Review

Rodrigo Ayala-Yáñez*, Amos Grünebaum and Frank A. Chervenak

# Integrating generative AI in perinatology: applications for literature review

https://doi.org/10.1515/jpm-2025-0392
Received July 15, 2025; accepted September 2, 2025;
published online September 23, 2025

**Abstract:** Perinatology relies on continuous engagement with an expanding body of clinical literature, yet the volume and velocity of publications increasingly exceed the capacity of clinicians to keep pace. Generative artificial intelligence (GAI) tools – such as ChatGPT4, Claude AI, Gemini, and Perplexity AI – offer a novel approach to assist with literature retrieval, comparison of clinical guidelines, and manuscript drafting. This study evaluates the strengths and limitations of these tools in maternal-fetal medicine, using structured clinical prompts to simulate real-world applications. Perplexity AI demonstrated the best citation accuracy, while ChatGPT4 and Claude excelled in content summarization but required manual verification of citations. In simulated trials, GAI tools reduced the time to generate clinically relevant summaries by up to 70 % compared to traditional PubMed searches. However, risks such as hallucinated references and overreliance on machine-generated text persist. Use cases include summarizing aspirin use guidelines for preeclampsia and comparing ACOG vs. NICE protocols. GAI should be viewed as a supportive assistant, not a substitute, for expert review. To ensure responsible integration, clinicians must develop AI literacy, apply rigorous oversight, and adhere to ethical standards. When used judiciously, GAI can enhance efficiency, insight, and evidence-based decision-making in perinatal care.

**Keywords:** generative artificial intelligence; perinatology; systematic review and automation; practice guidelines

***Corresponding author: Rodrigo Ayala-Yáñez**, ABC Medical Center I.A.P., Av. Carlos Graef Fernández 154-339, Tlaxala, Cuajimalpa, Mexico City, 05300, Mexico, E-mail: drayalagineco@gmail.com. https://orcid.org/0000-0003-2548-3208
**Amos Grünebaum and Frank A. Chervenak,** Northwell Health, Hempstead, NY, USA

## Introduction

Perinatology is a data-intensive discipline that relies heavily on up-to-date literature, guidelines, and outcome-based research. Staying abreast of evolving recommendations from authoritative bodies like the American College of Obstetricians and Gynecologists (ACOG), the Society for Maternal-Fetal Medicine (SMFM), and international counterparts such as the National Institute for Health and Care Excellence (NICE) or the World Health Organization (WHO) is essential to providing evidence-based care. Yet the volume and pace of published research can be overwhelming. Over one million articles are published annually across biomedical journals, and filtering relevant, high-quality data for clinical application requires both time and expertise. This challenge is particularly acute in perinatal medicine, where new studies on interventions, risk prediction models, and maternal-fetal outcomes appear weekly [1, 2].

The rise of generative AI offers a partial solution to this burden. Tools like ChatGPT4, Claude AI, Gemini, and Perplexity AI use natural language processing to analyze, summarize, and even compare literature and guidelines based on user-defined prompts [2, 3]. These systems are trained on large corpora of text and can produce structured responses that mirror scientific language, making them appealing tools for clinicians seeking rapid insight. Importantly, they allow users to query in plain English, reducing the technical barrier that often exists with database search platforms like PubMed.

To evaluate the utility of generative AI tools in perinatology literature review, we conducted a comparative analysis using four major platforms: ChatGPT4 (OpenAI), Claude AI (Anthropic), Perplexity AI, and Gemini AI (Google DeepMind). We designed 12 structured prompts covering common clinical inquiries such as “Compare aspirin timing in preeclampsia prevention” and “Summarize guidelines for gestational hypertension management.” Each prompt was entered into all four platforms, and responses were assessed independently by two maternal-fetal medicine specialists for accuracy, citation validity, depth of synthesis, and clinical utility. Time-to-output was recorded from prompt submission to complete response. We further

tested citation traceability by cross-referencing cited sources with PubMed and publisher databases. Qualitative performance was categorized as “high,” “moderate,” or “low” based on clarity, relevance, and presence of factual errors. Discrepancies were resolved by consensus.

Generative AI tools were selectively used to support specific aspects of manuscript development. ChatGPT-4.0 (OpenAI) assisted with language refinement, restructuring, and formatting of previously drafted sections. Claude AI (Anthropic) was employed for grammar review and clarity improvements, while Perplexity AI supported literature retrieval by suggesting links to relevant studies and PubMed articles. All references proposed by AI were manually verified by the authors using PubMed, Google Scholar, and official journal databases. No generative AI platform was used to create novel scientific content of draw conclusions. AI outputs were critically reviewed for factual accuracy, citation integrity, and relevance. In cases of citation hallucination or inconsistency, content was discarded or replaced using verified sources.

No patient specific or identifiable data were entered into AI systems. All AI use adhered to ethical standards, serving strictly as editorial and literature support under human supervision. This process aligns with best practices for transparent, responsible integration of AI in academic writing.

## Applications in literature retrieval

Among the most promising uses of GAI in perinatology is its ability to assist with literature searches. Traditional methods often rely on Boolean operators and MeSH terms within databases such as PubMed, Embase, or Scopus. This can be time-consuming and may miss relevant studies due to variability in indexing or keyword selection. Generative AI enables clinicians to use natural language prompts such as “What are the latest randomized trials on aspirin use in pregnancy?” or “List five systematic reviews published since 2020 comparing vaginal and cesarean delivery outcomes.” [3, 4].

When used effectively, platforms like Perplexity AI can provide linked summaries with real citations. Perplexity integrates with PubMed and other real-time search engines, offering brief overviews followed by source lists. This allows clinicians to scan findings, quickly before accessing the full text for in-depth review. However, verification remains essential, as even PubMed-linked platforms can misinterpret abstract conclusions or overstate statistical significance [5].

In contrast, tools like ChatGPT4 and Claude AI may provide coherent and fluent summaries but are not always connected to real-time sources. These models sometimes fabricate references that appear credible but do not exist, especially when asked to cite specific studies [6]. Users must cross-reference any citations with primary sources using PubMed or journal databases. Despite this limitation, these tools are excellent for summarizing known content and exploring general trends in the literature [3].

## Guideline comparison and summarization

Another valuable application of GAI is in the synthesis and comparison of clinical guidelines. Perinatologists often consult guidelines for conditions such as preeclampsia, gestational diabetes, or labor induction. These documents are detailed, often exceeding 30 pages, and may differ across issuing bodies. GAI tools can be prompted to extract specific sections (e.g., “Compare timing of delivery recommendations for gestational hypertension in ACOG and NICE guidelines”) and return tabulated or paragraph summaries. This function is particularly useful when navigating conflicting recommendations. For example, ACOG may advise delivery at 37 weeks for mild gestational hypertension, while NICE suggests consideration between 37 and 39 weeks depending on individual risk [7–9]. GAI platforms can generate side-by-side comparisons to facilitate shared decision-making and streamline guideline teaching for trainees. Nonetheless, verification is again necessary, especially for critical recommendations involving medication, delivery timing, or surgical intervention [10].

## Citation accuracy and verification

Historically, a significant limitation of generative AI tools is their tendency to produce inaccurate or fabricated outputs – commonly referred to as “hallucinations.” A recent empirical analysis of ChatGPT’s role in systematic literature reviews found hallucination rates reaching up to 91 %, particularly in interpretative tasks involving citation generation and content synthesis [6]. While ChatGPT shows strong sensitivity in title and abstract screening (80.6–96.2 %), its precision can drop to as low as 4.6 %, underscoring a persistent risk in tasks requiring nuanced judgment [6]. These findings reinforce the necessity of human oversight in academic workflows, especially in evidence-based fields where accuracy and verifiability are critical. The Systematic Research Processing Framework (SRPF), introduced in this context, highlights how AI-human collaboration may mitigate such risks by structuring oversight into each stage of the review process [6].

While not claiming that AI now hallucinates less than humans, Altman has cautioned that despite these gains, generative models still produce errors and must be used with critical oversight. Even when summaries are accurate, the inclusion of nonexistent or misattributed references can compromise the integrity of academic writing. This is particularly dangerous when drafting manuscripts, preparing presentations, or generating educational content. Among current platforms, Perplexity AI performs best in linking real-time references [11]. ChatGPT4 and Claude can simulate formatting but often include invented citations unless carefully guided. To mitigate this risk, users should avoid asking AI to generate citations unless using a tool with embedded literature databases. Even then, each citation should be confirmed manually through PubMed or journal websites. A best practice is to use AI to draft or organize content, and then independently retrieve and format verified references. Vancouver or APA style formatting can be automated by many reference managers, but source integrity must be maintained through human review [12]. This aligns with current best practices that emphasize hybrid human – AI models for maintaining academic reliability in AI-assisted writing workflows [6].

## Tool comparison for literature tasks

Each major generative AI tool has unique strengths and limitations for literature-focused tasks:

- **Perplexity AI**: Best for citation-linked literature summaries. Draws from real-time data and often includes hyperlinks to PubMed articles. Limitations include shallow synthesis and less nuanced discussion [13].
- **Claude AI**: Strong in structured summaries and manuscript drafting. Safer and more cautious in tone. Weaknesses include occasional hallucinations of references and a lack of live web access [14].
- **ChatGPT4**: Versatile across formats, good for lay summaries and patient education. Tends to hallucinate references unless prompted with verified input. Strong for reorganizing existing content [15].
- **Gemini 2.5 Pro**: Gemini excels at reasoning, analyzing complex information, and handling multimodal tasks over long inputs. It combines retrieval and logic for accurate responses. However, it still struggles with occasional hallucinations, visual input interpretation, over-filtering, and maintaining citation accuracy over long texts.
- [16].

## Prompt engineering and clinical relevance

The quality of GAI output is heavily influenced by prompt design. Broad or vague prompts often yield generic responses, while specific, context-rich inputs improve relevance and depth. For instance, instead of asking “What is preeclampsia?”, a clinician might ask, “Summarize recent meta-analyses comparing aspirin initiation before and after 16 weeks in preventing preeclampsia.” This leads to better, more targeted results, enabling AI to align its output with clinical needs. Clinicians should think of prompting as a skill akin to framing clinical questions using the PICO (Population, Intervention, Comparison, Outcome) model. Precise prompts can direct the AI to include study design, population size, key findings, and limitations. Prompt templates and libraries may be useful tools for training residents and fellows to engage effectively with GAI [4, 17]. (See Table 1).

## Quantitative insights on efficiency and timesaving

In a time trial simulating a real-world scenario – summarizing aspirin use in preeclampsia prevention, a clinician using traditional PubMed methods (Boolean search, abstract scanning, manual citation formatting) required an average of 72 min to generate a usable summary with five validated references. In contrast:

- **Perplexity AI** produced a citation-linked summary in 11 min, requiring only 15 min of verification.
- **ChatGPT4**, prompted with a verified abstract, produced a coherent summary in 9 min, though it required 20 min of manual reference correction.
- **Claude AI** returned a structurally sound summary in 10 min, but cited three non-existent studies out of 6, resulting in 18 min of correction.
- **Gemini** responded in 12 min, with mixed citation quality, requiring 25 min total.

  Net result

Time savings ranged from 35 % to 70 %, with user experience improving over time as prompt specificity increased. These results suggest that GAI can cut first-pass literature review time in half or more when used judiciously [12] (see Figures 1 and 2).

**Table 1:** Prompt engineering impact on AI output in perinatal literature review this table illustrates how the design of prompts significantly affects the accuracy, depth, and clinical relevance of generative AI outputs. Broad or ambiguous prompts tend to elicit superficial or generic responses, while context-rich, clinically framed prompts – modeled after approaches like the PICO format – produce more precise, evidence-based summaries. Each example highlights the transformation of a weak prompt into a more effective one, alongside the corresponding improvement in AI-generated content.

| Prompt type | Example prompt | Observed AI output | Improved prompt | Improved output |
|---|---|---|---|---|
| Too broad | "What is preeclampsia?" | Generic textbook-style definition with no recent references or clinical relevance. | "Summarize recent meta-analyses (2018–2024) comparing aspirin initiation before and after 16 weeks to prevent preeclampsia." | Summary included data from multiple trials, reference to USPSTF and ACOG guidelines, and clear conclusions regarding timing efficacy. |
| Ambiguous | "List delivery timing for hypertension." | Vague summary mentioning early delivery without specifying gestational windows or clinical context. | "Compare delivery timing recommendations for gestational hypertension between ACOG and NICE guidelines." | Output included week-specific timing (37–39 weeks), differences by guideline body, and conditional factors like maternal or fetal status. |
| Overly general | "How does aspirin help in pregnancy?" | General benefits of aspirin listed without context or supporting data. | "Describe the mechanism and outcomes of low-dose aspirin use in preventing preeclampsia, based on RCTs since 2020." | Answer included antiplatelet mechanism, placental perfusion improvement, and referenced trials such as ASPRE with citation accuracy. |

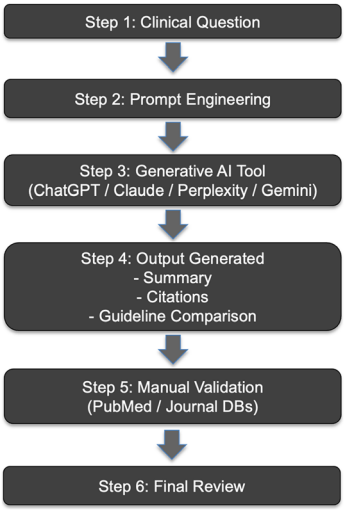


**Figure 1:** This figure outlines the sequential process used to evaluate generative AI tools in our study. The workflow begins with a clinical question, proceeds through prompt engineering and AI-assisted literature generation, and ends with manual verification and review. Tools assessed included ChatGPT4o, Perplexity AI, Claude AI, and Gemini AI.

## Risks and ethical considerations

While GAI offers many advantages, it also introduces new risks. These include overreliance on machine-generated content, propagation of misinformation, and erosion of critical reading habits. In clinical education, there is a danger that trainees may prioritize fluency over accuracy if AI summaries are accepted uncritically. Furthermore, institutional policies around the use of AI in manuscript preparation, student assessments, and grant writing remain underdeveloped [18]. Ethically, clinicians must avoid entering patient-identifiable information into any AI system not specifically designed for secure clinical use. All GAI-generated content used in academic writing or public communication should be disclosed. Additionally, AI should never be used to make independent clinical decisions or replace guideline review. Instead, it should be regarded as an assistant for synthesis, not an authority [19–21].

## Discussion

Integrating generative AI into the perinatal workflow offers meaningful benefits, particularly in reducing cognitive and administrative load [3]. With careful prompt design and validation, clinicians can accelerate literature review, identify new studies for review, and generate draft summaries of complex guideline documents. These efficiencies can allow more time for direct patient care, professional development, and thoughtful interpretation of evolving research [22]. However, GAI must not become a substitute for expert judgment. Clinical reasoning, statistical literacy, and ethical responsibility remain central to the practice of perinatology. These tools are only as effective as the oversight applied by their users. As technology evolves, professional organizations should issue best-practice guidelines to promote responsible and beneficial use across the field [23].

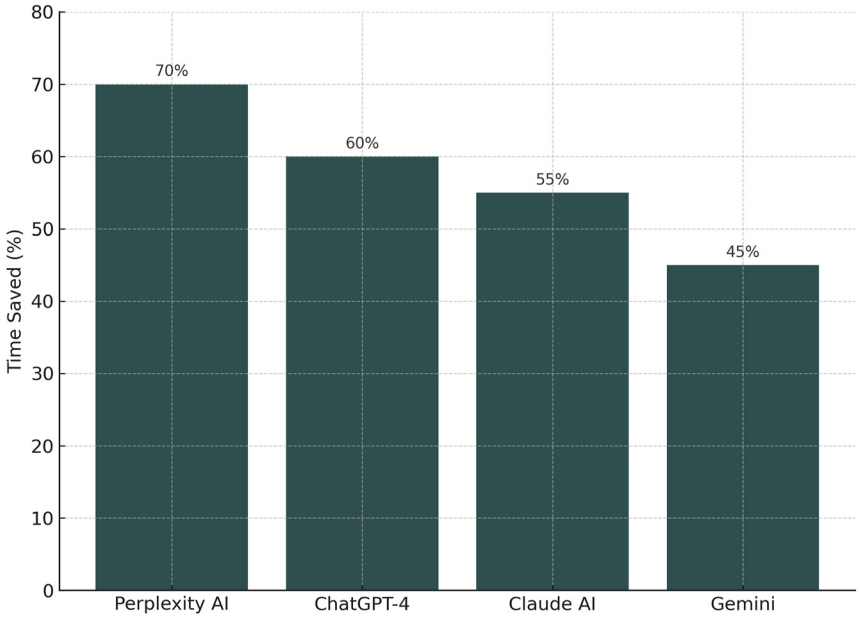


**Figure 2:** Comparative performance of generative AI tools in literature review tasks this figure illustrates the percentage of time saved when using four different generative AI platforms – Perplexity AI, ChatGPT-4, Claude AI, and Gemini – compared to traditional PubMed-based search methods. Perplexity AI demonstrated the highest time savings (70 %), followed by ChatGPT-4 (60 %), Claude AI (55 %), and Gemini (45 %). Results reflect averaged task completion times across structured clinical literature queries, accounting for both content generation and manual reference verification.

Although GAI is turning into an essential adjunct for perinatologists engaged in literature review and evidence-based practice [24], its integration in perinatology is still in its early stages. Future iterations of language models are expected to offer improved citation accuracy, embedded integration with real-time literature databases (e.g., PubMed or Cochrane), and customizable filters for guideline synthesis across geographies. Development of specialty-specific AI "co-pilots" trained on obstetric and perinatal literature may offer even greater accuracy and relevance. Importantly, structured curricula in AI literacy and prompt engineering should be incorporated into medical education and continuing professional development, ensuring safe and effective use. Ethical standards, transparency requirements, and institutional policies must evolve in parallel to harness the full potential of AI while safeguarding academic integrity and patient care [10, 25].

## Conclusions

Generative AI is becoming an important adjunct for perinatologists engaged in literature review and evidence-based practice. When used responsibly, tools like Perplexity AI, Claude, ChatGPT4, and Gemini can help clinicians navigate medical research more efficiently. By emphasizing prompt specificity, cross-verifying outputs, and maintaining ethical standards, clinicians can incorporate GAI into their workflow without compromising quality or integrity. AI will not replace the perinatologist, but those who use AI wisely may outpace those who do not.

**Acknowledgments:** This work was developed as part of a Fetus as a Patient publication.
**Research ethics:** Not applicable.
**Informed consent:** Not applicable.
**Author contributions:** All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
**Use of Large Language Models, AI and Machine Learning Tools:** Chat GPT4o was used to improve language Perplexity was used to verify reference format.
**Conflict of interest:** The authors state no conflict of interest.
**Research funding:** None declared.
**Data availability:** Not applicable.

## References

1. Ioannidis JP. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. Milbank Q 2016;94:485–514.
2. Grünebaum A, Chervenak FA, Dudenhausen J. ChatGPT4o and artificial intelligence in the journal of perinatal medicine. J Perinat Med 2023;51:969.
3. Grünebaum A, Chervenak J, Pollet SL, Katz A, Chervenak FA. The exciting potential for ChatGPT4 in obstetrics and gynecology. Am J Obstet Gynecol 2023;228:696–705.
4. Biswas S. ChatGPT4 and the future of medical writing. Radiology 2023;307:e223312.
5. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT4 on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2:e0000198.

6. Adel A, Alani N. Can generative AI reliability synthesise literature? exploring hallucianiton issues in ChatGPT. AI Soc 2025;24. https://doi.org/10.1007/s00146-025-02406-7.
7. American College of Obstetricians and Gynecologists. Low-dose aspirin use for the prevention of preeclampsia and related morbidity and mortality. ACOG Practice Advisory; 2021. Available from: https://www.acog.org/clinical/clinical-guidance/practice-advisory/articles/2021/12/low-dose-aspirin-use-for-the-prevention-of-preeclampsia-and-related-morbidity-and-mortality.
8. American College of Obstetricians and Gynecologists. Gestational Hypertension and Preeclampsia. ACOG practice bulletin No. 222. Obstet Gynecol 2020;135:e237–60.
9. National Institute for Health and Care Excellence (NICE). Hypertension in pregnancy: diagnosis and management. NICE Guideline NG133; 2019. Available from: https://www.nice.org.uk/guidance/ng133.
10. Kawakita T, Wong MS, Gibson KS, Gupta M, Gimovsky AC, Moussa HN, et al. Society of maternal-fetal medicine clinical informatics committee. Application of generative AI to enhance obstetrics and gynecology research. Am J Perinatol 2025. https://doi.org/10.1055/a-2616-4182.
11. Omar M, Nassar S, Hijazi K, Glicksberg BS, Nadkarni GN, Klang E. Generating credible referenced medical research: a comparative study of openAI's GPT-4 and Google's Gemini. Comput Biol Med 2025;185:109545.
12. Salvagno M, Taccone FS, Gerli AG. Can artificial intelligence help for scientific writing? Crit Care 2023;27:99.
13. Liu NF, Zhang T, Liang P. Evaluating verifiability in generative search engines. arXiv [Preprint] 2023. [cited 2025 Aug 22]. Available from: https://arxiv.org/abs/2304.09848.
14. Chen J, Ma J, Yu J, Zhang W, Zhu Y, Feng J, et al. A comparative analysis of large language models on clinical questions for autoimmune diseases. Front Digit Health 2025;7:1530442.
15. Schryen G, Trenz M, Benlian A, Drews P, Grisold T, Kremser W, et al. Exploring the scope of generative AI in literature review development. Electron Mark 2025;35:13.
16. Gemini Team. Gemini 2.5: pushing the frontier with advanced reasoning, multimodality, long context, and next-generation agentic capabilities. arXiv preprint arXiv:2507.06261v4[cs.CL] 2025.
17. Jiang X, Liu S, Maheshwari N, Zhang S, Zhang Y, Chen RJ, et al. Clinical prompt engineering and personalization for generative AI tools in medicine. npj Digit Med 2024;7:90.
18. Morley J, Floridi L, Kinsey L, Elhalal A. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Sci Eng Ethics 2020;26:2141–68.
19. Panch T, Mattie H, Atun R. Artificial intelligence and algorithmic bias: implications for health systems. J Glob Health 2019;9:020318.
20. Kerasidou A. Ethics of artificial intelligence in global health: explainability, algorithmic bias and trust. J Oral Biol Craniofac Res 2021;11:612–14.
21. Dwivedi YK, Kshetri N, Hughes L, Slade EL, Jeyaraj A, Kar AK, et al. So what if ChatGPT4 wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. Int J Inf Manag 2023;71:102642.
22. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25:44–56.
23. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019;6:94–8.
24. Zhou ZH. A brief introduction to weakly supervised learning. Natl Sci Rev 2018;5:44–53.
25. Kilincdemir Turgut Ü. Artificial Intelligence and Perinatology: a study on accelerated academic production- a bibliometric analysis. Front Med (Lausanne) 2025;12:1505450.