevalent) extracted first and then qualitized as a separate step, replacing numbers with text descriptors (e.g., "majority," "more than half," "significantly higher") or assigning descriptive labels. Five reviews were chosen as worked examples (risk stratification in cancer screening, physical activity after coronary bypass, COVID-19 and eating disorders, pediatric anaphylaxis information needs, midwife-led care in LMICs), each with extracted versus qualitized data shown in Table 1. The guidance recommends: verbatim extraction wherever possible; extraction of ALL relevant factors regardless of statistical significance or percentage cutoffs; constructing a narrative only when primary studies present results solely in tables/graphs; cross-checking narrative against numerical results, with discrepancies handled via supplementary data or author contact and, failing that, prioritizing numerical data while noting the discrepancy; double extraction or validation by someone experienced in quantitative research; and a published extraction table distinguishing verbatim from modified data. Qualitization is formally redefined as "the integrative process in which extracted quantitative and qualitative data from primary studies are analyzed for identifiable patterns and systematically grouped into descriptive categories developed from mixed evidence sources," occurring after extraction as part of integration, not during it.

Conclusions and takeaways: The core conceptual move is separating extraction (faithful capture of primary data) from qualitization (an integrative grouping exercise). Reviewers should use primary authors' own narrative representations rather than reinterpreting numbers, so extraction yields "integration-ready datasets." Keeping the numeric values alongside narrative text in integrated findings prevents loss of granularity and lets readers trace integrated statements back to the underlying statistics. The authors acknowledge inherent variability in MMSRs, that the approach may need tailoring, and that saturation of observed approaches was not certain; future work will address assessing certainty/confidence in MMSR findings.

Limitations and caveats: The sample of 11 reviews was drawn from the first 30 citing papers by Google Scholar relevance ranking — a convenience sample favoring highly cited, possibly unrepresentative works. The authors themselves report uncertainty about whether thematic saturation was reached, so the three approaches may not cover the full spectrum of practice. The guidance is normative (expert consensus) rather than empirically validated; no evidence is presented that the recommended approach improves review quality or reproducibility. Scenarios for constructing narratives are hypothetical (Table 2), and the advice to prioritize numbers over narratives when they conflict rests on an untested assumption that numbers are less biased. As a pdftotext extraction, Table 1′s layout is partly scrambled but readable; no serious artefacts elsewhere.

Practical implications for a non-academic researcher using PubMed and AI: For anyone synthesizing mixed evidence (e.g., survey percentages plus interview quotes), the rules generalize: extract verbatim, keep numbers attached to the words that interpret them, don't drop statistics when writing thematic summaries, and document every modification. An AI assistant can help draft narrative sentences from numeric tables (the reviewer-constructed narratives the paper recommends) but must be checked against the source numbers — exactly the discrepancy-checking discipline the paper mandates. AI-generated "descriptors" like "majority" from

percentages should be traceable, and a final extraction table flagging verbatim vs. AI-modified entries mirrors the paper's transparency requirement.

Notable quotes:

- "Data extraction in MMSRs should produce integration-ready datasets, not fragments requiring post-processing."
- "The integrative process in which extracted quantitative and qualitative data from primary studies are analyzed for identifiable patterns and systematically grouped into descriptive categories developed from mixed evidence sources."
- "This omission poses a risk of losing vital information during the process of data integration."

Connections: This is the methodology-guidance anchor of the batch: the JBI manual it expands is the same standards base that Townsend (2017): A competency framework for librarians involved in systematic reviews and Farris (2024): Curriculum framework for systematic review competencies (librarians) train librarians against. Its "collecting and combining data" concerns parallel the "C" phase of Parker (2022): Literature reviews: tips from knowledge synthesis librarians's PIECES framework and the data-management competency in Townsend. Its insistence on extracting all relevant data regardless of arbitrary statistical cutoffs is a useful corrective to the filter-and-hedge precision mindset implicit in Westrick (2021): Medical librarians' practices in locating clinical trials's search-filter survey. It has no direct link to the teaching papers (Minuti (2018): Teaching searching skills with a flipped classroom, Boden (2018): Learning while doing: MLA Systematic Review Project evaluation) but shares with them the theme that well-defined procedures must be taught explicitly rather than absorbed implicitly.

Spencer (2018): Roles for librarians in systematic reviews: a scoping review — Scoping review identifying 18 roles for librarians in systematic reviews (JMLA 2018;106(1):46–56)

What the paper is and why it matters: Spencer (St. Luke's Hospital) and Eldredge (University of New Mexico) conducted a scoping review asking: what roles do librarians and information professionals actually play in systematic reviews? Before this paper, the literature consisted mainly of case studies and anecdotal accounts; no study had attempted a comprehensive inventory. The paper matters because it reframed librarians from "expert searchers" to multifunctional systematic review team members, giving both librarians (a checklist of services to offer) and researchers (a menu of what to ask for) a shared vocabulary. It is approved for MLA's Independent Reading Program and became one of the most-cited reference points in the librarian-SR literature — it is explicitly cited by Farris (2024): Curriculum framework for systematic review competencies (librarians) and Parker (2022): Literature reviews: tips from knowledge synthesis librarians in this batch.

Methods in detail: A scoping review (defined by the authors as a broad, exploratory reconnaissance to map key characteristics), with searches run in February 2017. Multiple strategies were combined: (1) four PubMed searches — journal-title-restricted searches (JMLA, Bulletin of the MLA, Medical Reference Services Quarterly, Health Information & Libraries Journal) combined with systematic-review text words and PubMed filters (sysrev_methods[sb], systematic[sb]); a MeSH-based search ("Information Services," "Information Storage and Retrieval" AND librarian*/informationist terms) NOTed against the first; a "Librarians" MeSH search; and a combined "Review Literature as Topic" AND "Librarians" MeSH search; (2) two searches each in LISTA and CINAHL combining systematic review keywords with librarian/information-specialist keywords; (3) hand-searching of poorly indexed library journals (JCHLA website,

JEAHIL via PDF text search, Hypothesis, Evidence Based Library and Information Practice); (4) text-word searching of MLA annual meeting poster and paper abstracts for 2002–2016; and (5) indexes of two health sciences librarianship textbooks edited by M. Sandra Wood. Inclusion was deliberately broad: any role a librarian might perform in an SR, guided by two prior case studies (Beverley 2003; Harris 2005). Exclusions: book reviews, CE advertisements, SR-update articles, article reviews. Mendeley was used for citation management, with results in separate inclusion/exclusion folders and a Word document for deduplication screening.

Key results and numbers: The searches yielded 310 relevant articles, book chapters, presented papers, and posters, from which 18 distinct roles were identified and summarized alphabetically with a supporting bibliography. Well-documented roles included searching (the dominant role, with subcategories: general, databases/resources, grey literature, protocol development, search strategies, subject-specific searches), source selection, evaluation of search strategies, teaching, citation management, and de-duplication. Less-documented roles included planning (setting expectations at project outset), question formulation (traced to the reference interview; supported by Eldredge's RCT showing librarian training improved public health practitioners' question articulation), peer review of search strategies (per IOM recommendation and Sampson's PRESS guideline), reporting and documentation (with Yoshii's finding that none of 56 Cochrane reviews contained all seven required search-reporting elements), formalized SR services, impact/outcomes research (e.g., Koffel's survey showing librarian partnership increased search quality, and Rethlefsen's work linking librarian co-authorship to higher-quality search reporting), indexing of database terms, search filters/hedges, research agenda setting, systematic reviews in librarianship, collaboration, and technological/analytical tools (capture–mark–recapture, Bradford's Law). Table 1 of the article (major communication channels reporting on librarian roles) lost its numeric counts in the pdftotext extraction, though channel names (JMLA, HILJ, JCHLA, MRSQ, JEAHIL, EBLIP, J Clin Epidemiol, Systematic Reviews, MLA News, Hypothesis, posters, papers, book chapters) are visible.

Conclusions and takeaways: Librarians play central roles on systematic review teams "including roles that go beyond searching." Some roles named in early case studies (data abstraction, bias assessment, critical appraisal, data synthesis, document supply, report writing) were not further described elsewhere and were deliberately excluded

from the 18-role list — an honesty about the evidence base. Grey literature mattered: many roles were documented only in conference posters and papers, reinforcing the argument that professional knowledgebases require grey-literature searching. The authors urge librarians performing undocumented roles to publish their experiences, and hospital librarians with limited time to use the inventory as a menu of feasible partial services.

Limitations and caveats: Scoping reviews by design don't appraise study quality or synthesize effect sizes, so the inventory describes what is reported, not what is effective. The authors acknowledge likely missed references where librarians weren't acknowledged, roles weren't documented, or indexing eluded the search strategies. Roles overlap, and many sources cover multiple roles, so the 18-role taxonomy is an organizing heuristic rather than a validated instrument. No frequency data (how often each role occurs) is systematically presented in the extractable text. Minor extraction artefacts: Table 1′s counts are missing, and the reference list ordering is scrambled in places.

Practical implications for a non-academic researcher using PubMed and AI: The 18-role list doubles as a self-assessment checklist for a solo, non-academic researcher acting as their own "information specialist": question formulation, source selection, de-duplication, search-strategy evaluation, peer review (e.g., via PRESS), reporting/documentation, and citation

management are all roles an individual can perform deliberately with AI support. It also warns what AI cannot easily replace — knowledge of indexing quality, filter validation, and grey-literature channels (which the paper shows are where much practitioner knowledge hides).

Notable quotes:

- "Librarians play central roles in systematic review teams, including roles that go beyond searching."
- "He found that working with a librarian increased the quality of searches, particularly when the librarian was knowledgeable about systematic reviews via either training or past experience."
- "Thus, the presence of many meeting papers and posters reinforced the frequent observation that one needs to consult the grey literature when accessing a professional knowledgebase."

Connections: This paper is the hub of the batch's librarian literature: Farris (2024): Curriculum framework for systematic review competencies (librarians) explicitly instructs curriculum builders to review "the roles… described by Spencer and Eldredge"; Parker (2022): Literature reviews: tips from knowledge synthesis librarians cites it as justification for involving an information specialist; Townsend (2017): A competency framework for librarians involved in systematic reviews operationalizes its roles into competencies; Boden (2018): Learning while doing: MLA Systematic Review Project evaluation shares co-author Eldredge and documents librarians learning to fill these roles experientially; Westrick (2021): Medical librarians' practices in locating clinical trials surveys practitioners inside its largest role ("searching"); and Minuti (2018): Teaching searching skills with a flipped classroom exemplifies its "teaching" role aimed at future physicians.

Townsend (2017): A competency framework for librarians involved in systematic reviews — Competency framework for librarians involved in systematic reviews (JMLA 2017;105(3):268–275)

What the paper is and why it matters: A commentary from seven University of Michigan Taubman Health Sciences Library informationists (Townsend, Anderson, Ginier, MacEachern, Saylor, Shipman, Smith) presenting the first published competency framework specifically for librarian involvement in systematic reviews. It matters because it converts the informal expertise implied by role inventories into a structured, assessable standard: six competencies with associated skills and knowledge "indicators," each assessable at three levels of an adapted Miller's Pyramid. It was written to build local capacity amid rising demand, but was deliberately designed for external reuse — and indeed became the foundation for the MLA Systematic Review Services Specialization's skill list, as Farris (2024): Curriculum framework for systematic review competencies (librarians) documents in this batch.

Methods in detail: The team convened in November 2015 with the purpose of identifying and defining core competencies for informationists/librarians involved in SRs. Membership was purposeful: informationists at varying experience levels, all integrated into SR work via instruction, consultation, or team membership; five taught or assessed the "Systematic Reviews: Opportunities for Librarians" flipped-classroom workshop, one led the Michigan Medicine Clinical Care Guideline Development Search team, and three trained doctoral nursing students. The process had four steps: (1) each member independently generated lists of skills and knowledge items core to successful SR integration; (2) members thoroughly reviewed methodology and reporting standards — Cochrane Handbook, JBI Reviewer's Manual, PRISMA, National Academies (IOM) Standards, Campbell Collaboration, and CRD guidance — to add skills and ensure consistency;

(3) results were collated and indicators grouped by theme into overarching competency areas, combining skills where appropriate; (4) subgroups defined each competency's scope and refined its indicators. An initial plan to stratify indicators by expertise level ("novice/intermediate/expert") was abandoned because librarian roles on SR teams resist such stratification; instead the team adapted Miller's Pyramid for Clinical Assessment (original levels: Knows, Knows How — cognitive; Shows, Does — behavioral) into three applied levels: Knows, Shows, Does. No external validation, Delphi process, or user testing of the framework is reported — it is a consensus product of one team.

Key results and numbers: Six competencies were identified: (1) "Systematic review foundations" — how/why SRs are used in health sciences, finding SRs, methodological distinctions between SRs and other review types; (2) "Process management and communication" — team communication, project management, negotiating the librarian's role (searcher vs. consultant), authorship/acknowledgment criteria, typical timelines and team composition, tools, and keeping collaborators apprised of search progress; (3) "Research methodology" — matching question to review type, SR methodology and best practices, protocol development (question refinement, inclusion/exclusion criteria, data management plan, registration), critical appraisal, study selection with inter-rater reliability and risk of bias; (4) "Comprehensive searching" — preliminary searches, database selection including grey literature, keyword/controlled-vocabulary term generation, search filters and validation (sentinel articles), advanced multi-database strategies, saved searches/alerts, and transparent replicable documentation; (5) "Data management" — citation managers (EndNote, RefWorks, Mendeley), SR software (Distiller, Covidence, Rayyan), extraction tools/forms, archiving and PRISMA data collection; (6) "Reporting" — reporting standards, transforming final strategies into replicable publication formats, communicating the search per PRISMA. The paper gives no participation numbers; its "results" are the framework table itself (whose grid layout is garbled in the pdftotext extraction but reconstructable). Suggested assessments: PRESS form for "Does"-level comprehensive searching; PRISMA comparison for Reporting; observed consultations for Shows; coursework/quizzes for Knows.

Conclusions and takeaways: The framework is intentionally flexible: it sets no required levels per indicator, because librarians' appropriate competence varies with role, environment, and institutional priorities. Its intended uses are self/institutional assessment, gap identification, targeted professional development, workforce capacity building for SR services, and administrator planning. The authors note the MLA Systematic Review SIG had begun independently developing competencies in June 2016 and anticipate convergence — inviting continued discussion and evolution as librarians take on new SR roles and as the framework is adapted to other review types requiring systematic searching.

Limitations and caveats: The framework derives from a single institution's consensus (seven informationists at one university) with a literature-standards cross-check but no external validation, no Delphi survey, no inter-rater testing, and no field evaluation. The adapted Miller's Pyramid levels are asserted, not psychometrically established; the suggested assessment methods (observed consultations, PRESS forms) are illustrative only. By 2026 the named software (Distiller, Covidence, Rayyan) and guidelines (PRESS 2015, PRISMA 2009) are dated. Extraction note: Table 1′s three-column pyramid grid is badly garbled in the text file, requiring careful reconstruction.

Practical implications for a non-academic researcher using PubMed and AI: The six competencies are a ready-made self-training syllabus for a solo researcher doing systematic-style searching: foundations (know what an SR is), searching (controlled vocabulary, filters, sentinel-article validation), data management (deduplication, screening tools, archiving), and reporting

(PRISMA-S-style documentation). AI can help you "Show" and begin to "Do" — e.g., drafting search strings, PRESS-style critique of your own strategy — but the framework's core insight stands: competence must eventually be demonstrated against standards, not merely professed to a chatbot.

Notable quotes:

- "Competence can be measured using an adaptation of Miller's Pyramid for Clinical Assessment, either through self-assessment or identification of formal assessment instruments."
- "Expert searching, including comprehensive and replicable searches, remains a core way that librarians demonstrate their value."
- "In many instances, researchers want to conduct their own searches but need significant assistance with resources to consult as well as search strategies to employ."

Connections: Directly operationalizes Spencer (2018): Roles for librarians in systematic reviews: a scoping review's role inventory into trainable, assessable competencies; explicitly anticipated the MLA SIG process that later produced the SRSS curriculum implemented by Farris (2024): Curriculum framework for systematic review competencies (librarians) (which cites Townsend as its objectives source). Westrick (2021): Medical librarians' practices in locating clinical trials's finding that even expert librarians lack study-design-methodology knowledge maps precisely onto the framework's "Research methodology" competency gap. Boden (2018): Learning while doing: MLA Systematic Review Project evaluation's experiential learners were implicitly working toward the "Does" level without the framework's structure; Minuti (2018): Teaching searching skills with a flipped classroom and Farris (2024): Curriculum framework for systematic review competencies (librarians) show the "Knows/Shows" levels being taught in flipped and modular formats respectively.

Farris (2024): Curriculum framework for systematic review competencies (librarians) — Case report: building the "Essential Searching Skills" curriculum for MLA's Systematic Review Services Specialization (JMLA 2024)

What the paper is and why it matters: A case report by three librarians (Farris at MD Anderson Cancer Center, Lebo at University of South Dakota, Price at NIH Library) describing the design, delivery, and iterative refinement of "Essential Searching Skills for Librarians on Systematic Review Teams," a required Level I core course in the Medical Library Association's Systematic Review Services Specialization (SRSS), launched in 2022. It matters because it is a documented, working example of converting the 2017 competency framework (Townsend (2017): A competency framework for librarians involved in systematic reviews in this batch) into a delivered curriculum across eleven cohorts, and because it states the professional premise of the whole batch: librarian co-authorship correlates with higher-quality reported search strategies, and methodological guidance (Cochrane, JBI, Campbell, National Academies) recommends consulting an information retrieval expert.

Methods in detail: The MLA SR Workgroup defined course objectives aligned to Townsend et al.'s six competencies; the SRSS itself comprises 26 fundamental skills across two Learning Pathways (Level I, Level II), satisfied through required and elective courses — specialization candidates take eight required core courses and four electives. The authors' course was structured into four modules mapped to objectives: Module 1 (consultation and pre-searching: preliminary searches, finding existing/in-progress reviews, keyword and controlled-vocabulary term harvest-

ing from non-systematic articles, the reference interview, restrictions and bias debates, database selection, grey literature, hand-searching, preprints, drug/chemical searching); Module 2 (the search: question frameworks including PICO, concept nesting, controlled vocabulary with MeSH and Emtree examples, syntax, filters/hedges, live PubMed demonstration, and a breakout-group activity building one of three search concepts); Module 3 (search peer review using the PRESS 2015 guideline, PRESSforum, and the CADTH PRESS assessment form); Module 4 (search reporting using the PRISMA-S extension's 16 items in four parts). Delivery: a four-hour in-person pilot at MLA's New Orleans meeting (May 2022), then online as two 2.5-hour sessions (October 2022; seven sessions across 2023; March 2024; in-person again May 2024 in Portland), roughly thirty registrants per cohort, with Slack for communication, an OSF open course site (readings, handouts, slides, activity instructions), five-question identical pre-/post-tests, and MLA feedback forms. Iteration: concept nesting was re-explained after an early cohort's confusion, and small-group time was lengthened after positive feedback.

Key results and numbers: Eleven cohorts were delivered over 2022–2024, each with approximately thirty registered participants. Table 2 (pre/post-test analysis, garbled in extraction) reports a confidence question ("How confident are you at building a comprehensive literature search?", 1–5 scale) rising from a mean of 3.32 pre-test to 3.96 post-test (+0.64, mapped to SRSS skills 8–14). The remaining four knowledge items show positive "difference" values — Cochrane Handbook's three recommended resources (+10), appropriate drug-resource selection (+4), PRESS statement elements (+8), PRISMA-S major categories (+16) — though the extraction loses the units and raw column values, so these should be read as improvements with the exact denominators uncertain. Feedback themes were consistently positive: instructors curating content for both new and experienced searchers, enthusiasm for breakout rooms and the activity, and applicability to professional work. Critiques: course length perceptions varied both ways, desire for individual search feedback, and frustration with the number of communication tools. An emerging theme across all cohorts was realizing "there are no absolute answers in systematic searching." No formal statistical analysis was performed.

Conclusions and takeaways: The instructors' guiding principle was "Does this add value?" They emphasize searcher autonomy in how searches are built and documented, provided the resulting strategies are sensitive and reproducible. Librarians-teaching-librarians is framed as pivotal for peer knowledge sharing and service development. The case report is offered as a transferable framework: would-be instructors should review Townsend's competencies and Spencer and Eldredge's roles when designing curricula. The concluding claim is normative and strong: developing proficient searching skills and instruction "is essential for mitigating bias, ensuring credibility, and upholding the integrity of published evidence syntheses." Conflicts of interest are disclosed: the authors are contracted by MLA to teach the course.

Limitations and caveats: This is a self-reported case report of the authors' own course, with no control condition, no objective skill assessment beyond an identical five-question pre/post-test, no response-rate or missing-data reporting for the test/feedback data, and no long-term retention or workplace-performance follow-up. Pre/post improvements could reflect course-specific test familiarity rather than durable skill. Table 2′s extraction is garbled (units unclear). Generalizability beyond the SRSS context is untested, and the proprietary course content is intentionally not disclosed.

Practical implications for a non-academic researcher using PubMed and AI: The course design is effectively a free blueprint: PubMed is the deliberate platform of choice "because it is fully accessible without restriction" and carries MeSH — the same logic a solo researcher should adopt. The module sequence (pre-search term harvesting → concept nesting → PRESS-

style self-review → PRISMA-S-style reporting) is exactly the workflow an AI-assisted individual should replicate, using AI as the breakout-group sparring partner. The finding that learners must accept "no absolute answers" is a useful inoculation against over-trusting any single AI-generated search string.

Notable quotes:

- "Review authors should work closely, from the start of the protocol, with an experienced medical/healthcare librarian or information specialist" (quoted from the Cochrane Handbook).
- "What matters the most is that the documented searches are sensitive in nature and easily reproducible."
- "An emerging theme on feedback from all cohorts was the realization that there are no absolute answers in systematic searching."

Connections: The most explicit node: it builds its objectives directly on Townsend (2017): A competency framework for librarians involved in systematic reviews and instructs readers to also review Spencer (2018): Roles for librarians in systematic reviews: a scoping review. Its Module 3 (PRESS) and Module 4 (PRISMA-S) correspond to the search-peer-review and reporting roles in Spencer's taxonomy and the "Reporting" competency in Townsend. Parker (2022): Literature reviews: tips from knowledge synthesis librarians's advice on preliminary searching and information-specialist involvement is the team-facing mirror of this course. Westrick (2021): Medical librarians' practices in locating clinical trials's survey identifies the study-design knowledge gaps a course like this could fill; Boden (2018): Learning while doing: MLA Systematic Review Project evaluation represents the informal, unpaid precursor to this formalized training; Minuti (2018): Teaching searching skills with a flipped classroom is the parallel teaching of the same PubMed/MeSH/PICO skills to medical students rather than librarians.

Boden (2018): Learning while doing: MLA Systematic Review Project evaluation — Program evaluation of the MLA Systematic Review Project: learning by doing at scale (JMLA 2018;106(2))

What the paper is and why it matters: Boden (University of Saskatchewan), Ascher (New York Medical College), and Eldredge (University of New Mexico) evaluated the Medical Library Association Systematic Review Project — an unprecedented experiment in which nearly 200 volunteer librarians worldwide formed 15 autonomous teams to conduct systematic reviews answering the profession's 15 top-ranked research questions (from a 2011 Delphi study). It matters because it is one of the few evaluations of a large-scale, volunteer-driven "learning by doing" model for building evidence-synthesis capacity in a profession, documenting both the promise (publications, awards, skills, networks) and the costs (slow progress, attrition, uneven leadership) of training through real projects rather than courses.

Methods in detail: A program evaluation via an eleven-item online survey (multiple-choice and short-answer) distributed by email to all project volunteers, with responses collected over two weeks beginning at the end of January 2016 and reminders sent one week and one day before closing. Because only nine of fifteen team leaders responded to a pre-survey roster-update request, contact information for six teams came from the original 2013 volunteer roster. Of the 199 original volunteers, an estimated 14 withdrew before starting and 20 withdrew early; the survey went to 198 individuals (one team leader was an author and was excluded). The University of Saskatchewan ethics board deemed the evaluation exempt. Analysis: frequency statistics for closed questions; thematic analysis of short-answer responses (six questions) using Braun and

Clarke's inductive approach, with two reviewers independently coding and three authors reaching agreement on themes. Outputs were collated and de-duplicated by type, though unique output counts could not be reliably computed from anonymous responses (cross-checked against a 2017 status report in Hypothesis).

Key results and numbers: Eighty of 198 potential respondents completed the survey (40% response rate): 8 team leaders and 72 team members. Team status among respondents: 22% on completed reviews; 53% on reviews in progress (30% contributing fully, 23% intermittently); 5% on restarted teams; 16% unable to work to completion; 4% blank. Estimated for the full 198: roughly 15% on completed teams, 20% on teams with new leaders, 65% still working. A January 2017 status report counted 3 teams completed and published (one winning the 2015 MLA Ida and George Eliot Prize), 8 making significant progress, and 4 very early or restarting; at least one team switched to a scoping review. Withdrawal reasons: other work commitments (n=8), project took longer than expected (n=6), job change (n=2), personal reasons (n=1), plus free-text reports of frustration, poor communication, and lack of centralized tools. 84% said the project should be repeated (66% for both answering questions and facilitating learning; 10% answers only; 8% learning only), 10% said no, 6% blank. 84% wanted to participate in another SR (76% in all aspects; 8% as search expert only). Thematic analysis yielded five themes: learning and experience (hands-on learning, practical skills, appreciation of SR complexity, leadership development), interpersonal/networking (international contacts, future support points), teamwork (dynamics, communication difficulties across time zones and hospital firewalls, pivotal team-leader roles, project-management needs), outcomes (publications, presentations across six countries, awards, credibility, tenure/promotion evidence), and barriers (no training/funding infrastructure, uneven participant knowledge, geographic dispersion, and ambiguous research questions causing huge result sets and long screening).

Conclusions and takeaways: The short answer to whether teams succeeded is "a qualified 'yes,' but slowly." Clear evidence exists that librarians gained skills, confidence, credibility, and networks through experiential learning, though learning quality varied with team dynamics and leader competence. Recommendations for future iterations: manage expectations, boundaries, and commitment continuously; establish strong communication routes through and beyond leaders; and balance team membership for subject knowledge, methodological expertise, and strong leadership. The training model was heterogeneous across teams — loosely "learning by doing" or experiential learning rather than any single formal approach — and the authors call for prospectively designed studies of experiential learning and possibly a coordinated educational component to amplify it.

Limitations and caveats: A self-selected 40% response sample from a single initiative; respondent/non-respondent characteristics could not be directly compared (restarted-team members likely slightly underrepresented); anonymous responses made output counts non-verifiable; the survey was not designed for generalizability to other learning-by-doing implementations; and team status figures for the full population were rough estimates. The evaluation occurred three years in, before most teams finished, so outcome measures are interim. Extraction note: Figure 1 (respondent counts per short-answer question) is not present in the text file, and one block-quote about ambiguous research questions appears displaced between the outcomes and barriers sections.

Practical implications for a non-academic researcher using PubMed and AI: The paper is a warning about the hidden costs of doing synthesis alone or with loosely committed collaborators: ambiguous questions inflate searches and screening time (one respondent's "result set huge, our screening time long"), and lack of leadership/infrastructure stalls projects. A solo

researcher can extract the countermeasures: sharpen the question before searching, pilot inclusion criteria, set deliverables and deadlines, and use tools for communication and workflow. It also validates learning by doing — building your first evidence synthesis with AI support may teach more than any tutorial, provided expectations and timelines are realistic.

Notable quotes:

- "It was the 'learning by doing' aspect that was the greatest value added of this experience."
- "TEAMWORK!!! We all learned so much from each other!"
- "I think the ambiguity of our research question has led to several challenges… our interpretation of the question and search were very broad, our result set huge, our screening time long, etc."

Connections: Eldredge is also co-author of Spencer (2018): Roles for librarians in systematic reviews: a scoping review, and the roles volunteers struggled to fill map onto Spencer's taxonomy and Townsend (2017): A competency framework for librarians involved in systematic reviews's competencies (process management/communication emerges as the biggest weakness). Its experiential-learning findings explain the demand that Farris (2024): Curriculum framework for systematic review competencies (librarians)'s formalized course and the SRSS later met — Boden's respondents wanted exactly the training infrastructure Farris describes. Westrick (2021): Medical librarians' practices in locating clinical trials's author was itself trained via the MLA Research Training Institute, the mentoring counterpart to this project. Minuti (2018): Teaching searching skills with a flipped classroom shows the same "learning then doing" logic applied to students, and Parker (2022): Literature reviews: tips from knowledge synthesis librarians's emphasis on planning and team communication echoes Boden's hardest-won lessons.

Westrick (2021): Medical librarians' practices in locating clinical trials — Survey: medical librarians' knowledge and practices in locating clinical trials for systematic reviews (JMLA 2021)

What the paper is and why it matters: A cross-sectional survey by Westrick (Rush University Medical Center librarian, an MLA Research Training Institute fellow) and Buchholz (Michigan State University nursing professor) examining three under-studied questions: how much information researchers give librarians when asking for clinical-trial searches, which resources librarians use to learn about study-design methodology, and how confident librarians are in that knowledge. It matters because locating trials of varied designs (not just RCTs) is a core but under-documented librarian task — the authors state this is the first survey on the topic — and because it exposes, even among self-identified expert searchers, a genuine knowledge gap with direct consequences for the completeness of systematic review searches.

Methods in detail: A cross-sectional, anonymous, 21-question survey administered electronically for one month (April 2020) via seven email listservs reaching medical librarians (list in Appendix B; response rate not calculable because of possible cross-listing duplication). Eligibility was restricted: respondents had to make search-term decisions for systematic reviews and affirm they had been credited as co-author or acknowledged by name on a systematic review published in a peer-reviewed journal. Survey development included review by three librarian experts and one measurement expert for accuracy, relevancy, and meaning before finalization. Structure: two screening questions, then 19 items covering demographics/occupation (4), researcher input (5), search-string development (2), librarian confidence (6), resources (1), plus an open-comment question, answered via Likert scales, multiple choice, and open text. Administration in RedCap;

analysis in Excel. Rush University Medical Center IRB deemed the study exempt (IRB ID #20012004-IRB01).

Key results and numbers: 120 librarians completed the survey. Demographics: 92% "Librarian," 6% "Information Specialist"; 47% university, 36% academic/university medical center, 9% hospital, 2% government, 5% other; 70% US, 11% Canada, 17% Europe, 2% Australia, 1% Asia. Estimated time to create a search strategy ranged from 2 to 100 hours. Researcher input: half of librarians had been asked to limit a search to clinical trials; of those researchers, 47% often/always specified the trial types, and fewer than half of those requests were RCT-only; 53% of librarians always/often knew their research's evidence categorization (prognostic, causation, therapeutic, etc.), and 44% said that knowledge affected their strategy. Search-string development (multi-select): hedges/pre-formulated filters were most popular (55%), own-built strings >one-third, database filters such as PubMed's Clinical Trials filter (27%), keyword exclusions (27%), and 23% simply did not search for trials; 34% used one method, 22% two or three, 13% four, 7% five. Confidence: 45% rated themselves experts, 48% proficient, 7% competent (no novices/beginners); 88% agreed/strongly agreed they could locate clinical trials; 83% agreed their knowledge of trial types was sufficient; but only 49% could locate "interventional studies" and 45% "longitudinal studies" without researching the design first. Resources: 40% wrote unnamed "textbooks" or "articles"; most-named source was JAMA's Users' Guide to the Evidence (9 mentions, 10%); colleagues/researchers 24%; PubMed/MeSH ~20%; internet/Google/Wikipedia 17%; LibGuides 16%; CEBM/JAMA/Cochrane each 10%.

Conclusions and takeaways: Even expert searchers — a deliberately elite sample — lack a definitive reference source for study-design methodology ("I don't have a single good place to look") and show a clear need and desire for such information, with one respondent begging for a class on study-design methodology for search-strategy designers. Researchers provide minimal, inconsistent guidance about the trial types they want, which hampers the use of category-dependent hedges (e.g., SIGN filters requiring diagnosis/epidemiology/outcome categories). Restricting searches to trials can omit pertinent evidence, and the 23% who don't search for trials at all may sometimes be right. The paper recommends that librarians push researchers to specify trial types, and suggests future examination of SR-course curricula for study-design content.

Limitations and caveats: The eligibility restriction (co-author/acknowledged on a published SR) makes the sample deliberately unrepresentative of all librarian searchers — and likely overstates the profession's confidence. No response rate was calculable, and the April 2020 COVID-19 shift to remote work may have affected responses. The survey did not cover trial registries (ClinicalTrials.gov), scoping reviews, or protocol registries (PROSPERO, OSF) that often specify intended trial types. Self-reported confidence and practice measures are subject to social-desirability bias, and the free-text resource question yielded many unidentifiable answers (40% unnamed "textbooks/articles"). Figures 1–4 are not present in the extraction.

Practical implications for a non-academic researcher using PubMed and AI: Before searching, explicitly decide and state which study designs you want — the paper shows this single communication step is routinely skipped and materially changes search strategy. PubMed's Clinical Queries and trial filters are one method among several; hedges (e.g., the Cochrane RCT filter) are powerful but category-dependent, so an AI assistant should be asked to explain design categories before applying them. Finally, treat "I feel confident" as weak evidence: even 88%-confident experts needed to research what "longitudinal" meant — a caution against trusting gut feeling or a fluent AI answer about study design.

Notable quotes:

- "If I want to revisit a study type . . . I don't have a single good place to look."

- "Sometimes it seems that a study can have more than one [category]; i.e., prognostic AND diagnostic."
- "I really want to take a class on study design methodology. If there were a class/webinar like this — study design methodology for librarians designing search strategies for SRs — I'd take it in a heartbeat!"

Connections: Empirically validates the "Research methodology" competency gap in Townsend (2017): A competency framework for librarians involved in systematic reviews's framework and supplies the pedagogical demand that Farris (2024): Curriculum framework for systematic review competencies (librarians)'s course partially answers (its drug/chemical searching module echoes the same EBM-source territory). Its hedges/filters content lands within the largest of Spencer (2018): Roles for librarians in systematic reviews: a scoping review's 18 roles. Its author's MLA RTI mentorship links to the training-infrastructure theme of Boden (2018): Learning while doing: MLA Systematic Review Project evaluation and to Minuti (2018): Teaching searching skills with a flipped classroom's teaching of Clinical Queries to students. Parker (2022): Literature reviews: tips from knowledge synthesis librarians's planning-phase advice — align question with review type — is the team-level fix for the researcher-input failures this survey documents.

Minuti (2018): Teaching searching skills with a flipped classroom — Flipped-classroom teaching of PubMed searching skills to medical students (Med Ref Serv Q 2018;37(2):119–131)

What the paper is and why it matters: Librarians at Albert Einstein College of Medicine's D. Samuel Gottesman Library (Minuti, Sorensen, Schwartz, King, Glassman, Habousha) describe designing and iterating a flipped-classroom module teaching first- and second-year medical students how to search the literature for evidence-based articles. It matters for two reasons: it shows librarians embedding themselves early in medical curricula (previously these skills were taught only in third-year clerkships) to satisfy LCME accreditation Standard 6.3 on self-directed, lifelong learning, and it is an unusually candid account of implementation — mandatory-but-unenforced tutorials, broken microphones and buzzer systems, student cheekiness in anonymous polls, and poor long-term retention of the taught skills.

Methods in detail: Constraints drove the design: one-hour, one-shot sessions for each cohort of 190+ students, held in a new 36-table Education Center with round tables, tabletop microphones, and wireless flat-screens, with the podium in the room's middle. With an instructional-design specialist from the Office of Medical Education, librarians chose a two-pronged flipped approach: (1) self-paced online tutorials built with already-licensed tools — LibGuides (content sites), LibWizard (gradable quizzes, interactive step-by-step exercises embeddable beside the live PubMed screen), and Adobe Captivate (narrated videos) — released a week before class, nominally mandatory, due three days before the session so quiz exports into Excel could reveal knowledge gaps for tailoring; and (2) in-class reinforcement. The tutorials followed a "see-one, do-one, teach-one" model: video demonstration, individual PubMed exercise, then group application. Two tutorials: "Smart Searching" (first-years; three modules — Search, Find, Evaluate — teaching PICO formulation, search-term selection, MeSH searching in PubMed, PubMed refinement tools, full-text retrieval, applicability appraisal, built on reused third-year EBM clinical cases) and "Searching the Literature for Clinical Evidence" (second-years; adds PubMed's Clinical Queries; in 2017 second-years were directed to the new module only, with the first three as optional refreshers). Classroom sessions: 2016 used group case work (PICO → MeSH search → discussion; second-years also critiqued a located randomized trial with EPHEM faculty present);

after students complained of redundancy, 2017 added a six-question Jeopardy game (Search/Find/Evaluate categories, buzzer, granola-bar prize) and PollEverywhere anonymous audience-response questions.

Key results and numbers: Exercise completion (2016): first-years (n=191 required) 70–76% across the five exercises; second-years (n=183) 85–99% plus 98% on the Clinical Queries exercise. 2017: first-years (n=190) 74–86%; second-years (n=191) 86% on the required Clinical Queries exercise, but less than half completed the optional refresher exercises. Evaluation surveys (via SurveyMonkey, emailed by the Office of Medical Education): 2016 response rates 52% (first-years) and 61% (second-years); 92% and 95% respectively found the tutorial satisfactory for teaching PubMed searching, but only 74% and 72% found the classroom session reinforcing. 2017 response rates collapsed to 8% and 24% (reason unknown); among respondents, tutorial satisfaction was 82% (first-years) and 71% (second-years); only 41% agreed Jeopardy was fun for reinforcement and 42% found PollEverywhere useful for testing knowledge; the authors state no significant conclusion can be drawn from 2017. Documented problems: mis-worded quiz questions, unenforced "mandatory" tutorials with no penalty mechanism, technology failures (microphone amplification in 2016; broken buzzer system in 2017 forcing a PollEverywhere substitute), persistent redundancy complaints even after the 2017 redesign, unprofessional anonymous poll answers (display was switched off), and third-year clerkship teachers observing that students retained very little from the prior year's sessions while being far more receptive once patient care made the skills consequential.

Conclusions and takeaways: The online component succeeded: students largely enjoyed it, mostly answered correctly, and learned at their own pace; the material is repurposable for residents and graduate students. The live session was the weak link — technology glitches and redundancy detracted from educational value, though in-person collaboration retains value. The project demonstrated librarian value to faculty (first-time inclusion in the first- and second-year curriculum; Einstein meeting two LCME requirements; faculty impressed with tutorial quality). Planned retention fixes: require second-years to redo all exercises with harder cases, anchor cases to real third-year patients, and add a third-year self-assessment of skill retention; also allow a full week between tutorial deadline and class to enable data-informed tailoring.

Limitations and caveats: This is a descriptive program report, not an evaluation: no objective learning outcomes were measured (only quiz completion, quiz answers, and subjective satisfaction surveys with collapsing response rates), and the authors themselves disclaim conclusions from 2017′s 8%/24% response rates. Contamination and cohort effects abound (the 2017 second-years had prior exposure), "mandatory" was unenforced, and confounders like the novel room's acoustics are inseparable from pedagogy. The single-institution, single-discipline setting limits transferability. The text is otherwise clean, with only minor extraction quirks (a stray "none defined" line near the abstract).

Practical implications for a non-academic researcher using PubMed and AI: The tutorial contents are a free, self-paced PubMed curriculum still available at the cited LibGuide URLs — PICO framing, MeSH term selection, PubMed filters/refinement, Clinical Queries, and applicability appraisal are precisely the skills a non-academic researcher needs, and the "see-one, do-one, teach-one" sequence is a good template for self-instruction with an AI tutor in the "teach-one" role. The retention finding is the key warning: skills learned before they are needed decay; plan refresher practice immediately before real searching projects rather than trusting one-off learning.

Notable quotes:

- "Students begin by watching a video demonstration of the searching process. In the next step, they work through a search exercise on their own. Finally, during the classroom session, students work in small groups and help each other to complete a new assignment."
- "Teaching in such a huge room is difficult even under the best of circumstances."
- "Librarians who teach EBM during the third-year clerkship noticed in 2017 that students retained very little knowledge from the online and in-class sessions of 2016."

Connections: Embodies the "teaching" role documented by Spencer (2018): Roles for librarians in systematic reviews: a scoping review (which cites the flipped-classroom modality) and teaches, at beginner level, the same PubMed/MeSH/PICO/Clinical Queries content that Farris (2024): Curriculum framework for systematic review competencies (librarians) teaches to librarians at professional level — including the identical flipped/modular pedagogy and pre/post assessment idea. Its discovery that learners need skills "when they need them" resonates with Boden (2018): Learning while doing: MLA Systematic Review Project evaluation's finding that experiential learning on real projects works best, and with Westrick (2021): Medical librarians' practices in locating clinical trials's evidence that even experienced searchers need targeted design-methodology refreshers. Townsend (2017): A competency framework for librarians involved in systematic reviews's framework describes exactly the competencies (question formulation, comprehensive searching, reporting) this module tries to instill; Parker (2022): Literature reviews: tips from knowledge synthesis librarians's PICO and planning advice is the research-team version of the same content.

Parker (2022): Literature reviews: tips from knowledge synthesis librarians — Literature reviews: tips from knowledge synthesis librarians; the PIECES framework (JGME 2022)

What the paper is and why it matters: A short practical column by two knowledge-synthesis librarians — Robin Parker (Dalhousie University, Evidence Synthesis and Information Services Librarian) and Lindsey Sikora (University of Ottawa, Head of Research Support, Health Sciences) — opening the Journal of Graduate Medical Education's Literature Review Series. Aimed at GME educators and researchers, it distills advice "based on our experience as knowledge synthesis librarians" (and, per the conclusion, from supporting "hundreds of reviews") into the PIECES framework from Foster and Jewell's 2017 book *Assembling the Pieces of a Systematic Review: A Guide for Librarians*: Planning the review, Identifying studies and resources, Evaluating and appraising the evidence, Collecting and combining data, Explaining the synthesis, and Summarizing the findings. It matters as a compact, authoritative bridge between librarian expertise and the health-professions-education researchers who most often attempt reviews without one.

Methods in detail: Not an empirical study — a structured experiential guidance piece organized around the six PIECES phases, each with core content and "Pro Tips." Planning: develop and refine a research question aligned to the review type (narrow/predefined for systematic reviews vs. broad/evolving for state-of-the-art reviews), optionally using frameworks like PICO; build a team; choose a synthesis method; run a preliminary search to confirm the review's need and sense the volume and types of sources (iteratively feeding back into question refinement); inventory available tools (citation management, review management, analysis, reporting) and team expertise per phase (via a table mapping phases to information specialists, review-methods experts, content experts, project managers, and research assistants, and to tools from email/Trello to Archie and JBI SUMARI); set inclusion/exclusion criteria (population, publication type, language, time period, study type, sample size). Identifying: database searching determined by

the question, supplemented by grey literature (conference abstracts, technical reports, association websites); work with an information specialist (IS) — invite one onto the team or as a consultant — who can develop the search to best practices and PRISMA-S reporting standards; ISs can be found at academic/hospital libraries, professional associations, and knowledge-synthesis networks, and will often provide free consultation to unfunded projects. Evaluating: screen for eligibility (systematic for SRs, purposive sampling for other types), appraise with tools like the Cochrane risk-of-bias tool or MERSQI where appropriate. Collecting: extract variables per the question and analysis plan; extraction templates for SRs/scoping reviews, inductive thematic approaches for interpretive reviews. Explaining: transparently draw together data from disparate sources, explicitly describing team decisions. Summarizing: coherent narrative with figures/tables per the review type's reporting standards.

Key results and numbers: No participant data — the deliverable is the framework plus six Pro Tips: (1) determine the ideal publication venue and target audience during planning to align question and purpose; (2) for GME topics, search education-specific databases such as ERIC and Education Research Complete, not just PubMed; (3) pilot inclusion criteria on a sample (e.g., 10% of retrieved records) to operationalize criteria and improve inter-rater reliability; (4) use review-management software for large citation volumes and collaborative screening; (5) pilot the data-extraction form on a handful of included studies to catch omissions and improve reliability; (6) use analysis software where applicable (RevMan for meta-analyses; qualitative analysis software for qualitative syntheses). A table maps all six phases to team-member expertise and tool types; a Box lists key resources (BEME, Campbell, Cochrane Handbook, JBI Manual, the Systematic Review Toolbox website, and three review-typology articles including Grant & Booth's 14-type typology); an online supplement provides a glossary. Note: the pdftotext extraction badly jumbles the reading order (the introduction, editor's note, table, and box interleave with body text), though the content is fully recoverable.

Conclusions and takeaways: Common elements cut across all review types, and the most successful review teams "have a good plan, use expert advice and technological tools, and establish excellent communication practices." Reviews are overlapping and iterative rather than linear, but must start with a concrete plan for conduct and reporting. Information specialists materially strengthen reviews through search design and PRISMA-S-grade re

porting, and cost is not a barrier for unfunded teams willing to invite an IS as collaborator or consultant. A well-conducted review, whatever its type, is decision-support evidence — for policy, programs, or assessment design — and its credibility depends on methodological transparency at every PIECES phase.

Limitations and caveats: All advice is expert opinion from two librarians' practice; no evidence citations support most specific claims (e.g., the 10% piloting figure), and no data on outcomes of following versus ignoring the tips. It is explicitly introductory — the JGME series' later installments cover review types in depth — so it should not be used as a stand-alone methods guide for a full systematic review. It is also GME-oriented: the ERIC-database tip is domain-specific. The framework itself is borrowed (Foster & Jewell 2017) rather than validated here. Extraction caveat: text order is significantly scrambled; quotes below were reconstructed from the jumbled flow.

Practical implications for a non-academic researcher using PubMed and AI: This is the batch's most directly usable checklist. For a solo PubMed-based researcher with AI assistance: run the preliminary search to test whether a review is needed; pilot screening criteria on ~10% of records; pilot your extraction form; use free review-management tools; and, in the AI era, treat the "information specialist" role as one you must deliberately perform — asking the AI to act

as an IS for search design and then checking its output against PRISMA-S's 16 reporting items. The Pro Tips convert directly into project-management steps.

Notable quotes:

- "We recommend inviting an IS to join the review team or to serve as a consultant."
- "Pro Tip 5: Piloting the data extraction form with a handful of included studies will catch omissions in the data collection process and improve the inter-rater reliability of the project."
- "The most successful review teams have a good plan, use expert advice and technological tools, and establish excellent communication practices."

Connections: Cites Spencer (2018): Roles for librarians in systematic reviews: a scoping review directly as evidence for involving an information specialist, and its phase-to-expertise table implicitly assumes the competencies Townsend (2017): A competency framework for librarians involved in systematic reviews formalized. Its "Collecting and combining data" phase is where Lizarondo (2025): Data extraction and transformation in mixed methods systematic reviews's extraction-and-qualitization guidance lives; its planning-phase emphasis (question clarity, piloting) answers the ambiguity problems Boden (2018): Learning while doing: MLA Systematic Review Project evaluation documented. Its PRISMA-S recommendation is exactly the content of Farris (2024): Curriculum framework for systematic review competencies (librarians)'s Module 4, and its preliminary-search advice mirrors Farris's Module 1; Westrick (2021): Medical librarians' practices in locating clinical trials supplies the empirical backdrop for why IS expertise matters, and Minuti (2018): Teaching searching skills with a flipped classroom shows where the next generation of researchers should first acquire these habits.

15. Addendum — supplementary sources

Godin (2015): Applying systematic review search methods to the grey literature: a… — the four-strategy grey-literature template behind checklist item 5.

Haby (2016): Best methodologies for rapid reviews for evidence-informed decision… — the rapid-review methods evidence: no agreed definition, shortcuts widely used with thin empirical backing, and an RCT suggesting rapid reviews improve clarity and accessibility for decision makers.

Lizarondo (2025): Data extraction and transformation in mixed methods systematic reviews — the JBI guidance on data extraction and qualitization that the AI extraction papers are automating.

Boden (2018): Learning while doing: MLA Systematic Review Project evaluation — the learning-at-scale evidence: ~200 volunteer librarians, 15 review teams, 84% would repeat it.

Johnson (2022): The Systematic Review Toolbox: tools to support evidence synthesis — the community tool catalogue (235 tools, 112 guidance documents; stakeholder engagement the biggest gap).

Godino (2023): How to structure Microsoft Excel documents for systematic reviews — the free six-step Excel audit-trail method (OCR-transcribed on its page).

16. Addendum II — process, quick-reference, and appraisal sources

Aromataris (2014): Constructing a search strategy and searching for evidence: a guide — the systematic-review-grade tutorial: the PICO logic grid, pilot searches to harvest keywords and MeSH terms, field codes, truncation, and the worked animal-assisted-therapy search string.

Elston (2019): Getting the most out of online literature searches — tips for advan… — the one-page quick reference (Boolean operators, field tags, truncation, Limits, Clinical Queries).

Pollock (2022): Methodological quality, guidance, and tools in scoping reviews (pro… — the registered protocol for mapping appraisal tools for scoping reviews.

Grimshaw (2012): Knowledge translation of research findings — the KT framing: five questions, and the argument that the unit of translation should be the up-to-date systematic review.

Minuti (2018): Teaching searching skills with a flipped classroom — the flipped-classroom education case (online tutorials + gamified reinforcement under one-hour constraints).

Godino (2023): How to structure Microsoft Excel documents for systematic reviews — see section 15.