PRISMA-ScR-report

PRISMA-ScR reporting document — scoping review of literature-searching methodology and AI-assisted evidence synthesis

Extension used: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR; Tricco et al., Ann Intern Med 2018;169:467–473). **Note on review type:** this work is a scoping-style mapping of methods literature with a deployed knowledge-product (a curated reading library and synthesis report in Capacities). Formal critical appraisal of sources was omitted by design; in its place a two-tier relevance/practicality screening was applied (rationale under item 11).

TITLE

1. Title. "How to search PubMed well with AI assistance (2026): a scoping mapping of literature-searching methodology and AI-assisted evidence synthesis, with a deployed 48-paper reading library and synthesis report."

Identified as a scoping review-style mapping.

RATIONALE

2. Rationale. Literature-searching methodology has a mature, consensus-based core (peer-reviewed search strategies, validated-ish filters, librarian roles) while AI-assisted screening, extraction, and strategy-building have surged since 2023 with thin empirical grounding. Practitioners — including non-academic researchers — lack a plain-language, evidence-linked pathway through both literatures. The rapid growth of AI-methods papers (e.g., a 95-study scoping review documenting post-2023 explosion; GPT-4 present in 74% of included diagnostic-AI studies) outpaces narrative coverage. A scoping mapping was chosen to chart the breadth of both literatures, identify gaps, and produce a usable knowledge product rather than a pooled estimate.

OBJECTIVES

3. Objectives.

1. Chart the breadth of literature on (a) PubMed/MEDLINE search methodology and (b) AI-assisted evidence synthesis (screening, extraction, interfaces, reporting).
2. Empirically characterize retrieval behaviour of candidate search strings against PubMed (live hit counts, pairwise combination matrices).
3. Produce a curated, practical, accountability-linked reading library (48 papers) with structured summaries and full texts, plus a synthesis report.

Research questions.

- RQ1: What is the state of knowledge on PubMed search methodology (strategy building, filters, peer review, supplementary searching)?
- RQ2: What is the state of knowledge on AI assistance across the evidence-synthesis workflow?
- RQ3: How robust is the evidence base for each, and what are the gaps?

METHODS

4. Protocol and registration. No protocol was registered a priori. The working protocol evolved iteratively in a documented workspace log (`mesh-archive.md` and pipeline scripts); all screening, harvesting, and curation decisions are timestamped in that log and in this document. This is a limitation (see item 17).

5. Eligibility criteria.

Inclusion:

- Papers on literature-searching methodology (strategy building, search filters and their validation, peer review of search strategies, supplementary searching, reporting standards) OR on AI assistance in evidence synthesis (screening, data extraction, interfaces, LLM-generated reporting outputs) OR on the human infrastructure of evidence synthesis (librarian roles, competencies, training).
- Any publication type with retrievable full text or abstract; any date; English-language sources (one source pair originally German with full English abstract/translation).

Exclusion (applied during curation):

- Specialty-specific clinical LLM application studies (e.g., radiography, orthopaedics, oncology-specific tool guides) — out of scope.
- Purely bibliometric field maps were retained only when methodologically instructive (dual-database design, science-mapping workflow).
- Five records were removed at final curation (2026-11-18): two specialty/policy pieces (family-medicine AI editorial-policy survey; dental librarian roles survey), one bibliometric medical-education mapping, one AI-in-cancer-research guide (full text judged too extensive for the workflow's scope; PDF retained in Zotero), one methods-protocol companion retained only in part. See item 13 for the removal list.

6. Information sources.

- **Bibliography of record:** a 273-entry Zotero collection assembled by the reviewer from prior reading, saved PubMed sets, and citation chasing (the "searched" pool).
- **PubMed/MEDLINE** via NCBI E-utilities (esearch/esummary/efetch), used for: per-DOI record verification, metadata extraction, and live search-combination measurements.
- **NLM MeSH records** (id.nlm.nih.gov RDF and NCBI efetch) for controlled-vocabulary verification (entry terms, tree positions, qualifiers).
- **Zotero storage PDFs** as the full-text source for transcription (pdftotext; Windows OCR for one scanned PDF).
- **Grey/methods infrastructure:** ISSG Search Filters Resource, Systematic Review Toolbox, EQUATOR Network (referenced within included sources rather than searched separately).

7. Search.

- **Controlled-vocabulary farming (bootstrapping).** MeSH descriptors, entry terms, tree positions, and qualifiers were farmed per-PMID from three seed PMID sets supplied by the reviewer (43 unique papers) and verified against NLM records. This produced the corpus vocabulary spine: *Information Storage and Retrieval; Systematic Reviews as Topic; Databases, Bibliographic; Artificial Intelligence; Databases, Factual; Review Literature as Topic; PubMed; Search Engine; Natural Language Processing; Large Language Models; Automation; Semantics; Generative Artificial Intelligence.*

- **Candidate strings.** Four thematic term blocks (guides; filters/QC; supplementary/reporting/KT; AI core; librarian craft) were combined into 13 self-contained OR-block strings, each bounded by an a-priori ≤150-hit precision cutoff after empirical tuning.
- **Empirical string characterization.** 420 pairwise AND combinations across term blocks were measured live in PubMed (hit counts recorded in the workspace archive, measured 2026-11-11; counts drift with database updates). A per-term OR schema was then fixed (the 13 strings).
- **Example string (one of 13):**

```
("Systematic Reviews as Topic"[Mesh] OR "Review Literature as Topic"[Mesh] OR "Information Storage and Retrieval"[Mesh] OR "Databases, Bibliographic"[Mesh] OR "Databases, Factual"[Mesh] OR "PubMed"[Mesh] OR "Search Engine"[Mesh] OR "MEDLINE"[Mesh] OR "Artificial Intelligence"[Mesh] OR "Generative Artificial Intelligence"[Mesh] OR "Natural Language Processing"[Mesh] OR "Large Language Models"[Mesh] OR "Automation"[Mesh] OR "Semantics"[Mesh])
```

(13 such strings were run separately; combined pool ≈ 6,400 records before de-duplication.)

- **Supplementary methods.** Per-DOI backward verification of the entire bibliography; citation chasing within the bibliography (pearl growing); no forward citation searching; no hand-searching of journals; no contact with experts.

8. Selection of sources of evidence.

- Screening was performed by one reviewer (the user) with AI-assisted pre-screening (deterministic keyword-tier scoring implemented as scripted, reproducible regex tiering — not a black-box classifier; all tier assignments reviewable in `screening-report*.txt`).
- **Tier 1 (relevance screen), n = 117 abstracts:** keyword-tier scoring (A = PubMed/search-method terms ×3; B = AI terms ×3; C = review-task terms ×2; D = mechanics ×1). Verdicts: 35 HIGH / 66 MEDIUM / 14 LOW / 2 OFF-SCOPE.
- **Tier 2 (title + MeSH screen), n = 132 records without PubMed abstracts:** title keywords + indexed-MeSH tiers. Verdicts: 35 HIGH / 78 MEDIUM / 18 LOW / 2 OFF-SCOPE.
- **Merged master pool:** 70 unique HIGH records (35 + 35, de-duplicated); curation added curated MEDIUM keepers (quality-assessment and reporting-standards papers retained; clinical-topic LLM applications removed; 7 removed) → full pool of 207.
- **Practicality screen:** the 207 were scored for hands-on usability (guides, checklists, tools, workflows) vs technical depth (benchmarks, architectures) → 53 practical records → final curation removed 5 (see item 13) → **48 included**.
- **Reference standard and error context.** Human abstract-screening error rates in comparable settings are ~10.8% (Wang et al., included in the corpus); the single-reviewer screening here inherits comparable misclassification risk. Disagreements and artefacts were documented, not silently resolved.

PRISMA flow (counts).

```
Zotero bibliography (records identified) ..... 273
  → verified per-DOI against PubMed ..... 249 matched
      (24 not retrievable in PubMed: preprints, non-indexed venues,
       one transient API failure)
Records with abstract available ..... 117
Records in PubMed without abstract ..... 132 (+2 later transcribed
                                              from newly added PDFs)
After relevance screen (117 abstracts) ..... 35 HIGH / 66 MED / 14 LOW / 2 OFF
After title+MeSH screen (132 no-abstract) ..... 35 HIGH / 78 MED / 18 LOW / 2 OFF
Merged unique HIGH pool ..... 70
Curation (+ curated MEDIUM keepers, -7 clinical-topic
  LLM applications) → full pool ..... 207
```

```
Practicality screen → practical set ..................... 53
Final curation (5 removed 2026-11-18) → included ....... 48 papers
  (52 Capacities article pages initially built;
   4 trashed, 1 slimmed to summary-only)
```

9. Data charting process.

- A data charting form was the **Capacities article page template**: metadata (thematic phase, practicality score, citation, DOI), structured summary (Type / Aim / Methods / Results / Conclusions & takeaways), reading-notes section, and full transcribed text.
- Charting was performed by the reviewer with AI-assisted transcription and drafting; summaries were drafted from full texts and revised by the reviewer. The charting form was not piloted across multiple reviewers (single-reviewer limitation, item 17).

10. Data items.

- Bibliographic: authors, year, title, journal, DOI, PMID, publication type.
- Corpus-alignment: thematic phase (1–4), practicality score (0–16 scale from the practicality screener), screening tier and flags.
- Content items charted in summaries: aims, methods (search strategies described, tools, datasets, metrics), key quantitative results, conclusions, limitations, practical implications.
- Full-text items: complete transcribed text (47 of 48 pages; 1 page summary-only by curation).

11. Critical appraisal of sources. *Omitted, with rationale.* PRISMA-ScR permits omission of formal critical appraisal in scoping reviews. In its place: (a) a two-tier relevance/practicality screen determined inclusion emphasis; (b) internal-consistency checking of source documents during deep reading (documented artefact ledger: contradictory counts across abstract/results in four sources; garbled tables; OCR artefacts); (c) the reviewer's curation decisions recorded with reasons. No formal risk-of-bias instrument was applied to included sources.

12. Synthesis of results. Narrative and structured. Outputs: (1) per-paper structured summaries and deep-dive analyses charted onto 48 Capacities pages; (2) a thematic synthesis organized in six themes (canonical search process; filters/quality control; supplementary searching/KT; AI screening & extraction; AI field maps; librarian infrastructure); (3) an integrated five-layer working model for PubMed + AI searching; (4) a six-tension analysis; (5) a 10-item practical checklist for non-academic researchers. The synthesis lives as the workspace "State of the Field Report" page (linked from the Project).

RESULTS

13. Selection of sources of evidence. See the PRISMA flow above. Of 273 bibliography records: 249 verified in PubMed per-DOI; 117 yielded full abstracts; 132 lacked abstracts in PubMed; 24 were not retrievable (preprints, non-indexed venues, one transient API failure). After screening and curation, **48 papers were included** as the final reading library. During the project, five papers were removed by reviewer curation (two specialty applications, one bibliometric medical-education mapping, one AI-guide full text, one librarian-roles survey) and the library was rebuilt accordingly; removals are documented with dates and reasons in the workspace log.

14. Characteristics of sources of evidence. 48 included papers, publication years 2005–2026 (median ~2021; 13 papers from 2025–2026, reflecting the AI-surge). Types: 9 plain-language tutorials/guides; 10 filter/quality-control studies (4 Cochrane methodology reviews, 2 SWAR-type empirical comparisons, 1 PRESS guideline, 1 citation-searching cross-section, 1 HTA monograph, 1 scoping protocol); 13 AI-workflow papers (4 screening/extraction evaluations, 2

interface/curation platforms, 1 abstract-generation evaluation, 1 error-rate study, 1 methods-caution study, 4 field maps/guides); 16 librarian-infrastructure papers (roles, competencies, curricula, surveys, tool catalogues). Full corpus characteristics are charted on the 48 article pages (each with phase, practicality score, structured summary, full text where available).

15. Synthesis of results. Full synthesis is deployed as the workspace report ("State of the Field Report — Literature Searching Methodology & AI"). Headline findings:

- The search process is consensus-standardized across nine guidance documents (eight key stages), but supplementary searching is its least standardized stage.
- Human abstract-screening error (~10.8%) sets a realistic benchmark; LLM pipelines now reach human-comparable screening sensitivity (~90%, $\varkappa$ 0.71) and extraction F1 up to 93–98 on concordant two-LLM designs.
- Text-mining strategy builders are faster but less sensitive; methodological filters degrade under external validation; no filter for observational or diagnostic-accuracy designs combines high sensitivity with usable precision.
- LLM evaluation methodology is immature (repeat-prompting correlation can inflate effective sample size >100-fold).
- Reporting standards exist (PRISMA-S, PRESS, GAMER) but journal endorsement of AI-specific reporting guidelines is rare (5% in surveyed specialties).
- The librarian multiplier is evidence-backed (18 roles; 6 competencies; curricula; program evaluation).
- Gaps: LLM-assisted strategy building is unevaluated; filter validation is dated; vocabulary bootstrapping (affiliation-anchored MeSH farming) is undocumented in the literature; GAMER/PRISMA-S adoption is unmeasured.

DISCUSSION

16. Summary of evidence. See the deployed synthesis report (sections 1–16: canonical process; filters/QC; human infrastructure; AI wave; risks/integrity/reporting; integrated five-layer model; six structural tensions; gaps; practical checklist). The corpus supports a human-in-the-loop architecture: vocabulary farming → Boolean/MeSH skeleton → tool-assisted screening/extraction with human arbitration → PRESS/PRISMA-S/GAMER accountability.

17. Limitations.

- **Single-reviewer screening** with scripted (not independent double) pre-screening; misclassification risk comparable to the ~10.8% human baseline documented in the corpus.
- **Language restriction** (English-language corpus; one German source with English abstract).
- **Database restriction** (PubMed/MEDLINE-centred; Embase, Scopus, Web of Science, grey-literature databases not systematically searched — consistent with the corpus's own finding that supplementary searching is under-standardized).
- **Extraction artefacts:** 24 of 48 full texts carry documented pdftotext/OCR artefacts (garbled tables, OCR character confusions); a consolidated artefact ledger is maintained in the report (§14.0). Four source documents contain internal inconsistencies (contradictory counts, p-value drift) that automated ingestion would propagate.
- **Selection bias:** the seed bibliography was assembled from the reviewer's prior reading (judgment sample, not random); relevance tiers are keyword-based, not semantic.
- **Temporal drift:** PubMed hit counts and MeSH assignments drift; measurements are dated 2026-11-11/18.

- **No registered protocol; no dual independent review** of screening decisions.

18. Implications of results.

- For researchers: the practical checklist (report §11) is directly usable; vocabulary farming should precede string building; automation belongs in assistive roles with human arbitration.
- For methodology research: the affiliation-anchored MeSH-farming bootstrap (Phase 0) and multi-instrument capture–recapture estimation of unobserved relevant literature are untested and now specified; both are candidate SWARs.
- For journals: GAMER/PRISMA-S endorsement is rare — adoption would close the accountability gap this corpus documents.

FUNDING

20. Funding. None. No external funding was received for this work. The author declares no conflicts of interest. All pipeline scripts, screening reports, and data files are retained in the local workspace.

Appendix A — PRISMA-ScR checklist mapping (item → location)

PRISMA-ScR item	Reported here	Section
1 Title	Title line	top
2 Rationale	Rationale	§2
3 Objectives	Objectives + RQs	§3
4 Protocol & registration	declared not registered	Methods §4
5 Eligibility criteria	Inclusion/exclusion	Methods §5
6 Information sources	Zotero, PubMed E-utilities, NLM MeSH, PDFs	Methods §6
7 Search	MeSH farming, 13 strings, example string, supplementary methods	Methods §7
8 Selection	Two-tier screen + curation + PRISMA flow	Methods §8
9 Charting process	Capacities page template, AI-assisted drafting	Methods §9
10 Data items	Bibliographic/corpus/content/full-text items	Methods §10
11 Critical appraisal	Omitted — rationale given; replaced by tiered screen + internal-consistency checking	Methods §11
12 Synthesis	Narrative + structured charting + deployed report	Methods §12

13 Selection of sources	PRISMA flow counts	Results §13
14 Characteristics	48 papers, years, types	Results §14
15 Synthesis of results	Headline findings; full report	Results §15
16 Summary of evidence	Deployed synthesis report	Discussion §16
17 Limitations	6 limitation classes	Discussion §17
18 Implications	Researchers / methodology research / journals	Discussion §18
19 (optional) Funding & conflicts	Funding	§20
20 (optional) — n/a for updated guidance	—	—

Appendix B — Included papers (48)

Phase 1 (plain-language guides, 9): Watson 2020; Blümle 2019; Motschall 2005; Aromataris 2014; Elston 2019; Muin 2006; Kim 2022; Cooper 2018; Booth 2016. Phase 2 (filters & quality control, 10): McGowan 2016 (PRESS); Escobar Liquitay 2023; Beynon 2013; Li 2019; Lefebvre 2017; Dullea 2026; Paynter 2021; Briscoe 2020; Luo 2025 (GAMER); Pollock 2022. Phase 3 (AI-assisted workflow, 13): Lund 2024; Li 2025; Khan 2025; Pilz 2024; Mahboubi 2025; Wang 2020; Gallo 2025; Yarar 2026; Su 2025; Ma 2026; Garry 2026; Demir 2025; Perez-Lopez 2024. Phase 4 (librarian infrastructure, 16): Godin 2015; Haby 2016; Grimshaw 2012; Begagić 2025; Lizarondo 2025; Spencer 2018; Townsend 2017; Farris 2024; Boden 2018; Westrick 2021; Minuti 2018; Parker 2022; Johnson 2022; Verma 2024; Godino 2023; Laynor 2024.

Removed at curation (2026-11-18, with reasons): O'Brien 2025 (specialty editorial-policy survey); Schvaneveldt 2021 (dental librarian roles survey); Liu & Huang 2026 (medical-education bibliometric); Abdel-Wahab 2019 (health-website standards — full text judged too extensive; PDF retained); Perez-Lopez 2024 (AI guide — full-text transcription removed, summary retained).

Companion artifacts: Capacities Project/Book/48 article pages + State of the Field Report; final-search-corpus.bib (53), final-search-corpus-full-207.bib (207), pubmed-corpus-44.bib (44); mesh-archive.md; screening reports and deep-dive files avaiable at this link: https://drive.proton.me/urls/MMMR0PX0XR#hZ3xkAJOiLss