

# Selecting AI Tools for Medical Research: A Practical Framework for Effective Utilization

 Ansel Kaplan EROL<sup>1</sup>,  Almıla EROL<sup>2</sup>

<sup>1</sup>Georgia Institute of Technology, School of Computer Science, Atlanta, GA, USA

<sup>2</sup>Adjunct Faculty, Psychiatry & Psychology, Mayo Clinic, Rochester, MN, USA

**A**rtificial intelligence (AI) is assuming an increasingly prominent role in medicine, with important implications for medical research, scholarly publishing, and the preparation of research manuscripts (1). A growing array of AI tools is becoming available to support research, underscoring the importance of guidance to help researchers select tools appropriate to each stage of the research process. This editorial provides practical, stepwise guidance for selecting AI tools across the research process. It does not attempt a comprehensive survey of all AI tools available to medical research but highlights the most relevant tools and outlines good practices for their use.

While AI presents wide-ranging opportunities, medical journals have responded by developing policies and guidelines to define its responsible use and disclosure (2). Although efforts to develop consensus guidelines for the ethical use of AI are still ongoing, transparency regarding AI use is widely recognized as central to research integrity. Researchers and authors are not only expected to disclose and describe the use of AI in detail but also required to critically evaluate AI-generated outputs at every stage of the research process and retain full responsibility for all content, decisions, and conclusions produced with AI assistance. General principles of AI use for research are depicted in Box 1.

## Box 1. Principles of AI use across all research stages

- Use several tools together; no single tool covers the whole process.
- For facts and references, use tools that search real papers, not tools that write from memory.
- Give more human review where a mistake costs more. For example, grammar needs little review; decisions about which studies to include, or about cause and effect, need full review.
- Keep a record of AI use: tool, version, date, your prompts, the raw output, and what you corrected.
- AI does not remove methodological work. It turns doing into checking, and requires more documentation, not less.
- AI search tools find mostly English-language literature, so careless use widens the gap for regional and non-English research.
- Most evidence about these tools comes from the companies that sell them. Independent studies are needed.

## Highlights

- Many AI tools are available to support medical research.
- This paper outlines how to leverage the most relevant AI tools for medical research.
- In medical research, smart, responsible and transparent use of AI tools is essential.
- AI tools can accelerate early research tasks but cannot replace human input.

The research process consists of multiple stages: identifying a research question, conducting a comprehensive literature review, formulating research questions and hypotheses, designing the research to test the hypothesis, collecting data via experimental, observational or other methods, analyzing the data, interpreting findings, and writing the manuscript (3). The stages of the research process and the AI tools that can assist in each stage are explained in detail below and summarized in Table 1.

**1) Identifying a research question:** Identifying a research question starts with selecting a topic and identifying unanswered questions. You can start with a question checklist such as PICO (patient, intervention, comparison, outcome) (4), or FINER (feasible, interesting, novel, ethical, relevant) (5). These are simple templates that turn a broad interest into one clear question. AI tools then help you test that question.

• **AI chat tools:** AI chat tools such as **ChatGPT** (<https://chatgpt.com>), **Claude** (<https://claude.ai>), **Gemini** (<https://gemini.google.com>) can be used to write several versions of your question and to argue against each one. Ask yourself, “What result would show this idea is wrong?” Weakness: these tools tend to agree with the user and may claim research gaps without proof. Verify claimed gaps with real search.

• **Consensus** (<https://consensus.app>): An AI search engine that reads published papers and tells you whether the evidence says yes, no, or mixed. Use it to see if your question already has an answer. Weakness:

**Cite this article as:** Erol AK and Erol A. Selecting AI Tools For Medical Research: A Practical Framework for Effective Utilization. Arch Neuropsychiatry 2026;63:655–658. doi: 10.29399/npa.29595

for many papers it sees only the abstract, so it can miss differences between studies.

**2) Literature review:** To refine the direction of the research, it is essential to review the current knowledge and to identify debates, contradictions, and gaps. Use medical databases (e.g. PubMed, Embase, PsycINFO, ClinicalTrials.gov) as the foundation and use AI to add speed. The databases hold the information but have limited AI features. The AI tools work on top of them.

- **Semantic Scholar** (<https://www.semanticscholar.org/>): A free search site covering more than 200 million papers (6). It writes short summaries and shows which citations matter most. Use it to find papers. Weakness: the short summaries leave out methods and limitations.

- **ResearchRabbit** (<https://www.researchrabbit.ai/>): A free tool that finds papers related to a few papers you are interested in, to fill in gaps missed by key-word search. Weakness: the user must start with good papers to achieve good results through this tool.

- **Elicit** (<https://elicit.com/>): Searches papers by meaning instead of keywords, pulling details (sample size, outcome, effect) from each paper into a table. Use it to build your first evidence table. Weakness: Elicit can miss relevant papers, so it is not enough for a systematic review on its own.

- **NotebookLM / Gemini Notebook** (<https://notebooklm.google.com/>): You upload the papers you have selected, and it answers questions using only these papers (7). This design prevents invented references. Use it to compare studies, organize notes, and draft summaries of your reading. Weakness: occasional errors parsing tables and requires good paper selection for effective use.

- **Screening tools: Rayyan** (<https://www.rayyan.ai/>), **ASReview** (<https://asreview.nl/>), **Covidence** (<https://www.covidence.org/>) are screening tools for review teams. They rank abstracts so the most likely matches appear first (8,9). Use them when you must screen thousands of records; keep two human reviewers and decide in advance when to stop. Weakness: over time, reviewers trust the ranking too much and read less carefully (10).

- **scite** (<https://scite.ai/>): Shows how later papers cite a study: do they support it, contradict it, or only mention it (11)? Use it before you build on a key claim, because later work may have overturned it. Weakness: it is probabilistic and can occasionally misclassify.

**3) Formulating research questions and hypotheses:** Formulating a research question requires narrowing broad interests into specific, researchable questions. You need to write a clear problem statement to define the purpose and context of your study and come up with a hypothesis. ClinicalTrials.gov (<https://clinicaltrials.gov/>) and PROSPERO (<https://www.crd.york.ac.uk/prospERO/>) can be searched to make sure no one is already doing the same study.

- **AI chat tools** (see item 1): They can be used to list possible confounders and to turn the problem statement into a testable hypothesis. Weakness: these tools often describe a link as a cause, and they call ideas “novel” without checking.

- **Google Co-Scientist** (<https://deepmind.google/blog/co-scientist-a-multi-agent-ai-partner-to-accelerate-research/>): A multi-agent system built on Gemini that generates candidate hypotheses, critiques them in internal debates, and ranks them in a tournament. Some of its proposals have been validated in laboratory experiments (12). Use it when the research goal is defined but the space of possible hypotheses is large. Weakness: it draws on open-access literature only, access is currently limited, and its proposals require expert screening for plausibility and feasibility. Access is via Google Cloud for Organizations.

**4) Research design:** Designing a research study includes choosing appropriate methods. Plan your research design, including variables, controls, and data collection procedures. This step requires creativity and AI helps least here, and mistakes here cost the most: a design error cannot be fixed later.

- **AI chat tools** (see item 1): They can be used to compare study designs (cohort, case-control, trial), draft a protocol outline, and play the role of a harsh reviewer. Tools that can run code will also draft sample-size calculations. Weakness: cannot replace expert statistical analysis or ethics review.

- **AI reporting guidelines: SPIRIT-AI** (13), **CONSORT-AI** (14), **TRIPOD+AI** (15) should be used if the study tests an AI system itself.

**5) Data collection:** The research plan should be applied systematically. Keep detailed notes and track all procedures. Adjust methods if necessary while maintaining scientific rigor. Identifiable patient data must never be entered into a public AI tool (16).

- **Transcription tools: Whisper** (<https://github.com/openai/whisper>) is an example which runs on a local computer so recordings never leave the institution. It turns interview recordings into draft text (17). Weakness: errors gather around drug names, psychiatric terms, numbers, and negative statements (“no history of...”). Always verify correctness.

- **Clinical text tools: Amazon Comprehend Medical** (<https://aws.amazon.com/comprehend/medical/>), and **Azure Text Analytics for Health** (<https://learn.microsoft.com/azure/ai-services/language-service/text-analytics-for-health/overview>) are cloud services that find diagnoses, drugs, and personal identifiers in clinical text and match them to standard codes such as The International Statistical Classification of Diseases (ICD) codes (18). Weakness: you have to test them first on a sample your team has checked by hand, and use them only inside systems your institution approves.

- **ATLAS.ti** (<https://atlasti.com/>): Can be used for qualitative analysis. It suggests first-draft categories for interview text. Treat the suggestions as drafts. Weakness: AI categories can be generic and can lose the context of what a patient actually said. A person must build the final coding scheme, and a second person should recode part of the data as a check.

**6) Data analysis:** An important rule is that raw patient-level data must never be shared with public AI tools, even after names are removed. Combinations of ordinary demographic and clinical variables are often sufficient to re-identify individuals (19). Data entered into public tools may also be retained or used for model training and could in principle be accessed and analyzed by others before the original investigators publish. Use only deployments approved by the institution under a data processing agreement, or models running on local infrastructure, and prefer aggregate or synthetic data for exploratory work in public tools. Appropriate statistical or qualitative analyses must be used. Do the final statistics in standard statistical software. AI helps around it.

- **AI Chat Tools** (see item 1): Those listed can write and execute analysis code on tables you upload. Use it to explore data and find quality problems. Weakness: scripts produced should be tested and reproduced to validate accuracy.

- **H2O** (<https://h2o.ai/>): An autoML that builds many prediction models automatically and ranks them (20). Use it only as a comparison point. Weakness: the top-ranked model is not proof of clinical value. Mistakes in data processing can inflate the score.

**Table 1.** The stages of the research process and available artificial intelligence (AI) tools

| Research stage                               | AI Tools                                                                                                             | Use                                                                                                                            |
|----------------------------------------------|----------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------|
| Research question identification             | AI chat tools (ChatGPT, Claude, Gemini, etc.)<br>Consensus                                                           | Draft and stress-test candidate questions<br>Check whether the evidence already answers candidate questions                    |
| Literature review                            | Semantic Scholar<br>ResearchRabbit<br>Elicit<br>NotebookLM<br>Screening tools (Rayyan, ASReview, Covidence)<br>scite | Find and map papers<br>Extract evidence tables<br>Rank abstracts for screening<br>Check whether later work supports key claims |
| Research question and hypothesis formulation | AI chat tools (ChatGPT, Claude, Gemini, etc.)<br>Google Co-Scientist                                                 | Rewrite questions in testable forms<br>Generate and rank candidate hypotheses                                                  |
| Study design                                 | AI chat tools (ChatGPT, Claude, Gemini, etc.)                                                                        | Compare designs<br>Draft protocol outlines<br>Simulate a hostile reviewer                                                      |
| Data collection                              | Whisper<br>Amazon Comprehend Medical<br>Azure Text Analytics for Health<br>ATLAS.ti                                  | Transcribe recordings<br>Extract clinical entities<br>Draft qualitative coding schemes                                         |
| Data analysis                                | ChatGPT data analysis<br>GitHub Copilot<br>H2O                                                                       | Explore data<br>Draft and translate code<br>Benchmark prediction models                                                        |
| Interpretation of findings                   | NotebookLM<br>scite                                                                                                  | Compare results with selected studies'<br>Recheck whether landmark findings still stand                                        |
| Manuscript preparation                       | AI chat tools (ChatGPT, Claude, Gemini, etc.)<br>AI editing tools (Paperpal, Writefull, Trinka)<br>Zotero            | Restructure and tighten text<br>Check grammar and journal requirements<br>Catch retracted references                           |

**7) Interpretation of findings:** This stage is where you integrate your results with prior knowledge, compare them with your hypothesis and the existing literature, and evaluate their significance, limitations, and implications. A useful instruction for any AI tool at this stage is “Label each conclusion as: direct result, combined result across studies, inference, or speculation. Do not treat a significant p-value as clinical importance.”

- **NotebookLM** (<https://notebooklm.google.com>): Upload your included studies and your own results, and ask how they compare. Weakness: it can write one smooth story even when the studies disagree. You must check the differences yourself.
- **scite** (<https://scite.ai>) (see also item 2): Can be used before you compare your findings with a landmark study, to check whether later work has contradicted that study.

**8) Manuscript preparation:** Write a structured research paper including introduction, methods, results, and discussion sections (21). Follow ethical guidelines. Authorship follows International Committee of Medical Journal Editors (ICMJE) rules (22). AI is never an author, and the authors stay responsible for everything, including every reference. Do not use AI tools to cite references and opt for standard reference managers instead, as AI can cause issues related to the reliability and precision of references (2).

- **AI chat tools** (see item 1): Can be used to reorganize sections, shorten text, and write plain-language summaries. Give it one firm rule: “add no facts, numbers, or references”. Weakness: if allowed, the tool fills gaps with invented detail.
- **AI editing tools:** **Paperpal** (<https://paperpal.com>), **Writefull** (<https://www.writefull.com>), **Trinka** (<https://www.trinka.ai>) are trained on academic text. They check grammar, consistency, and journal requirements. Very useful for authors whose first language is not English. Weakness: review each change one by one. A rewrite can quietly change “may cause” into “causes.”

- **Zotero** (<https://www.zotero.org>): A reference manager with retraction alerts. It helps to catch withdrawn papers.

Across research stages, practical AI tools can accelerate tasks such as locating literature, proposing methodological options, and editing manuscripts. These systems expand the range of possibilities, but they do not replace scientific judgment. Researchers must still evaluate suggestions, determine study eligibility, assess evidence certainty and causality, and approve the final manuscript. As AI takes on more of the work of generating text, code, and candidate hypotheses, verifying these outputs becomes an essential part of the research process.

Efforts to develop reporting guidelines for AI use in research are underway (23). In the meantime, responsible use rests on best practices. These practices include transparent documentation of tools, versions, prompts, and corrections; careful tool selection with a preference for source retrieval when facts and references are at stake; and author responsibility for every claim (22). AI can support research but responsibility for rigor remains human.

Psychiatric research requires particular care. The field’s complex constructs, cultural variability, and heterogeneous populations make it vulnerable to AI outputs that oversimplify or reinforce bias, while the sensitivity of patient data raises additional privacy concerns. These risks call for careful oversight, but they also offer opportunities for psychiatric researchers to contribute. The field’s experience in studying automation bias and assessing reliability and validity can inform independent evaluations of these tools. As the tools evolve, the guiding principles remain straightforward: match the tool to the research task, scale human review to the consequences of error, and retain clear responsibility for the final work.

**Keywords:** Artificial intelligence, AI-assisted research, Framework, Medical research, Research, Scientific writing

## REFERENCES

- Hepdurgun C. The present and future of artificial intelligence applications in psychiatry. *Arch Neuropsychiatry* 2024; 61:1–2. [\[Crossref\]](#)
- Flanagin A, Perlis RH, Bibbins-Domingo K. Updated guidance for author use of AI in medical publication. *JAMA* 2026; 336: 926–928. [\[Crossref\]](#)
- Erol A. How to Conduct Scientific Research? *Arch Neuropsychiatry* 2017; 54:97–98. [\[Crossref\]](#)
- Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. *ACP J Club*. 1995;123:A12–3. [\[Crossref\]](#)
- Cummings SR, Browner WS, Hulley SB. Conceiving the research question and developing the study plan. In: Browner WS, Newman TB, Cummings SR, Grady DG, Huang AJ, Kanaya AM, Pletcher MJ Eds, *Designing clinical research*, 5th ed, Philadelphia, Lippincott Williams and Wilkins, 2022, p. 17–25.
- Fricke S. Semantic Scholar. *J Med Libr Assoc* 2018; 106:145–147. [\[Crossref\]](#)
- Shor R, Greene EA, Sumberg L, Weingrad AB. AI tools in academia: evaluating NotebookLM as a tool for conducting literature reviews. *Psychiatry* 2026; 89:82–91. [\[Crossref\]](#)
- Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. *Syst Rev* 2016; 5:210. [\[Crossref\]](#)
- van de Schoot R, de Bruin J, Schram R, Zahedi P, de Boer J, Weijdemans F, et al. An open source machine learning framework for efficient and transparent systematic reviews. *Nat Mach Intell* 2021; 3:125–133. [\[Crossref\]](#)
- Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J Am Med Inform Assoc* 2012; 19:121–127. [\[Crossref\]](#)
- Nicholson JM, Mordaunt M, Lopez P, Uppala A, Rosati D, Rodrigues NP, et al. scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. *Quant Sci Stud* 2021; 2:882–898. [\[Crossref\]](#)
- Gottweis J, Weng WH, Daryin A, Tu T, Palepu A, Sirkovic P, et al. Towards an AI co-scientist. *arXiv preprint* 2025; arXiv:2502.18864. [\[Crossref\]](#)
- Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ; SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med* 2020; 26:1351–1363. [\[Crossref\]](#)
- Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med* 2020; 26:1364–1374. [\[Crossref\]](#)
- Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024; 385:e078378. [\[Crossref\]](#)
- World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Geneva: World Health Organization; 2024. <https://www.who.int/publications/i/item/9789240084759>. Accessed September 12, 2026.
- Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I. Robust speech recognition via large-scale weak supervision. *Proceedings of the 40th International Conference on Machine Learning (ICML) 2023*; 28492–28518.
- World Health Organization. ICD-11 International Classification of Diseases 11th Revision. <https://icd.who.int/browse/2025-01/mms/en#486722075>. Accessed Sept. 12, 2026.
- Rocher L, Hendrickx JM, de Montjoye YA. Estimating the success of re-identifications in incomplete datasets using generative models. *Nat Commun* 2019; 10:3069. [\[Crossref\]](#)
- LeDell E, Poirier S. H2O AutoML: scalable automatic machine learning. 7th ICML Workshop on Automated Machine Learning (AutoML) 2020.
- Erol A. Basics of writing original research papers. *Arch Neuropsychiatry* 2024; 61:193–194. [\[Crossref\]](#)
- ICMJE. Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals (Updated January 2026). <https://www.icmje.org/recommendations/> Accessed Sept. 12, 2026.
- International Science Council. AI disclosure in research: towards a global reporting standard. <https://council.science/our-work/ai-disclosure-in-research/> Accessed Sept. 12, 2026.