Pediatric Rheumatology

RESEARCH ARTICLE Open Access

Reliability of a generative artificial intelligence tool for pediatric familial Mediterranean fever: insights from a multicentre expert survey

Saverio La Bella[1,2,3*], Marina Attanasi[1], Annamaria Porreca[4], Armando Di Ludovico[1,2], Maria Cristina Maggio[5], Romina Gallizzi[6], Francesco La Torre[7], Donato Rigante[8], Francesca Soscia[9], Francesca Ardenti Morini[9], Antonella Insalaco[10], Marco Francesco Natale[10], Francesco Chiarelli[1*], Gabriele Simonini[11], Fabrizio De Benedetti[10], Marco Gattorno[3] and Luciana Breda[1,2]

Abstract

Background Artificial intelligence (AI) has become a popular tool for clinical and research use in the medical field. The aim of this study was to evaluate the accuracy and reliability of a generative AI tool on pediatric familial Mediterranean fever (FMF).

Methods Fifteen questions repeated thrice on pediatric FMF were prompted to the popular generative AI tool Microsoft Copilot with Chat-GPT 4.0. Nine pediatric rheumatology experts rated response accuracy with a blinded mechanism using a Likert-like scale with values from 1 to 5.

Results Median values for overall responses at the initial assessment ranged from 2.00 to 5.00. During the second assessment, median values spanned from 2.00 to 4.00, while for the third assessment, they ranged from 3.00 to 4.00. Intra-rater variability showed poor to moderate agreement (intraclass correlation coefficient range: -0.151 to 0.534). A diminishing level of agreement among experts over time was documented, as highlighted by Krippendorff's alpha coefficient values, ranging from 0.136 (at the first response) to 0.132 (at the second response) to 0.089 (at the third response). Lastly, experts displayed varying levels of trust in AI pre- and post-survey.

Conclusions AI has promising implications in pediatric rheumatology, including early diagnosis and management optimization, but challenges persist due to uncertain information reliability and the lack of expert validation. Our survey revealed considerable inaccuracies and incompleteness in AI-generated responses regarding FMF, with poor intra- and extra-rater reliability. Human validation remains crucial in managing AI-generated medical information.

Keywords Artificial intelligence, AI, Pediatric rheumatology, Familial mediterranean fever, Generative artificial intelligence, FMF

*Correspondence:
Saverio La Bella
saveriolabella@outlook.it
Francesco Chiarelli
chiarelli@unich.it

Full list of author information is available at the end of the article

Background

In the medical field, artificial intelligence (AI) is garnering increasing attention as it is becoming a widely used tool for clinical activities and research. Non-generative AI applications such as supervised classification models involve analyzing and interpreting existing data. In contrast, generative AI applications and tools, such as generating patient education materials, create new original content. Generative AI incorporates regional differences and randomisation to tailor content (mostly text, images, and videos) according to language trends, cultural preferences, and locally relevant texts. Machine learning and several AI models have recently been largely used, with often good outcomes in rheumatology and pediatric rheumatology. These contemporary technologies can potentially reshape our concept of diagnosis and management, also in pediatric rheumatology disorders. Indeed, machine learning algorithms may assist physicians and other health professionals to provide more accurate and timely clinical decisions, discover novel biomarkers, and customize individual treatment plans using extensive datasets and advanced analytics [1].

Several generative AI tools are free and easily accessible, allowing patients and caregivers to interact with them for insights into medical topics. In today's fast-paced world, generative AI tools can be perceived as fast and easily accessible reservoirs of knowledge for patients and their families. Moreover, individuals with limited access to health services may consider generative AI tools as a free means to obtain health-related information while minimizing expenses. However, because generative AI tools typically draw knowledge from web sources, the information may vary in quality, sometimes presenting inaccuracies and obsolescence. Despite these premises, medical organizations have mostly not yet adopted an official position on AI, its potential benefits, and associated risks. These considerations also apply to pediatric rheumatology.

Familial Mediterranean fever (FMF) is a rare inherited disease caused by pathogenic variants in the *MEFV* gene [2]. Although traditionally prevalent in regions bordering the Mediterranean Sea, FMF has now spread globally largely due to extensive migration trends in the past and recent history [3, 4]. Given that its onset mostly occurs in individuals younger than 20 years, this disease is often recognized by pediatric rheumatologists. As the current EuroFever/Paediatric Rheumatology INternational Trials Organisation (PRINTO) classification criteria for hereditary periodic fevers consider genotype as a leading factor for FMF diagnosis, ongoing research frequently aims to establish the pathogenicity of undefined *MEFV* variants [5]. Consequently, patients may face uncertainty regarding their condition. Furthermore, while pathogenic *MEFV* variants affecting exon 10 have been associated with more severe clinical manifestations, a comprehensive genotype-phenotype correlation remains elusive [6]. Renal amyloidosis is one of the most feared complications among FMF patients, and both clinical and subclinical inflammation often require lifetime treatment with colchicine [6]. However, in cases where colchicine fails to achieve sufficient control of inflammation, therapeutic strategies may necessitate the use of interleukin (IL)-1 inhibitors as adjunctive or alternative drugs [7, 8].

Thus, patients and caregivers may utilize generative AI tools to access valuable information regarding etiopathogenesis, clinical manifestations, risk factors for a severe disease course, diagnosis, and treatment. However, generative AI tools yield varied responses to identical inquiries when interrogated, due to the randomization inherent in the data acquisition process and the mechanisms of machine learning. Consequently, patients may encounter heterogeneous information compounded by the lack of expert validation, thereby increasing the likelihood of encountering misleading or inaccurate information provided by generative AI models.

The aim of this study was to conduct an expert, blinded, multicentric survey to assess the accuracy and consistency of responses generated by a widely used generative AI tool regarding pediatric FMF.

Methods

Study design and expert selection

We utilized a prominent, free, generative AI tool, Microsoft Copilot with Chat-GPT 4.0, to gather insights on several aspects of FMF in pediatric patients. On November 24, 2023, a total of 15 distinct questions, approved by experts on pediatric FMF, were posed to the generative AI tool in English. Each question was repeated three times, resulting in a total of 45 responses (the complete list of responses is available as Supplementary Data). After each question, the generative AI tool was reset to prevent it from recalling previously provided information. The inquiries, focusing on multiple topics related to pediatric FMF, were formulated either in broad terms ("What is FMF?") or in a more specific manner ("Should colchicine be continued for life in patients with FMF?") (Table 1).

A web-based survey was distributed to targeted experts, namely pediatric rheumatologists affiliated with the Italian Society of Pediatric Rheumatology (ReumaPed) and operating within tertiary-level pediatric rheumatology centres associated with PRINTO. The survey eligible participants were required to possess a more than 10 years of clinical experience in managing pediatric patients with autoinflammatory diseases, particularly FMF.

To assess how accurate Microsoft Copilot with Chat-GPT 4.0 answers are at each session, we asked experts

Table 1 The 15 questions on pediatric familial Mediterranean fever submitted to Microsoft Copilot with Chat-GPT 4.0

N.	Questions
1	What is FMF?
2	Is FMF contagious?
3	What is the pathogenesis of FMF?
4	How widespread is FMF?
5	Can I have FMF with just one mutation?
6	Which are the main symptoms of FMF?
7	Which are the current classification criteria for FMF?
8	What should I do if I have FMF?
9	Do I need to have regular visits if I have FMF?
10	Can the onset of FMF in adulthood be possible?
11	Which are the most dangerous risks for FMF holders?
12	Which are the most important risk factors for developing a severe FMF phenotype?
13	Is FMF curable?
14	Should colchicine be continued for life in patients with FMF?
15	Which are the approved biologics for children with FMF?

Abbreviations Familial Mediterranean fever: FMF

what their judgment was in terms of accuracy of what was stated in the responses from the generative AI tool. The nine participating experts were prevented from accessing the ratings provided by their colleagues, ensuring a blind rating mechanism. Generative AI responses underwent assessment using a rating system based on a Likert-like scale ranging from 1 to 5, structured as follows:

1. Not relevant.
2. Relevant but not acceptable due to substantial inaccuracy.
3. Relevant with minor inaccuracy.
4. Relevant and accurate but incomplete.
5. Relevant, accurate and complete.

To ensure sufficient time for a comprehensive assessment and rating process, experts were requested to provide their ratings within a timeframe of up to two months. In addition, we asked experts how much confidence they had in using generative AI tools for medical purposes before and after completion of the survey (Fig. 1). No ethical approval was required due to the structure of the study.

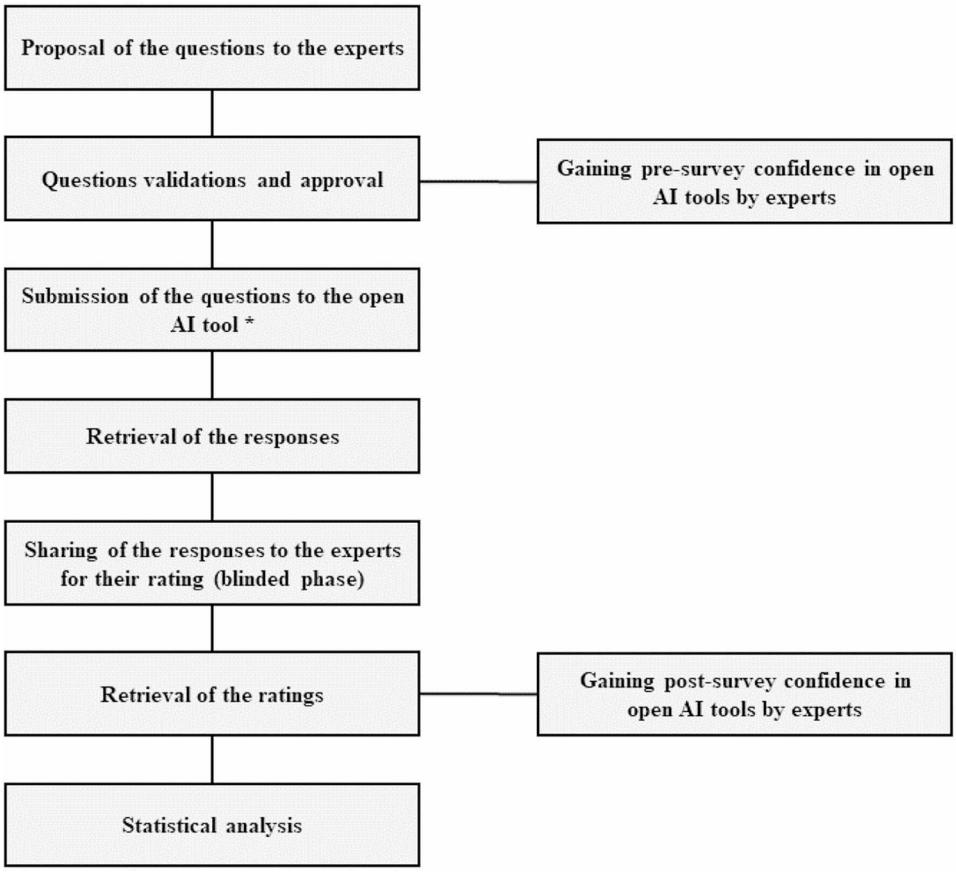

Fig. 1 Flow chart of the survey methodology. * The 15 questions were prompted each three times to the generative AI tool Microsoft Copilot with Chat-GPT 4.0

Statistical analysis

Ordinal variables were presented as median with first and third quartiles. The Friedman test evaluated the significant differences among readers over the sessions. Dunn's post hoc test, with a Bonferroni correction, was used to compute multiple pairwise comparisons. The intraclass correlation coefficient (ICC) was used to assess session agreement. An ICC < 0.5 was considered as poor, ≥ 0.5 and < 0.75 as moderate, ≥ 0.75 and < 0.9 as good and ≥ 0.9 as excellent agreement. Inter-rater agreement in graded responses was evaluated for each session using Krippendorff's alpha reliability coefficient. Alpha values closer to −1 are defined as inverse agreement, 0 is complete disagreement, and 1 is complete agreement. All tests were two-sided, and a level of statistical significance was set at $p < 0.05$. All the statistical analyses were performed using the R environment for statistical computing and graphics version 3.5.2 (R Foundation for Statistical Computing, Vienna, Austria; https://www.R-project.org/).

Results

The evaluation of the responses revealed notable insights into the expert assessments and the performance of the generative AI tool on pediatric FMF and unveiled substantial intra-rater variability as well as variability among expert judgments. This variability, manifested in the differences observed in first, second, and third assessments, underscores the complexity inherent in evaluating responses to multifaceted medical queries.

A metric for assessing intra-expert reliability was given by ICCs, which showed values ranging from low to moderate across various answers. ICCs ranged from a minimum of -0.151, indicating slight intra-rater disagreement, to a maximum of 0.534, suggesting moderate intra-rater agreement. However, only two out of nine experts exhibited statistically significant intra-rater variability. Median values for overall responses at the initial assessment ranged from 2.00 to 5.00. During the second assessment, median values spanned from 2.00 to 4.00, while for the third assessment, they ranged from 3.00 to 4.00 (Table 2).

Notably, while some questions elicited relatively higher agreement, others displayed more pronounced divergence in assessments. The findings unraveled a spectrum of variability in expert assessments, suggesting differing levels of concordance regarding the fidelity of the generative AI platform responses across the pediatric FMF inquiry spectrum. This variability stemmed from a confluence of intrinsic and extrinsic factors inherent in both expert proficiency and evaluation dynamics. Additionally, Krippendorff's alpha coefficients, highlighting the inter-expert reliability, indicated a modest level of agreement that varied marginally across assessment iterations. Indeed, the observed trend of decreasing Krippendorff's alpha coefficients across expert ratings warrants careful consideration and speculation regarding potential underlying factors. The initial value of 0.136 (at the first response), followed by subsequent declines to 0.132 (at the second response) and 0.089 (at the third response), suggests a diminishing level of agreement among experts over time. Regarding the trust of experts in generative AI tools for medical scope, four out of nine experts have increased their trust, two out of nine have decreased their trust, and three out of nine have preserved their trust (Fig. 2). However, the pre-survey levels of trust were generally low, with a slight increase in the post-survey assessment.

Discussion

The emergence of AI in the medical field represents an exciting new frontier; indeed, AI tools can provide several functionalities that could be potentially revolutionary. For example, when asked about the potential

Table 2 Median [first; third] quartiles of the likert-like scale ranging from 1 to 5 evaluations over three sessions by nine experts generated by Microsoft Copilot with Chat-GPT 4.0 search engine

Expert *N.*	First Response (*)	Second Response	Third Response	*p*-value	ICC (95%)
	N = 15	*N* = 15	*N* = 15		
1	5.00 [4.00;5.00]	3.00 [3.00;4.00]	3.00 [2.00;3.00]*	**0.004**	0.080 (-0.087 to 0.372)
2	4.00 [3.50;4.00]	4.00 [4.00;5.00]	4.00 [4.00;4.00]	0.174	0.242 (-0.062 to 0.594)
3	5.00 [4.00;5.00]	4.00 [3.00;5.00]	4.00 [2.50;4.00]	0.056	0.437 (0.136 to 0.726)
4	4.00 [3.00;4.00]	4.00 [4.00;5.00]	4.00 [3.00;4.50]	0.195	0.505 (0.204 to 0.769)
5	5.00 [3.00;5.00]	4.00 [4.00;4.50]	4.00 [4.00;5.00]	0.779	-0.137 (-0.341 to 0.220)
6	4.00 [4.00;5.00]	4.00 [3.00;4.00]	3.00 [2.50;4.00]*	**0.002**	0.270 (0.007 to 0.593)
7	3.00 [3.00;4.00]	4.00 [3.00;4.00]	3.00 [3.00;3.00]	0.063	0.269 (-0.020 to 0.606)
8	3.00 [2.00;4.00]	3.00 [2.00;4.00]	3.00 [2.00;3.50]	0.595	0.534 (0.224 to 0.789)
9	2.00 [1.50;3.00]	2.00 [1.50;4.00]	3.00 [1.50;4.00]	0.361	-0.151 (-0.352 to 0.206)

These responses correspond to 15 unique questions related to familial Mediterranean fever in children

The Intraclass Correlation Coefficient (ICC) an its' 95% Confidence Interval (CI) values are provided to repeatability over the sessions

The *p*-value derived from Friedman test

* There is a statistically significant difference using post-hoc analysis

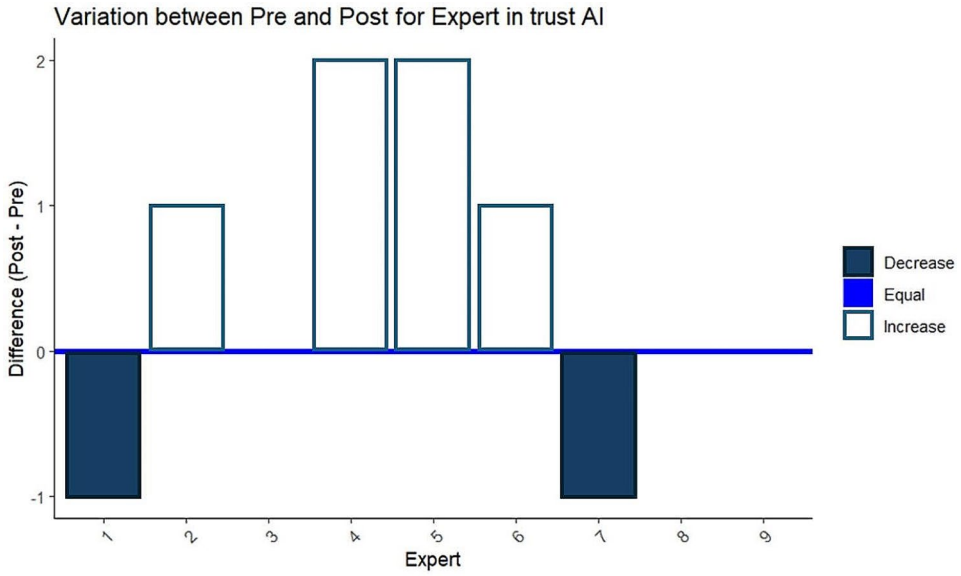

Fig. 2 Difference in expert trust in generative AI tools for medical scope before and after completing the survey

applications of AI in pediatric rheumatology, the popular generative AI tool Microsoft Copilot with Chat-GPT highlighted numerous opportunities:

1. Early diagnosis. AI can evaluate patient history and phenotype for early diagnoses and better outcomes, with minimized long-term complications.
2. Treatment optimization. AI tools can evaluate treatment responses and disease outcomes, thus helping clinicians to set individualized treatment plans and improving efficacy, minimizing side effects.
3. Predictive analytics. AI tools can analyze various factors to predict disease progression with substantial accuracy, helping clinicians to intervene timely.
4. Imaging analysis. AI tools can be used for the interpretation of imaging studies such as X-rays, ultrasound, and magnetic resonance imaging (MRI), helping clinicians to detect minor abnormalities and monitor disease course with relevant accuracy.
5. Patient education. Generative AI tools can help patients and their families to access accurate and updated information about their disease, spanning from epidemiology to pathogenesis, clinical manifestations, diagnosis, treatment options, and perspectives.

However, serious challenges arise from unregulated information source processes and the absence of expert recognized and validated tools for data interpretation. With no validated methods for managing its perils and promises, there is a considerable risk for inaccurate or outdated interpretations of medical knowledge provided by AI. These challenges should be considered by the medical community as well as it should be imperative to fully comprehend and exploit the benefits of AI in healthcare settings, similar to other contemporary digital tools [9, 10]. Pediatric rheumatology is not exempt from these considerations; nonetheless, there have been several interesting studies conducted on the topic. To the best of our knowledge, the first pioneering research on computer-assisted diagnosis of rheumatologic diseases was published in 1983, when it was proposed a knowledge-based computer consultant system christened "AI/RHEUM" [11]. This system aimed to offer diagnostic support for 26 different rheumatologic disorders, reporting a diagnostic performance of 92%, as confirmed by further research [11, 12]. Nevertheless, a lower accuracy was documented in subsequent evaluations [13]. This first AI model garnered attention from the medical community, leading to the development of some related project such as an educational package named "AI/LEARN Network", consisting of graphical representation of the knowledge encompassed by AI/RHEUM [14]. A modified version AI/RHEUM for pediatric rheumatology was proposed in 1998, with a declared diagnostic rate of 92% [15]. Also, in recent years several intriguing studies have been conducted to assess the reliability of AI models in pediatric rheumatology. In 2019, researchers developed an AI method to address the educational gap among parents and children dealing with juvenile idiopathic arthritis. This method delivered a dialogue system facilitating caregivers' access to educational material, with good outcomes [16, 17]. In 2020, AI was utilized to compare segments of pre- and post-pamidronate whole-body MRI scans in patients under 16 years old with

chronic non-bacterial osteomyelitis. A machine learning algorithm was developed, showing promising sensitivity in detecting new lesions or resolution of existing ones. However, it failed to accurately classify stable disease [18]. Intriguing research on the topic was conducted in 2023, wherein AI was employed to differentiate children with juvenile dermatomyositis from healthy controls utilizing nailfold capillaroscopy images. The neural network model named "NFC-Net" exhibited good sensitivity and specificity, demonstrating its accuracy in predicting clinical disease activity [19]. A new diagnostic model for systemic-onset juvenile idiopathic arthritis was developed in 2024, providing a novel tool for aiding in the identification of the disease. This innovative model also allowed the recognition of four key genes that could serve as potential biomarkers for the disease [20].

FMF has also been the focal point of several studies employing AI. A missense variant metapredictor tool named Rare Exome Variant Ensemble Learner (REVEL) was employed to decrease the number of *MEFV* variants with unknown significance [21, 22]. The REVEL scores aligned with the consensus classification provided by experts. In addition, this model enabled the authors to propose a reclassification of 96 *MEFV* variants [23]. The opportunities and potential of AI in supporting diagnosis and treatment for FMF patients have been emphasized by reports of patients who have benefited from reaching a diagnosis through AI [24]. Moreover, AI models have successfully highlighted the role of some common *MEFV* variants through deep neural networks and machine learning approach [25, 26]. Thanks to the development of a machine learning approach, a strong geographic association has been identified between the modern-day origins of three prevalent *MEFV* variant and a specific geographical area spanning from North Africa to Europe, and West Asia [26]. In recent years, there has been a notable increase in research interest regarding the application of AI in the field of rheumatology [27–31]. However, challenges have also arisen in AI applications in FMF. For instance, recent research employing generative AI models to differentiate between patients with FMF and those with Deficiency of Interleukin-1 Receptor Antagonist (DIRA) did not exhibit higher accuracy compared to an expert physician [32].

Despite the growing interest in understanding the possibilities of AI for medical scope (particularly for rare diseases such as FMF), there is currently no consensus from medical organizations on related benefits and risks. Our multicentric survey aimed to evaluate the accuracy of a popular generative AI tool for inquiries about pediatric FMF. Several considerations can be made about the responses provided; for example, despite the fact that responses are generally fluent, well-written, and rich in details, information is often inaccurate or incomplete (Table 3). When prompted about what FMF is, the popular generative AI tool responded that "genetic testing is not the only way to diagnose FMF [.]" (response 1 A); however, according to the current EuroFever/PRINTO classification criteria, this definition could be acceptable but it should be specified that genetic investigations should always be performed, when available, to achieve a diagnosis, with higher levels of accuracy, sensitivity, and specificity than clinical classification criteria [5]. The popular generative AI tool repeated in various responses (responses 1B, 2B, 2 C, etc.) that "people with FMF inherit a faulty copy of this gene from each parent" or "this mutation is passed from parents to their children in an autosomal recessive manner, meaning that both parents must be carriers for a child to have the disease"; nevertheless, patients with a heterozygous phenotype with likely pathogenic or pathogenic variants may be diagnosed with FMF [5]. Other inaccuracies could be found, such as "the gene mutation causes pyrin to be activated by bacterial modifications in Rho GTPases" because this molecular mechanism is also present in normal conditions (response 3 A), or "in people with FMF, the mutation in the *MEFV* gene makes pyrin less effective" because pyrin is hyperactivated in FMF (response 3B), or "FMF can also occur in people who have mutations in other genes that affect the same pathway" (response 5 C). Other significant inaccuracies detected were "therefore, it is possible to have FMF with just one mutation, but it is rare and unpredictable" (response 5 C), while there is a relevant number of FMF patients with a heterozygous *MEFV* genotype. One of the most serious inaccuracies was found in response 7 A, where the generative AI tool reported that the current EuroFever/PRINTO classification criteria for FMF are "based on a scoring system that assigns points to different clinical features, laboratory tests, and genetic tests. A patient is diagnosed with FMF if he or she has a score of 13 points or more […]". This sentence has major inaccuracy because this classification criteria require one or two (based on the *MEFV* genotype) among recurrent fever, abdominal pain, chest pain, and arthritis, and therefore are not based on a score of 13 points [5]. Also, substantial inaccuracies are present in the other responses to the question focused on the current classification criteria, such as in response 7B the lack of the current EuroFever/PRINTO classification criteria, and the wrong definition of the Yalcinkaya-Ozen classification criteria "a patient is diagnosed with FMF if they have a score of 10 or more points, based on the presence and frequency of fever, serositis, arthritis, skin rash, and *MEFV* gene mutations" while they in reality are suggestive for FMF with at least two among fever, abdominal pain, chest pain, oligoarthritis, and a positive family history. In response 7 C are also reported wrong, lower, rates of sensitivity and specificity for the EuroFever/

Table 3 The major inaccuracies found among the 45 responses provided by the generative artificial intelligence tool on pediatric familial Mediterranean fever

Response	Inaccuracies	Comment
1 A	Genetic testing can confirm the presence of a mutation in the *MEFV* gene, but it is not always conclusive, as there are many different types of mutations that can cause FMF. Some people may have a mutation but not show any symptoms, while others may have symptoms but not have a mutation. Therefore, genetic testing is not the only way to diagnose FMF, and it should be interpreted along with the clinical and family history.	Genetic investigations should always be performed, when available, to achieve a diagnosis. There are several variants recognized as pathogenic (*https://infevers.umai-montpellier.fr/web/search.php?n=1*). Moreover, a non-confirmatory genotype may lead to a diagnosis according to the EuroFever/PRINTO classification criteria. Having *MEFV* variants alone is not sufficient for a diagnosis of FMF without clinical manifestations.
2B	The diagnosis of FMF is based on the clinical criteria, which include the history of recurrent febrile attacks, the family history of the disorder, and the response to treatment. There are also genetic tests that can detect the mutation in the *MEFV* gene, but they are not always conclusive, as there are many different types of mutations that can cause FMF.	
9 A	However, genetic testing is not mandatory and does not replace the clinical diagnosis of FMF.	
1 A, 1B	Fever, which can last from a few hours to several days	Fever typically lasts up to 3 days. Longer periods are usually more suspectable for other autoinflammatory diseases or diverse etiologies (infectious, neoplasms, etc.).
1B	Skin rash, which can appear as red spots or patches on the lower legs	Skin rashes are not only observed on the lower legs.
1B	People with FMF inherit a faulty copy of this gene from each parent, making them unable to control inflammation properly.	A diagnosis of FMF can also be made with one *MEFV* pathogenic or likely pathogenic variant, resulting in a heterozygous genotype.
2B	This mutation is passed from parents to their children in an autosomal recessive manner, meaning that both parents must be carriers for a child to have the disease	
2 C	This mutation is inherited in an autosomal recessive manner, which means that both parents must be carriers of the mutated gene for a child to have the disease	
1 C	Genetic testing can confirm the presence of mutations in the *MEFV* gene, but it cannot predict how severe the symptoms will be or how often the attacks will occur	Patients with a homozygous *MEFV* genotype usually have a more severe disease course, as well as some variants affecting exon 10, such as the missense M694V variant.
2 C	A genetic test can confirm the presence of the *MEFV* mutation, but it may not detect all the possible variant	Genetic tests employed nowadays have reached good accuracy levels. Moreover, a genetic test should always be performed when available.
3B	In people with FMF, the mutation in the *MEFV* gene makes pyrin less effective, leading to increased activation of the inflammasome, a complex of proteins that triggers the release of IL-1β	Despite a not fully understood mechanism, the pyrin inflammasome is usually hyperactivated in FMF patients.
5 A	You can have FMF with just one mutation, but it is less common and less predictable than having two mutations	The real prevalence of heterozygous FMF patients is unknown, probably higher than that of homozygous patients due to underdiagnosis or milder clinical phenotypes.
5 C	This is called autosomal dominant or digenic inheritance, respectively. These cases are rare and may have milder or atypical symptoms compared to the classic form of FMF.	
7 A	These criteria (EuroFever/PRINTO classification criteria) are based on a scoring system that assigns points to different clinical features, laboratory tests, and genetic tests. A patient is diagnosed with FMF if he or she has a score of 13 points or more.	EuroFever/PRINTO classification criteria for FMF require at least one or two of the following: duration of episodes 1–3 days, arthritis, chest pain, and abdominal pain. Moreover, the EuroFever/PRINTO clinical classification criteria require at least six out of nine clinical features (both presence and absence).
7 C	According to the web search results, the current classification criteria for FMF are based on a set of clinical and genetic features that were developed by an international panel of experts in 2018. The criteria include: - The presence of at least one typical FMF symptom, such as fever, abdominal pain, chest pain, arthritis, erysipelas-like erythema, or protracted febrile myalgia. - The presence of at least one *MEFV* gene mutation that is known to be associated with FMF, or a positive family history of FMF in a first-degree relative. - The exclusion of other diseases that can mimic FMF, such as infections, malignancies, autoimmune diseases, or other autoinflammatory syndromes.	
7 C	The (EuroFever/PRINTO) criteria have a sensitivity of 93.1% and a specificity of 86.7% for the diagnosis of FMF.	Sensitivity and specificity rates are inconsistent.

Table 3 (continued)

Response	Inaccuracies	Comment
7B	A patient is diagnosed with FMF if they (Tel Hashomer) have two major criteria, or one major and two minor criteria, or five minor criteria	According to the Tel Hashomer criteria, a patient is diagnosed with FMF with at least one major criterion, or two minor criteria, or one minor criterion plus at least five supportive criteria, or one minor criterion plus at least four of the "first" five supportive criteria. Incomplete attacks are classified as minor criteria and not as major criteria. In addition, the time of 6 h to 3 days is inconsistent for incomplete attacks. Erysipelas-like erythema, skin rash, aphthous stomatitis, and chest pain relieved by sitting up or leaning forward are inconsistent with the Tel Hashomer criteria.
7 A	According to these (Tel Hashomer) criteria, a patient is diagnosed with FMF if he or she has: - Two or more major criteria; or - One major and two minor criteria; or - Five minor criteria. The major criteria are: - Typical attacks of peritonitis, pleuritis, pericarditis, or monoarticular arthritis of the hip, knee, or ankle, lasting for 1–3 days and resolving spontaneously. - Fever only, lasting for 1–3 days and resolving spontaneously. - Incomplete attacks of the abdomen, chest, or joints, lasting for 6 h to 3 days and resolving spontaneously. The minor criteria are: - Favorable response to colchicine prophylaxis. - Erysipelas-like erythema. - Recurrent or chronic aphthous stomatitis. - Chest pain relieved by sitting up or leaning forward. - Family history of FMF.	
7B	The Tel Hashomer criteria are the oldest and most widely used criteria for FMF diagnosis. They include major and minor criteria, such as fever, abdominal pain, chest pain, joint pain, skin rash, and family history of FMF	
7B	The Yalcinkaya-Ozen criteria are newer and more specific criteria for FMF diagnosis. They are based on clinical features, genetic testing, and response to colchicine treatment. A patient is diagnosed with FMF if they have a score of 10 or more points, based on the presence and frequency of fever, serositis, arthritis, skin rash, and *MEFV* gene mutations[1].	A patient is diagnosed with Yalcinkaya-Ozen criteria with at least two out of five criteria. The score of 10 or more points is inconsistent, as is the presence of skin rash and *MEFV* variants among the criteria.
9 A	According to the EULAR recommendations for the management of FMF, you should have a follow-up visit every 3 to 6 months, or more frequently if you have frequent or severe attacks, poor compliance, or adverse effects from colchicine	According to the current EULAR recommendations for the management of FMF, after colchicine has been initiated, patients should be followed closely for 3–6 months. Moreover, response, toxicity, and compliance should be monitored every 6 months. Monitoring C reactive protein, serum amyloid A protein, or both at least every 3 months is required during dose escalation in patients with active disease to determine the necessary colchicine dose. Blood tests should be performed 3 months after colchicine dose reduction.
9B	According to the EULAR recommendations for the management of familial Mediterranean fever, you should have a follow-up visit every 3–6 months to monitor your disease activity, colchicine adherence and side effects, and serum amyloid A levels	
9 C	According to the EULAR recommendations for the management of FMF, you should have a follow-up visit every 6 to 12 months, or more frequently if you have frequent or severe attacks, or if you have signs of amyloidosis	
11B	FMF holders are people who have inherited the mutated gene that causes FMF, but they may or may not develop symptoms of the disease	Clinical manifestations must be present to meet the EuroFever/PRINTO classification criteria for FMF. Having pathogenic *MEFV* variants is not sufficient for a diagnosis of FMF.
11 C	FMF holders are people who have this condition or carry the gene mutation that causes it.	
15 A	These biologic drugs are not approved specifically for FMF, but they are approved for other inflammatory diseases, such as rheumatoid arthritis, juvenile idiopathic arthritis, and cryopyrin-associated periodic syndromes. Therefore, they can be used off-label for FMF under the guidance of a specialist.	Canakinumab is approved by the FDA for FMF. Anakinra and canakinumab are approved by the EMA for FMF, and both are considered well tolerated, safe, and effective.
15B	Anakinra is not FDA-approved specifically for FMF, but it has been used off-label in some patients who do not respond to colchicine or other biologics	
15 C	Rilonacept and anakinra are not FDA-approved specifically for FMF, but they may be effective in some cases	

Abbreviations Familial Mediterranean fever, FMF; Paediatric Rheumatology INternational Trials Organisation, PRINTO; interleukin, IL; European Alliance of Associations for Rheumatology, EULAR; Food and Drug Administration, FDA; European Medicines Agency, EMA

PRINTO classification criteria. Other relevant inaccuracies are detectable in the biologic treatment available for children with FMF, as in response 15 A it was reported that "several biologic drugs that may be effective for children with FMF who do not respond to colchicine, but they are not approved for this indication" or "these biologic drugs are not approved specifically for FMF but they are approved for other inflammatory diseases, such as rheumatoid arthritis, juvenile idiopathic arthritis, and cryopyrin-associated periodic syndromes. Therefore, they can be used off-label for FMF." In addition, in response 15B, it is reported that "anakinra is not the U.S. Food and Drug Administration-approved specifically for FMF, but it has been used off-label in some patients who do not respond to colchicine or other biologics". Actually, anakinra has been proven to be effective in large studies. Moreover, the information only focused on the FDA but did not mention the European Medical Agency (EMA) or other international and national companies. Anakinra and canakinumab are, indeed, approved by the EMA for children with FMF. Several are examples of incomplete information; for example, only IL-1 is cited among the various responses to the pathogenesis, while IL-18 is never mentioned.

The variability observed among expert judgments, as reflected in the intra-rater variability and inter-expert reliability metrics, underscores the intricacies inherent in assessing the fidelity of AI-generated responses to complex medical queries. While some questions garnered relatively higher agreement among experts, others exhibited more pronounced divergence in assessments, highlighting the heterogeneous nature of response interpretation within pediatric FMF. The variability emerged in expert evaluations can be attributed to several factors such as the differences of personal viewpoints and the heterogeneous levels of trust in generative AI prior and during the survey. Interestingly, it appeared that trust in the generative AI model decreased as the survey progressed, probably due to the lower-than-expected accuracy of response. Moreover, subjective interpretations and multiplicity in response comprehension contributed to incongruences in assessments, accentuating the need for robust validation methods and ongoing refinement of evaluation protocols. The diminishing level of agreement among experts over time, as evidenced by declining inter-expert reliability metrics, highlights the dynamic nature of response evaluation and the evolving discernment among experts throughout the assessment process. There are a number of plausible explanations for this trend, such as experts' better discernment due to experience with the assessment procedure or the appearance of subtle discrepancies in interpretation as evaluations advanced. Alternatively, it may reflect the inherent challenges associated with consistently evaluating responses generated by generative AI systems, necessitating ongoing refinement of evaluation protocols to mitigate potential sources of variability and ensure robust inter-expert reliability. One of the reasons of this variability may be the subjective prism through which experts approached their assessments, each imbued with unique perspectives, clinical insights, and scholarly backgrounds. Despite the valuable functionalities and potential of AI models, the absence of validated methods for determining the reliability and accuracy of AI-provided information is still prevalent. Addressing these concerns is imperative to fully harness the potential of AI in medical settings.

In this study, questions were prompted in English. However, the interactions between languages and diverse cultures play an important role in influencing AI responses. In fact, language shapes our perceptions and interpretations, which affect the responses and outputs of AI models. Cultural specificities within the language can impact communication, resulting in varied interpretations and responses from generative AI tools. Additional research should evaluate how differences in language, culture, and societal norms influence AI responses. A deeper comprehension of these aspects will not only improve our understanding of AI outputs but will also promote more respectful interactions with users, especially patients with different cultures and provenances. Some methodological limitations are needed to be discussed. Firstly, although a widely used generative AI tool was utilized for this survey, there are various other generative AI platforms that could yield similar health information. A broader and more complete survey investigating the possible disparities between various generative AI tools on pediatric rheumatology-related topics, particularly FMF, could be of interest. Secondly, the analysis was limited to 45 responses resulting from 15 questions reiterated three times; therefore, the efficacy of the randomization strategy could be limited, and a higher number of responses could be required for a more in-depth evaluation. Additionally, experts were selected exclusively from third-level pediatric rheumatology Italian centers, potentially limiting the diversity of perspectives and generalizability of our findings. An international survey could yield differing ratings owing to the varied viewpoints of experts across different countries. Despite meticulous selection and validation of the study protocol, questions, responses, and the Likert-like rating scale among experts, ratings are subject to personal interpretation and judgment. Consequently, the overall reproducibility of the experiment could be partially reduced. Moreover, the authors' personal confidence in generative AI tools may have influenced their assessments; however, their opinion has been blinded by others and collected before and after the survey assessment, resulting in the fact that

after completing the survey, experts trust has generally increased or been maintained (seven out of nine experts).

Conclusions

Despite generative AI showed promising applications, there are still several doubts on information reliability and inaccuracy. In order to optimize AI utilization in the medical field, further research should aim to address these contemporary digital challenges. Medical organizations should increase their efforts in developing dedicated protocols to validate AI uses in pediatric rheumatology. The importance of human oversight of AI-generated information is crucial as well as the need for integrating new digital tools with the solid clinical experience.

Abbreviations

AI	Artificial intelligence
FMF	Familial Mediterranean fever
PRINTO	Paediatric Rheumatology International Trials Organisation
IL	Interleukin
ReumaPed	Italian Society of Pediatric Rheumatology
ICC	Intraclass correlation coefficient
REVEL	Rare Exome Variant Ensemble Learner
DIRA	Deficiency of Interleukin-1 Receptor Antagonist
EMA	European Medical Agency

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12969-024-01011-0.

Supplementary Material 1

Author contributions

SLB and ADL had the idea for the article, performed the literature search, wrote the manuscript. AP provided the statistical analysis and realized tables and figures. MCM, RG, FLT, AI, MFN, FS, GS, DR and LB rated the responses provided by AI. MA, FC, FDB, FAM and MG coordinated, supervised, and approved the final version of the manuscript.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article. The authors have no relevant financial or non-financial interests to disclose.

Data availability

Data collected during the realization of this paper are available upon reasonable request to the authors.

Declarations

Ethics approval and consent to participate

Ethics approval and consent were not required because of the structure of the study and the absence of patients involved.

Consent for publication

Not available.

Conflict of interest

All authors report no conflicts of interest.

Author details

[1]Department of Pediatrics, "G. D'Annunzio" University of Chieti-Pescara, Chieti, Italy
[2]Division of Pediatric Rheumatology, "G. D'Annunzio" University of Chieti-Pescara, Chieti, Italy
[3]Division of Rheumatology and Autoinflammatory Diseases, IRCCS Istituto Giannina Gaslini, Genova, Italy
[4]Laboratory of Biostatistics, Department of Medical, Oral and Biotechnological Sciences, "G. D'Annunzio" University of Chieti-Pescara, Chieti, Italy
[5]University Department PROMISE "G. D'Alessandro", University of Palermo, Palermo, Italy
[6]Department of Medical of Health Sciences, Magna Graecia University, Catanzaro, Italy
[7]Department of Pediatrics, Giovanni XXIII Pediatric Hospital, University of Bari, Bari, Italy
[8]Department of Life Sciences and Public Health, Fondazione Policlinico Universitario A. Gemelli, Rome and Università Cattolica Sacro Cuore, Rome, Italy
[9]Department of Pediatrics, Sant' Eugenio Hospital, Rome, Italy
[10]Division of Rheumatology, Bambino Gesù Children's Hospital, Scientific Institute for Research and Health Care, Rome, Italy
[11]Rheumatology Unit, IRCCS Meyer Children's Hospital, Florence, Italy

Received: 3 June 2024 / Accepted: 29 July 2024
Published online: 23 August 2024

References

1. Sadeghi P, Karimi H, Lavafian A, Rashedi R, Samieefar N, Shafiekhani S et al. Machine learning and artificial intelligence within pediatric autoimmune diseases: applications, challenges, future perspective. Expert Rev Clin Immunol. 2024;1–18.
2. Schnappauf O, Chae JJ, Kastner DL, Aksentijevich I. The pyrin inflammasome in Health and Disease. Front Immunol. 2019;10:1745.
3. Ben-Chetrit E, Touitou I. Familial Mediterranean Fever in the World. Arthritis Rheum. 2009;61:1447–53.
4. La Bella S, Di Ludovico A, Di Donato G, Basaran O, Ozen S, Gattorno M, et al. The pyrin inflammasome, a leading actor in pediatric autoinflammatory diseases. Front Immunol. 2023;14:1341680.
5. Gattorno M, Hofer M, Federici S, Vanoni F, Bovis F, Aksentijevich I, et al. Classification criteria for autoinflammatory recurrent fevers. Ann Rheum Dis. 2019;78:1025–32.
6. La Bella S, Di Ludovico A, Di Donato G, Scorrano G, Chiarelli F, Vivarelli M, et al. Renal involvement in monogenic autoinflammatory diseases: a narrative review. Nephrol (Carlton). 2023;28:363–71.
7. Ozen S, Demirkaya E, Erer B, Livneh A, Ben-Chetrit E, Giancane G, et al. EULAR recommendations for the management of familial Mediterranean fever. Ann Rheum Dis. 2016;75:644–51.
8. De Benedetti F, Gattorno M, Anton J, Ben-Chetrit E, Frenkel J, Hoffman HM, et al. Canakinumab for the treatment of Autoinflammatory recurrent fever syndromes. N Engl J Med. 2018;378:1908–19.
9. La Bella S, Di Ludovico A, Mainieri F, Lauriola F, Silvestrini L, Ciarelli F et al. Quality and characteristics of Pediatric Rheumatology Content on Social Media: toward a new era of education for patients and caregivers? J Rheumatol. 2024;jrheum.2024-0039.
10. La Bella S, Breda L, Ravelli A. Gallia est omnis divisa in partes tres: Social Media Platforms as a New Educational Channel for Pediatric Rheumatology. J Rheumatol. 2024;jrheum.2024–0408.
11. Kingsland LC, Lindberg DA, Sharp GC. AI/RHEUM. A consultant system for rheumatology. J Med Syst. 1983;7:221–7.
12. Porter JF, Kingsland LC, Lindberg DA, Shah I, Benge JM, Hazelwood SE, et al. The AI/RHEUM knowledge-based computer consultant system in rheumatology. Performance in the diagnosis of 59 connective tissue disease patients from Japan. Arthritis Rheum. 1988;31:219–26.
13. Bernelot Moens HJ. Validation of the AI/RHEUM knowledge base with data from consecutive rheumatological outpatients. Methods Inf Med. 1992;31:175–81.
14. Lee AS, Cutts JH, Sharp GC, Mitchell JA. AI/LEARN network. The use of computer-generated graphics to augment the educational utility of a knowledge-based diagnostic system (AI/RHEUM). J Med Syst. 1987;11:349–58.

15. Athreya BH, Cheh ML, Kingsland LC. Computer-assisted diagnosis of pediatric rheumatic diseases. Pediatrics. 1998;102:E48.
16. Rose-Davis B, Van Woensel W, Stringer E, Abidi S, Abidi SSR. Using an Artificial Intelligence-based argument theory to Generate Automated Patient Education Dialogues for Families of Children with juvenile idiopathic arthritis. Stud Health Technol Inf. 2019;264:1337–41.
17. Rose-Davis B, Van Woensel W, Raza Abidi S, Stringer E, Sibte Raza Abidi S. Semantic knowledge modeling and evaluation of argument theory to develop dialogue based patient education systems for chronic disease self-management. Int J Med Inf. 2022;160:104693.
18. Bhat CS, Chopra M, Andronikou S, Paul S, Wener-Fligner Z, Merkoulovitch A, et al. Artificial intelligence for interpretation of segments of whole body MRI in CNO: pilot study comparing radiologists versus machine learning algorithm. Pediatr Rheumatol Online J. 2020;18:47.
19. Kassani PH, Ehwerhemuepha L, Martin-King C, Kassab R, Gibbs E, Morgan G, et al. Artificial intelligence for nailfold capillaroscopy analyses - a proof of concept application in juvenile dermatomyositis. Pediatr Res. 2024;95(4):981–7.
20. Ding P, Du Y, Jiang X, Chen H, Huang L. Establishment and analysis of a novel diagnostic model for systemic juvenile idiopathic arthritis based on machine learning. Pediatr Rheumatol Online J. 2024;22:18.
21. Bentham J, Cesare MD, Bilano V, Bixby H, Zhou B, Stevens GA, et al. Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128·9 million children, adolescents, and adults. Lancet. 2017;390:2627–42.
22. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: an Ensemble Method for Predicting the pathogenicity of rare missense variants. Am J Hum Genet. 2016;99:877–85.
23. Accetturo M, D'Uggento AM, Portincasa P, Stella A. Improvement of MEFV gene variants classification to aid treatment decision making in familial Mediterranean fever. Rheumatology (Oxford). 2020;59:754–61.
24. Hirsch MC, Ronicke S, Krusche M, Wagner AD. Rare diseases 2030: how augmented AI will support diagnosis and treatment of rare diseases in the future. Ann Rheum Dis. 2020;79:740–3.
25. Isildak U, Stella A, Fumagalli M. Distinguishing between recent balancing selection and incomplete sweep using deep neural networks. Mol Ecol Resour. 2021;21:2706–18.
26. Adato O, Brenner R, Levy A, Shinar Y, Shemer A, Dvir S, et al. Determining the origin of different variants associated with familial mediterranean fever by machine-learning. Sci Rep. 2022;12:15206.
27. Chinnadurai S, Mahadevan S, Navaneethakrishnan B, Mamadapur M. Decoding applications of Artificial Intelligence in Rheumatology. Cureus. 2023;15(9):e46164.
28. Hügle M, Omoumi P, van Laar JM, Boedecker J, Hügle T. Applied machine learning and artificial intelligence in rheumatology. Rheumatol Adv Pract. 2020;4(1):rkaa005.
29. Kothari S, Gionfrida L, Bharath AA, Abraham S. Artificial Intelligence (AI) and rheumatology: a potential partnership. Rheumatology (Oxford). 2019;58:1894–5.
30. Stoel B. Use of artificial intelligence in imaging in rheumatology - current status and future perspectives. RMD Open. 2020;6:e001063.
31. Adams LC, Bressem KK, Ziegeler K, Vahldiek JL, Poddubnyy D. Artificial intelligence to analyze magnetic resonance imaging in rheumatology. Joint Bone Spine. 2024;91:105651.
32. Pillai J, Pillai K. Accuracy of generative artificial intelligence models in differential diagnoses of familial Mediterranean fever and deficiency of Interleukin-1 receptor antagonist. J Transl Autoimmun. 2023;7:100213.

Publisher's Note