ARP Rheumatology 2025;4:76-79
https://doi.org/10.63032/HWCR2841

EDITORIAL

AI in medical research: boosting discovery or weakening critical search skills?

Nack A[1], Benavent D[2]

Two decades ago, a clinician searching for evidence on methotrexate in rheumatoid arthritis might have devoted an afternoon scrolling through manuscripts and photocopying abstracts for paper journals. Today the same search ends in seconds: an Artificial Intelligence (AI) driven platform parses natural language and pushes a curated list of articles to the top of the screen. The gain in speed is undeniable, yet the tradeoffs are less visible. If opaque algorithms now choose which papers reach us, evidence based medicine risks devolving into evidence filtered medicine. The challenge for rheumatology, and every field that stakes patient care on systematic appraisal, is to adapt automation without losing the human scrutiny that detects bias and contextualises findings. As we welcome these powerful tools, should we also insist on preserving the skill of reading beyond the first page of search results?

Until the late 20th century, literature searches in medicine primarily relied on manual methods. Prior to the advent of PubMed in the 1990s, rheumatologists were required to manually browse printed indexes like the Index Medicus, review scientific journals in paper format and directly consult experts to remain up-to-date[1]. The introduction of PubMed facilitated literature searches, allowing structured electronic database queries in Medline/PubMed, EMBASE, and the Cochrane Library, through Boolean operators, MeSH terms, and systematic filters[2]. Retrieval involved meticulous abstract screening and citation cross-referencing, which ensured thoroughness but required significant time and expertise. This process, while labor-intensive, encouraged critical engagement with the literature: researchers refined their strategies, assessed relevance directly, and retained full control over source selection, minimizing external biases and avoiding reliance on opaque algorithms. However, even precise queries could miss key studies due to inconsistent indexing, and manual searches proved increasingly difficult to scale with the growing volume of biomedical literature[1,2].

[1] Rheumatology Department, Germans Trias i Pujol University Hospital, Barcelona, Spain;
[2] Rheumatology Department, Bellvitge University Hospital, Barcelona, Spain.

Correspondence to: Diego Benavent
E-mail: d_benavent@hotmail.com

PubMed now hosts more than 38 million records, a corpus impossible to navigate unaided[3]. The rise of AI and machine learning has reshaped this process. Traditional keyword-based strategies are now supplemented by AI-powered tools such as ChatGPT, Elicit, Consensus, OpenEvidence, and Scite[4], which analyze citation patterns, semantic relationships, and relevance networks to prioritize results[5]. Large Language Model (LLM) interfaces, semantic search engines and citation network analysers promise to slash screening time, surface hidden connections and even extract structured evidence tables, augmenting and automatizing clinician abilities.

These systems offer notable advantages. They reduce search and screening time dramatically, letting users filter thousands of articles in seconds and expedite evidence synthesis[6]. Elicit, for instance, can extract and organize key findings from primary studies into structured summaries, helping researchers interpret evidence without reading full texts line by line. Consensus applies LLMs to the biomedical literature to determine whether existing studies support or contradict a given hypothesis, thereby assisting both retrieval and synthesis. OpenEvidence is tailored to clinical practice: it delivers concise, evidence-based responses grounded in guidelines and recent publications, streamlining point-of-care decision-making. Within ChatGPT, the DeepResearch functionality enables targeted exploration of scientific questions using curated sources, facilitating synthesis without navigating multiple databases manually[4].

In addition to speed, AI models excel at identifying semantic relationships between studies, revealing conceptual links and evidence clusters that manual searches may miss. These tools increasingly assist in evidence synthesis by summarizing, comparing, and organizing conflicting findings. Importantly, they don't replace clinical or methodological expertise, but can enhance it by accelerating routine steps and allowing users to focus more on interpretation and application[6]. Beyond information retrieval, generative AI may contribute cognitively and even emotionally to the research process, a potential "cybernetic teammate"[7]. In recent studies, individuals working with AI matched the performance of human teams in innovation tasks and reported more positive emotional experienc-

(C) 2025 Sociedade Portuguesa de Reumatologia. This is an open access article under the CC BY-NC-ND license.

es, suggesting that AI may reshape not only how we search and synthesize knowledge, but also how we collaborate and learn[7].

While AI enhances discovery, concerns remain. How do AI-powered search engines prioritize studies? Do these systems have access to full-text sources, or are they limited to open-access metadata? Are researchers losing the ability to critically evaluate literature? Indeed, liberation may come at a cost. Ranking algorithms learn from citation counts, journal impact and prior clicks; the result is a feed that privileges prestige and momentum[8]. Negative, underpowered or regionally funded studies may vanish from view, producing a corpus that looks cleaner than reality. Most tools still draw chiefly on abstracts; full texts behind paywalls are mined only when institutional licences permit[9]. Methodological nuances or subgroup analyses often reside exclusively within the full manuscripts, invisible to models that scrape titles and summaries. The user confronted with a polished paragraph may feel every stone has been turned, unaware that key studies never entered the algorithmic pipeline[10]. In this regard, overreliance on AI summaries may threaten critical thinking, reducing researchers' capacity to independently assess literature and design rigorous strategies[8].

Generative systems add another layer of risk. Chatbots tasked with bibliographic retrieval have fabricated references and misattributed quotations[11]. These systems frequently present answers with unwarranted confidence, even when uncertain, increasing the risk of misinformation in clinical and academic settings[12]. Users may also overtrust AI output simply because it appears authoritative: in a recent study, AI-generated responses were rated more favorably than human ones—until participants learned the source was artificial[13]. Worryingly, ChatGPT has also cited retracted scientific articles without flagging their status, presenting flawed information as vàlid[14]. Without oversight, these tools may influence not only how information is accessed but what is deemed valid, potentially impacting medical research and clinical practice in unintended ways.

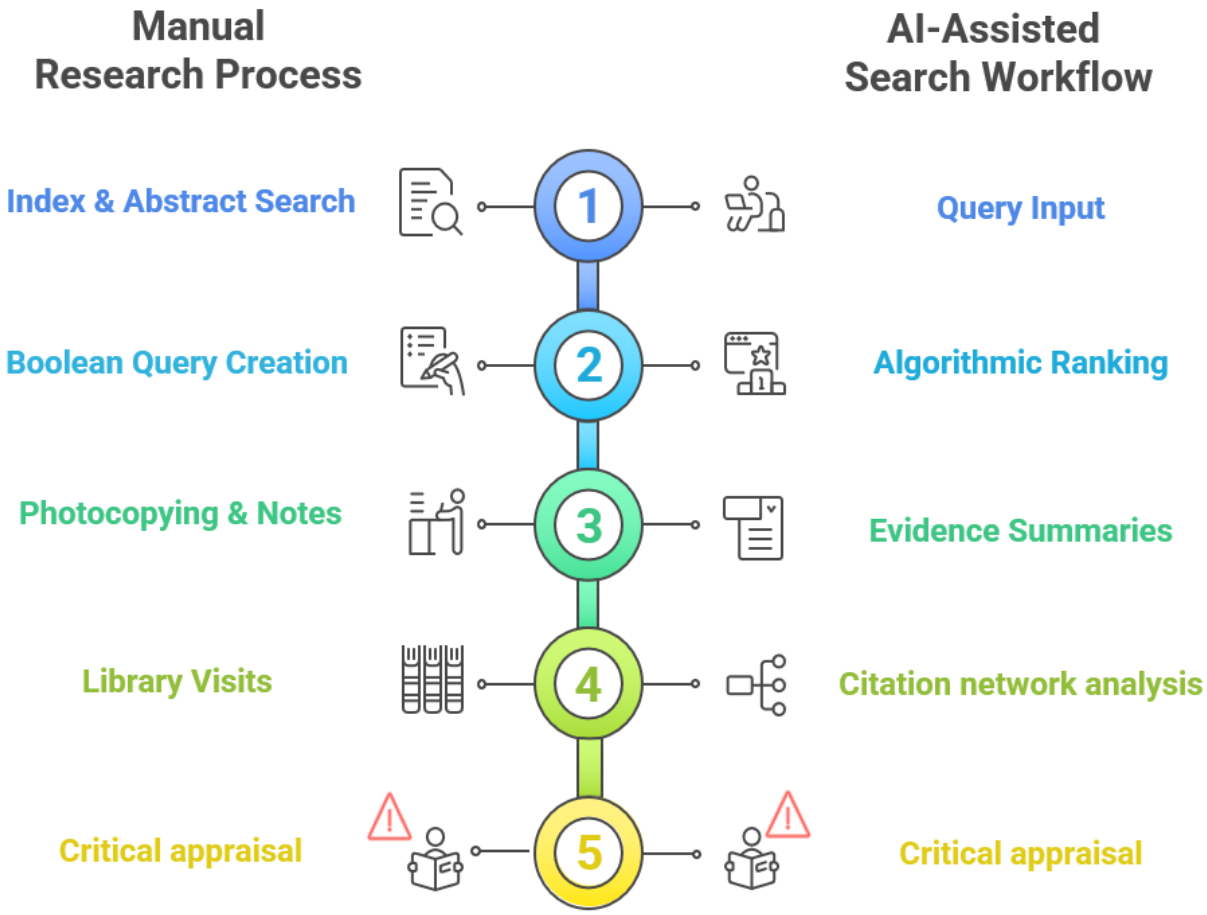

Figure 1. Evolution of the literature search.

Rather than rejecting AI, the solution may lie in integrating it critically into medical workflows[8]. Success depends on introducing robust digital-scholarship competencies (algorithmic literacy, critical appraisal of machine output and data-governance principles) early in medical training. Researchers must learn to validate AI-generated content against primary sources to ensure accuracy and avoid misinformation. Training should promote hybrid search strategies that combine Boolean logic with AI-powered tools to reduce bias and improve coverage. Equally essential is systematic instruction on algorithmic bias and model provenance, enabling future clinicians and scientists to interrogate opaque ranking systems and preserve methodological rigour in evidence synthesis[15]. Journals and funding bodies should, in parallel, require transparent disclosure of AI assistance in literature searches, reinforcing reproducibility and sustaining accountability across the research process[16].

In summary, are we truly ready to entrust generative language models to decide which studies earn a place in our systematic reviews today? Despite growing interest and promising advances, current AI-based tools are not yet sufficiently validated to replace traditional methods, particularly in the phase of literature assessment[17]. While some tools like ASReview, DistillerSR, and Rayyan have shown great utility in screening phases and LLMs offer semantic search capabilities, they lack external validation in autonomously handling systematic reviews; they often rely on partial corpora (abstracts without full texts) and conceal their ranking logic[17,18]. Comparative studies highlight gains in speed but also reveal risks of bias, reduced sensitivity, and opacity in ranking algorithms[18]. Therefore, AI should be integrated cautiously and complementarily, with human oversight, hybrid search strategies, and transparent reporting to preserve methodological rigor.

A PROPOSAL FOR RHEUMATOLOGY

Given the growing reliance on AI-assisted research, rheumatology should promote responsible use through targeted training. Residency programs should teach algorithmic literacy alongside epidemiological methods, reinforcing critical evaluation skills and ensuring trainees can verify and contextualise machinegenerated outputs while preserving their own critical judgment[15]. Comparative evaluations, such as AI-assisted reviews versus conventional systematic reviews, may help to map accuracy and bias[8]. Finally, scientific societies could publish guidance on responsible deployment, balancing efficiency with vigilance.

AI has transformed literature search, but the responsibility for interpreting evidence remains with clinicians. While it improves efficiency, AI may also introduce bias, over-filtering, and risks to critical appraisal. Its integration into medical practice must support—not replace—clinical judgment, and training should emphasize literacy, validation, and vigilance. Without oversight, AI may distort evidence-based medicine, prioritizing algorithmic outputs over scientific rigor. It remains a powerful tool, but clinical decisions must stay in human hands. Automation may determine what rises first on screen; judgement must decide what endures in practice.

REFERENCES

1. Douglas A, Capdeville M. From Index Medicus to the Palm of Our Hands-What's "App-ening" in Graduate Medical Education. J Cardiothorac Vasc Anesth 2020; 34: 2133-2135. https://doi.org/10.1053/j.jvca.2020.02.055
2. Goodfellow LT. An overview of how to search and write a medical literature review. Respir Care 2023; 68: 1576-1584. https://doi.org/10.4187/respcare.11198
3. https://pubmed.ncbi.nlm.nih.gov/about/ChatGPT. Accessed in May 30th 2025
4. Sequí-Sabater JM, Benavent D. Artificial intelligence in rheumatology research: what is it good for? RMD Open 2025; 11: e004309. https://doi.org/10.1136/rmdopen-2024-004309
5. Fiorini N, Canese K, Starchenko G, Kireev E, Kim W, Miller V, et al. Best Match: new relevance search for PubMed. PLoS Biol 2018; 16: e2005343. https://doi.org/10.1371/journal.pbio.2005343
6. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Syst Rev 2015; 4: 5. https://doi.org/10.1186/2046-4053-4-5
7. Dell'Acqua F, Ayoubi C, Lifshitz-Assaf H, Sadun R, Mollick ER, Mollick L, et al. The cybernetic teammate: a field experiment on generative AI reshaping teamwork and expertise. Harvard Business School Working Paper No. 25-043. https://ssrn.com/abstract=5188231. Accessed in May 30th 2025. https://doi.org/10.2139/ssrn.5188231
8. Feng Y, Liang S, Zhang Y, Chen S, Wang Q, Huang T, et al. Automated medical literature screening using artificial intelligence: a systematic review and meta-analysis. J Am Med Inform Assoc 2022; 29: 1425-1432. https://doi.org/10.1093/jamia/ocac066
9. OpenAI. How ChatGPT and our foundation models are developed. https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed Accessed in May 30th 2025.
10. Jin Q, Leaman R, Lu Z. PubMed and beyond: biomedical literature search in the age of artificial intelligence. EBioMedicine 2024; 100: 104988. https://doi.org/10.1016/j.ebiom.2024.104988
11. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. Sci Rep 2023; 13: 14045. https://doi.org/10.1038/s41598-023-41032-5
12. Tow Center for Digital Journalism. We compared eight AI search engines. They're all bad at citing news. Columbia Journalism Review. https://www.cjr.org/tow_center/we-compared-eight-ai-search-engines-theyre-all-bad-at-citing-news.php. Accessed in April 19th 2024.
13. Parshakov P, Naidenova I, Paklina S, Matkin N, Nesseler C. Users favor LLM-generated content-until they know it's AI. Inf Technol Tour. Published Online First: 2024. doi:10.1007/s40979-023-00146-z. https://doi.org/10.1007/s40979-023-00146-z

14. Gu T, Feng H, Li M, Wang G, Gu W. Alarm: retracted articles on cancer imaging are not only continuously cited by publications but also used by ChatGPT to answer questions. J Adv Res. Published Online First: 2025. doi:10.1016/j.jare.2025.03.020. https://doi.org/10.1016/j.jare.2025.03.020
15. Malerbi FK, Nakayama LF, Gayle Dychiao R, Zago Ribeiro L, Villanueva C, Celi LA, et al. Digital education for the deployment of artificial intelligence in health care. J Med Internet Res 2023; 25: e43333. https://doi.org/10.2196/43333
16. Kocak Z. Publication ethics in the era of artificial intelligence. J Korean Med Sci 2024; 39: e249. https://doi.org/10.3346/jkms.2024.39.e249
17. Chelli M, Descamps J, Lavoué V, Trojani C, Azar M, Deckert M, et al. Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: comparative analysis. J Med Internet Res 2024; 26: e53164. https://doi.org/10.2196/53164
18. Ge L, Agrawal R, Singer M, Kannapiran P, De Castro Molina JA, Teow KL, et al. Leveraging artificial intelligence to enhance systematic reviews in health research: advanced tools and challenges. Syst Rev 2024; 13: 269. https://doi.org/10.1186/s13643-024-02682-2