The impact of colchicine treatment on long-term complications of familial Mediterranean fever since its introduction in 1972: an evaluation of artificial intelligence performance

E. Ben-Chetrit[1], P.D. Levin[2], E. Ben-Chetrit[3]

[1]Infectious Diseases Unit, Shaare Zedek Medical Center, Jerusalem; [2]Department of Intensive Care, Shaare Zedek Medical Center, Jerusalem; [3]Rheumatology Unit, Department of Medicine, Hadassah Hebrew University Medical Center, Jerusalem, Israel.

Abstract

Objective

Familial Mediterranean fever (FMF) is a hereditary autoinflammatory disease characterised by recurrent febrile serositis and complicated in severe cases by AA amyloidosis, renal failure, infertility and adverse pregnancy outcomes. Since its introduction in 1972, colchicine has dramatically improved the prognosis of FMF.

Method

In this study, we try to explore the ability of three artificial intelligence (AI)-based medical information platforms: Open Evidence, ChatGPT-4, and ChatGPT-5.2, to summarise the impact of colchicine treatment on FMF-related complications. Each platform was presented with two identical clinically focused questions addressing AA amyloidosis complications, fertility, and pregnancy outcomes. Only the initial responses were analysed. Outputs were compared descriptively according to response length, structure, clinical content, contextualisation, citation characteristics, and overall presentation. References were manually verified, and textual similarity was assessed using TF-IDF cosine similarity analysis.

Results

All three systems correctly conveyed the major clinical benefits of colchicine in FMF. However, substantial differences were observed in depth, contextualisation and citation quality. Open Evidence generated concise and focused responses with directly accessible references but limited contextual discussion. ChatGPT-4 produced more detailed and historically contextualised summaries, whereas ChatGPT-5.2 generated the most comprehensive responses, including discussion beyond the scope of the questions. Most references corresponded to real publications, although occasional bibliographic inaccuracies were identified, particularly among ChatGPT-generated citations. Cosine similarity analysis demonstrated overlap in core medical concepts despite marked differences in phrasing and organisation.

Conclusion

AI-based medical information systems may provide rapid, clinically useful summaries of FMF and colchicine-associated outcomes. Nevertheless, human verification and expert interpretation remain essential because inaccuracies, hallucinations and citation inconsistencies persist. These tools should currently be regarded as complementary resources rather than replacements for conventional reviews or expert clinical judgment.

Key words

familial Mediterranean fever, artificial intelligence medical platforms, Open Evidence, Chat GPT.

Eli Ben-Chetrit, MD
Phillip D. Levin, MD
Eldad Ben-Chetrit, MD

Please address correspondence to:
Eli Ben-Chetrit
Infectious Diseases Unit,
Shaare Zedek Medical Center,
Shmuel (Hans) Beyth Street 12,
9103102 Jerusalem, Israel.
E-mail: elibc@szmc.org.il

Received on April 17, 2026; accepted in revised form on June 23, 2026.

© Copyright CLINICAL AND EXPERIMENTAL RHEUMATOLOGY 2026.

Competing interests: none declared.

Introduction

Familial Mediterranean fever (FMF) is a hereditary autoinflammatory disorder characterised by recurrent episodes of fever and acute serositis, including pleuritis, peritonitis, arthritis and pericarditis (1). Each attack typically lasts between 24 and 72 hours and resolves spontaneously. The frequency of attacks varies considerably, ranging from weekly episodes to only a few attacks per year.

The most serious complication of FMF is AA amyloidosis, which may lead to renal failure and dialysis dependency. Recurrent peritoneal inflammation may also result in intra-abdominal adhesions, leading to small bowel obstruction, fallopian tube obstruction, and infertility.

Since its introduction in 1972, colchicine has remained the cornerstone of FMF treatment (2, 3). Regular colchicine therapy has substantially reduced the frequency and severity of attacks and markedly decreased the incidence and progression of AA amyloidosis. In addition, colchicine reduces the risk of peritoneal adhesions, intestinal obstruction, and infertility.

Recently, we sought to review the impact of colchicine on the clinical course of FMF, particularly regarding AA amyloidosis, renal failure, dialysis requirement, intestinal obstruction, fertility, and pregnancy outcomes.

While initiating a conventional literature search through PubMed as an orientation tool, we simultaneously explored the performance of several artificial intelligence (AI)-based medical information platforms in summarising these issues: Open Evidence, ChatGPT-4, and ChatGPT-5.2. Open Evidence is a curated medical information platform that integrates AI-assisted retrieval from selected medical resources, including *The New England Journal of Medicine, JAMA*, Mayo Clinic and NCCN materials, whereas ChatGPT-4 and ChatGPT-5.2 are successive general-purpose large language models (LLMs) developed by OpenAI.

The present exploratory study aimed to compare how different AI-based systems respond to identical clinically focused questions regarding FMF complications and colchicine treatment. In particular, we sought to compare the depth, structure, citation characteristics, and overall presentation of the information generated by these systems. We did not intend to validate AI platforms against a formal systematic review methodology.

Methods

Each of the three AI platforms was presented with the same two questions. Only the initial responses generated by each platform were analysed; no follow-up prompts or refinements were used. This design was chosen to simulate a realistic single-query clinical use scenario.

Question 1. What changes have occurred in FMF since the introduction of colchicine in 1972 with respect to AA amyloidosis, renal failure, the need for dialysis, and small intestinal obstruction?

Question 2. What changes have occurred in FMF since the introduction of colchicine in 1972 with respect to fertility and pregnancy outcomes?

We selected two distinct clinical domains: renal complications related to AA amyloidosis and fertility/pregnancy outcomes, to assess whether AI-generated responses differed between a well-established FMF complication and more complex, multifactorial clinical outcomes.

The study was designed as an exploratory comparative evaluation rather than a formal validation study against a systematic review gold standard. PubMed searches were used only for general orientation and contextual familiarity with the literature; no formal systematic search strategy or predefined evidence synthesis protocol was used.

The responses generated by each platform were compared descriptively according to response length and structure, clinical content, clarity and organisation, citation presentation, journal quality of cited references (including Q1 journal status), and the inclusion of additional contextual information beyond the original question.

Medical accuracy was assessed through expert evaluation by clinicians (EBC, Eli BC) experienced in the management of FMF and amyloidosis. No

validated scoring instrument (such as DISCERN or CLEAR) or formal inter-rater agreement analysis was used; therefore, the qualitative assessment should be regarded as descriptive and exploratory only.

In addition, the references generated by the platforms were manually verified with regard to the existence of the citation, consistency of authorship, journal name, publication year, and general concordance with the cited medical statement.

We also assessed textual similarity among the outputs generated by Open Evidence, ChatGPT-4, and ChatGPT-5.2 using TF-IDF-based cosine similarity analysis (4-6).

Before analysis, all texts were converted to lowercase, punctuation was removed, and standard English stop words were excluded. The documents were vectorised using TF-IDF weighting, and cosine similarity values ranging from 0 to 1 were calculated.

The analysis was implemented in Python (version 3.10) using the scikit-learn library (version 1.3.0). ChatGPT-4 assisted in generating the Python code; however, the actual similarity calculations relied exclusively on standard deterministic scikit-learn algorithms rather than LLM-based interpretation.

Results and discussion

The complete responses generated by the three platforms are provided in the Supplementary material. All three systems conveyed the principal clinical concepts regarding colchicine-associated improvement in FMF outcomes. However, substantial differences were observed in response depth, structure, contextualisation, and citation presentation.

The Open Evidence platform provided relatively brief responses to both questions, totalling 273 words for the first and 221 words for the second. In contrast, ChatGPT-4 generated responses of 705 and 517 words, respectively, while ChatGPT-5.2 produced substantially longer outputs, comprising 1,012 and 1,013 words, respectively.

Overall, Open Evidence delivered concise answers that directly addressed the questions posed. The information was generally clear and well structured; however, authorship was not explicitly identified despite the inclusion of reference lists. Moreover, the overall quality of the cited literature appeared somewhat limited, with many references derived from lower-impact sources. Open Evidence provided 18 references, including 16 peer-reviewed journal articles and 2 FDA regulatory documents. Seven of the journal articles were published in Q1 journals.

In addition, Open Evidence did not provide broader contextualisation, synthesis, or complementary insights beyond the immediate scope of the questions. Nevertheless, its references were clearly presented and directly linked to the relevant content, facilitating immediate access and obviating the need for independent searches through external databases such as Google or PubMed.

ChatGPT-4 delivered substantially more detailed and contextualised responses. The information was structured into pre- and post-colchicine eras and cited authors who made seminal contributions to the introduction of colchicine therapy in FMF. Furthermore, ChatGPT-4 incorporated data on prognosis and additional factors influencing amyloidosis development, such as *MEFV* mutations, diagnostic delay, and treatment adherence. ChatGPT-4 generated 15 references, all from peer-reviewed journals, of which 9 were published in Q1 journals. However, the presentation of references was inconvenient, as users were required to independently locate each citation through Google or PubMed and manually compile reference lists, a time-consuming and inefficient process. ChatGPT-4 also included summaries and key messages, resulting in a total word count approximately threefold higher than that of Open Evidence.

ChatGPT-5.2 demonstrated further advancement beyond both ChatGPT-4 and OpenAI Evidence. It began with concise summary statements and, in response to the second question, differentiated fertility outcomes between males and females, both before and after the colchicine era. It further expanded beyond the original question by addressing the effects of colchicine on *in vitro* fertilisation (IVF). Additionally, it explained historical concerns regarding colchicine use during pregnancy and clarified why amniocentesis is no longer routinely recommended. Potential teratogenicity was discussed in detail and supported by accurate citations, including numerical data, percentages, and named studies. Finally, ChatGPT-5.2 produced 24 references, including contemporary treatment guidelines relevant to colchicine use, 15 of which were published in Q1 journals. As with ChatGPT-4, however, reference presentation remained suboptimal, requiring independent verification of citations.

Manual verification of references demonstrated that most citations generated by the three platforms corresponded to real publications and were generally consistent with the associated clinical claims. However, occasional inaccuracies in bibliographic details, incomplete citation formatting, and minor inconsistencies were identified, particularly among ChatGPT-generated references. These findings reinforce the importance of independent human verification of AI-generated citations before academic or clinical use.

A quantitative text comparison of responses addressing AA amyloidosis between Open Evidence and ChatGPT-4, using TF-IDF-based cosine similarity analysis, demonstrated low overall similarity (cosine score ≈ 0.19). This finding indicates that although both responses addressed the same clinical topic and reached similar conclusions, they were independently generated and differed substantially in structure, depth, and phrasing. As expected, comparison between ChatGPT-4 and ChatGPT-5.2 revealed moderate similarity (cosine score = 0.42), reflecting substantial overlap in core medical concepts while maintaining clear differences in organisation, emphasis, and level of detail rather than direct textual duplication.

For fertility- and pregnancy-related responses, the pairwise cosine similarity scores were as follows:

- Open Evidence *vs*. ChatGPT-4: 0.48
- Open Evidence *vs*. ChatGPT-5.2: 0.46
- ChatGPT-4 *vs*. ChatGPT-5.2: 0.49

These moderate similarity values suggest overlap in terminology and core concepts while still reflecting differences in structure, emphasis, and explanatory detail.

TF-IDF cosine similarity primarily evaluates lexical overlap rather than semantic accuracy, factual validity, or clinical usefulness. Therefore, these quantitative findings should be interpreted only as a supplementary illustration of textual similarity rather than a measure of clinical quality.

Large language models have been increasingly used in medicine and healthcare in recent years. As a major subclass of artificial intelligence, large language models (LLMs) hold considerable promise in medicine and its subspecialties, including rheumatology. Their diverse applications, ranging from improved diagnostic accuracy to personalised patient care, have made LLMs one of the most actively discussed developments in contemporary medicine (7-9). Several studies have demonstrated the contributions of LLMs to rheumatology practice, research, and medical education (10-11). Consequently, multiple investigations across different rheumatologic domains have evaluated the performance of various LLMs in responding to medical queries (12-14).

Today, LLMs provide rapid and convenient access to medical information that may facilitate earlier diagnosis and treatment. However, hallucinated or misleading outputs remain a major limitation, particularly in high-stakes medical contexts. Consequently, AI-generated medical information should always undergo critical human evaluation and verification.

The present study suggests that AI-based systems may provide rapid and clinically useful summaries of focused medical topics. However, caution is warranted because this may appear to be an "unbiased" method of information retrieval, whereas large language models can still reproduce existing biases. These biases may reflect the ways in which information has historically been collected, interpreted, and incorporated into their training data, thereby reinforcing conventional patterns of thinking. One possible strategy to address this issue is the use of a dedicated multi-agent system. In such a system, the review process is divided into clearly defined tasks, with each task assigned to a specialised LLM agent. This approach may help generate more balanced and critically evaluated outputs. Moreover, to achieve a comprehensive understanding of a topic, consulting multiple AI sources using the same query may be advantageous, allowing users to integrate complementary perspectives or select the responses best suited to their needs.

At present, AI information sources cannot replace conventional reviews or expert interpretation; rather, AI tools should be viewed as complementary resources assisting clinicians, educators, and researchers. Potential applications extend beyond information retrieval and include hypothesis generation, support for systematic reviews, pharmacovigilance, analysis of large clinical datasets, educational content generation, and clinical decision support.

The present study has several limitations. First, this was an exploratory descriptive comparison rather than a formal validation study against a systematic review gold standard. PubMed was used only as an orientation tool, and no structured systematic literature review involving additional databases such as EMBASE or Cochrane was performed. Second, the qualitative assessment relied largely on expert clinical judgment without the use of validated scoring instruments or formal inter-rater agreement analysis. Third, only the first response generated by each platform was evaluated. Because LLM outputs are stochastic, repeated prompting may generate different responses. Future studies should therefore assess response reproducibility and prompt sensitivity. Fourth, TF-IDF cosine similarity measures lexical overlap rather than semantic meaning, factual correctness, or clinical quality. Fifth, although reference verification was performed manually, hallucination rates were not systematically quantified using predefined criteria.

In summary, AI-based medical information platforms may provide rapid, flexible, and clinically useful summaries regarding FMF and colchicine-associated outcomes. Although substantial overlap exists in the underlying medical concepts presented by different platforms, important differences remain in depth, contextualisation, citation quality and presentation style.

These systems should be viewed as complementary tools rather than replacements for expert review articles, systematic evidence synthesis, or clinician judgment. Continued human oversight remains essential to ensure accuracy, minimise hallucinations, and critically interpret AI-generated medical information. It is possible that, with further progress and refinement, these tools may eventually replace narrative reviews and medical textbook chapters.

References

1. BEN-CHETRIT E, LEVY M: Familial Mediterranean fever. *Lancet* 1998; 351(9103): 659-64. https://doi.org/10.1016/S0140-6736(97)09408-7
2. GOLDFINGER SE: Colchicine for familial Mediterranean fever. *N Eng J Med* 1972; 287(25): 1302. https://doi.org/10.1056/nejm197212212872514
3. OZKAN E, OKUR O, EKMEKCI A *et al*.: A new approach to the treatment of periodic fever. *Med Bull Istanbul Med Fac* 1972; 5: 44-49.
4. LIN YS, JIANG JY, LEE SJ: A similarity measure for text classification and clustering. *IEEE Transactions on Knowledge and Data Engineering* 2016: 26: 1575-90. https://doi.org/10.1109/tkde.2013.19
5. ERDEN Y, TEMEL M H, BAGCIER F: Evaluating ChatGPT-4 for rheumatology patient education: a comparative analysis of readability, reliability, and similarity to the American College of Rheumatology's fact sheets. *Reumatologia* 2025; 63(5): 313-20. https://doi.org/10.5114/reum/207526
6. DEDE BT, GURSOY DE, OGUZ M, ALIANAK B, BAGCIER F: What Large Language Models offer about Familial Mediterranean Fever: An Analysis of Quality, Readability, Completeness, and Accuracy. *Mediterr J Rheumatol* 2025; 36(3): 410-16. https://doi.org/10.31138/mjr.261224.hfm
7. VENERITO V, BILGIN E, IANNONE F, KIRAZ S: AI am a rheumatologist: a practical primer to large language models for rheumatologists. *Rheumatology* (Oxford) 2023; 62(10): 3256-60. https://doi.org/10.1093/rheumatology/kead291
8. OMAR M, NAFFAA ME, GLICKSBERG BS, REUVENI H, NADKARNI GN, KLANG E: Advancing rheumatology with natural language processing: insights and prospects from a systematic review. *Rheumatol Adv Pract* 2024; 8(4): rkae120. https://doi.org/10.1093/rap/rkae120
9. VENERITO V, PUTTASWAMY D, IANNONE F, GUPTA L: Large language models and rheu-

matology: a comparative evaluation. *Lancet Rheumatol* 2023; 5(10): e574-e578. https://doi.org/10.1016/S2665-9913(23)00216-3

10. MANNSTADT I, MEHTA B: Large language models and the future of rheumatology: assessing impact and emerging opportunities. *Curr Opin Rheumatol* 2024; 36(1): 46-51. https://doi.org/10.1097/bor.0000000000000981
11. GROSSI E: The long journey of artificial intelligence in medicine: an overview. *Clin Exp Rheumatol* 2025; 43(5): 815-21. https://doi.org/10.55563/clinexprheumatol/oamfed
12. VENERITO V, GUPTA L: Large language models: rheumatologists' newest colleagues? *Nat Rev Rheumatol* 2024; 20(2): 75-6. https://doi.org/10.1038/s41584-023-01070-9
13. AYERS JW, POLIAK A, DREDZE M *et al*.: Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med* 2023; 183(6): 589-96. https://doi.org/10.1001/jamainternmed.2023.1838
14. COSKUN BN, YAGIZ B, OCAKGLU G, DALKILIC E, PEHLIVAN Y: Assessing the accuracy and completeness of artificial intelligence language models in providing information on methotrexate use. *Rheumatol Int* 2024; 44(3): 509-15. https://doi.org/10.1007/s00296-023-05473-5