DOI: 10.24938/kutfd.1708478 Kırıkkale Üniversitesi Tıp Fakültesi Dergisi 2025;27(2):211-216
Original Article *Özgün Araştırma*

CHAT GPT VS. RHEUMATOLOGISTS: DO WE STILL NEED THE CLINICIAN?

ChatGPT ile Romatologların Karşılaştırılması: Hala Klinisyene İhtiyaç Var mı?

Çisem YILDIZ[1] Batuhan KÜÇÜKALİ[1] Nuran BELDER[1] Merve KUTLAR[1] Nihal KARAÇAYIR[1] Pelin ESMERAY ŞENOL[2] Deniz GEZGİN YILDIRIM[1] Sevcan A. BAKKALOĞLU[1]

[1] *Division of Pediatric Rheumatology, Faculty of Medicine, Gazi University, ANKARA, TÜRKİYE*
[2] *Division of Pediatric Rheumatology, Mersin City Hospital, MERSİN, TÜRKİYE*

ABSTRACT

Objectives: Artificial intelligence (AI) encompasses systems designed to perform tasks that require human cognitive abilities, such as reasoning, decision-making, and problem-solving. Open AI's Generative Pre-Trained Transformer (GPT) model family, including ChatGPT, is widely recognized for its ability to generate human-like text and facilitate interactive discussions. ChatGPT has potential applications in diagnosis assistance and medical education in healthcare, yet its adoption raises concerns. Our study aims to evaluate ChatGPT's diagnostic performance in identifying autoinflammatory diseases compared to clinicians, exploring its potential as an accessible tool for physicians and patients.

Material and Methods: We evaluated the diagnostic performance of a publicly accessible AI model against two clinicians for identifying familial Mediterranean fever (FMF) and periodic fever, aphthous stomatitis, pharyngitis, and adenitis syndrome (PFAPA). Clinical data from 50 patients were presented anonymously in structured format to both the AI model and the clinicians. Diagnoses were compared to confirmed clinical diagnoses.

Results: A total of 50 patients were included in the study. The AI model suggested a rheumatologic diagnosis in 94% of cases but correctly diagnosed only 50% of them. In comparison, clinicians made accurate diagnoses in 76% and 70% of cases, respectively.

Conclusion: The development of AI has attracted significant attention in healthcare, as it has in other fields. However, AI-generated data may be incorrect, highlighting the importance of expert supervision. AI should complement, not replace physicians, enhancing their capabilities. Future research should evaluate AI performance across different fields and its impact on decision-making to ensure reliable use through standardized guidelines.

Keywords: *Artificial intelligence, autoinflammatory diseases, rheumatology.*

ÖZ

Amaç: Yapay zeka (YZ), insanın bilişsel yeteneklerini gerektiren görevleri yerine getirmek üzere tasarlanmış sistemleri ifade eder; bu görevler arasında akıl yürütme, karar verme ve problem çözme yer alır. OpenAI'nın Generatif Önceden Eğitilmiş Dönüştürücü (GPT) model ailesi, ChatGPT dahil, insan benzeri metin üretme ve etkileşimli tartışmalar yapabilme yeteneği ile geniş çapta tanınmaktadır. ChatGPT, tanı desteği ve tıbbi eğitimde sağlık alanında potansiyel uygulamalara sahipken, bu teknolojinin benimsenmesi bazı endişeleri de beraberinde getirmektedir. Bu çalışmanın amacı, ChatGPT'nin, otoinflamatuar hastalıkları tanımlama konusundaki tanısal performansını, klinisyenlerle karşılaştırarak değerlendirmek ve bunu hekimler ve hastalar için erişilebilir bir araç olarak incelemektir.

Gereç ve Yöntemler: Aşağıda belirtilen hastalıkların tanısını koymada bir yapay zekâ modelinin, iki klinisyenle karşılaştırılan tanısal performansı değerlendirilmiştir: Ailevi Akdeniz ateşi (AAA) ve periyodik ateş, aftöz stomatit, farenjit ve adenit sendromu (PFAPA). 50 hastanın klinik verileri anonim olarak yapılandırılmış bir formatta hem yapay zekâ modeline hem de klinisyenlere sunulmuştur. Tanılar, doğrulanmış klinik tanılarla karşılaştırılmıştır.

Bulgular: Çalışmaya toplam 50 hasta dahil edilmiştir. Yapay zeka modeli, vakaların %94'ünde romatolojik bir tanı önermiş, ancak bunların yalnızca %50'sini doğru bir şekilde teşhis etmiştir. Buna karşılık, klinisyenler sırasıyla %76 ve %70 oranında doğru tanı koymuştur.

Sonuç: Yapay zeka teknolojisinin gelişimi, sağlık hizmetleri dahil olmak üzere birçok alanda büyük ilgi uyandırmıştır. Ancak, yapay zeka ile üretilen veriler hatalı olabilir, bu da uzman denetiminin önemini vurgulamaktadır. Yapay zeka, hekimleri ikame etmek yerine tamamlayıcı bir araç olarak kullanılmalı ve hekimlerin yeteneklerini artırmalıdır. Gelecekteki araştırmalar, yapay zekânın farklı alanlardaki performansını ve karar verme süreçlerine etkisini değerlendirerek, standartlaştırılmış kılavuzlarla güvenilir kullanımını sağlamayı hedeflemelidir.

Anahtar Kelimeler: *Yapay zeka, otoinflamatuar hastalıklar, romatoloji.*

Correspondence / Yazışma Adresi: **Dr. Çisem YILDIZ**
Division of Pediatric Rheumatology, Faculty of Medicine, Gazi University, ANKARA, TÜRKİYE
Phone / Tel: **E-mail** / E-posta: drcisemyildiz@gmail.com
Received / Geliş Tarihi: 29.05.2025 **Accepted** / Kabul Tarihi: 13.07.2025

KÜ Tıp Fak Derg 2025;27(2):211-216
Doi: 10.24938/kutfd.1708478

INTRODUCTION

In its broadest sense, artificial intelligence (AI) refers to machines or computers designed to carry out tasks that typically require human intelligence.[1] These tasks include comprehension, perception, problem-solving skills, and judgment.[1,2] Using neural network algorithms trained on large datasets, AI can generate human-like text outputs and provide extensive information on various topics[3]. Open AI's Generative Pre-Trained Transformer (GPT) model family includes ChatGPT, which was released in November 2022. It is considered as one of the most advanced language models publicly available.[1,4,5] ChatGPT is a notable example in healthcare, recognized for its ability to produce text resembling human-like communication.[6]

Due to an extensive database, ChatGPT can generate reasonable and contextually appropriate responses to a wide range of questions and engage in interactive discussions, demonstrating and understanding of the complexities of human language.[1,4,7,8] Numerous applications of ChatGPT in medical practice and education have been proposed, yet its adoption has yielded mixed outcomes. [2,4,7,9] Preliminary studies suggest that AI language models can be beneficial when their limitations are well understood.[4] Although there is a significant interest in its potential to assist with diagnosis, and interpret medical reports, this also raises several concerns.[10-14]

A common limitation of these models is their susceptibility to generating incorrect information and fabricated outputs not grounded in actual training data.[3] Additional concerns include the potential confidentiality, and misuse.[4,15] Notably, both healthcare professionals and patients, as well as their families, can utilize ChatGPT and similar models for addressing health-related queries.[3,16] Although the use of the internet to search for health information is already widespread, large language models (LLMs) may become a more prominent source of information provided by such software.[3,16] Generative AI tools, due to the randomness inherent in their data collection processes and machine learning mechanisms, may produce varying responses to identical queries.[16] There remains a significant gap in studies evaluating the performance of ChatGPT in addressing medical questions.[6] This study aims to assess the diagnostic capabilities of ChatGPT in identifying autoinflammatory diseases, comparing its performance with that of clinicians, to provide insights into its utility as an easily accessible AI model for patients and healthcare providers.

MATERIALS AND METHODS

Study Design

We aimed to evaluate the performance of an artificial intelligence model, freely accessible to both parents and clinicians, in generating a list of potential diagnoses. The clinical characteristics of patients with two periodic fever syndromes, familial Mediterranean fever (FMF) and periodic fever, aphthous stomatitis, pharyngitis, and adenitis syndrome (PFAPA), were standardized into a structured format, and both the AI model and two rheumatology fellows were asked to provide the best possible diagnosis for these cases. The results were subsequently compared.

Study Description

A total of 50 patients (20 with FMF and 30 with PFAPA, diagnosed by international criteria were included in the study.[17] These patients had been under follow-up at our outpatient clinic for at least one year, with confirmed diagnoses and favourable treatment responses. The researchers documented the patients' histories in a standardized paragraph format (ÇY, BK) (Figure 1). Each paragraph included the patients' age, frequency of symptoms, presence of periodicity, the number of attacks in the past six months, and the characteristics of the episodes (cryptic tonsillitis, fever, cervical lymphadenopathy, joint pain, joint swelling, aphthous ulcers, rash, and diarrhea) (Figure1). The standardized paragraph, ensuring the confidentiality of the patient's identifying information, was presented to two pediatric rheumatology fellows (Clinician-I, Clinician-II) and ChatGPT, who were asked to specify the most likely diagnosis.

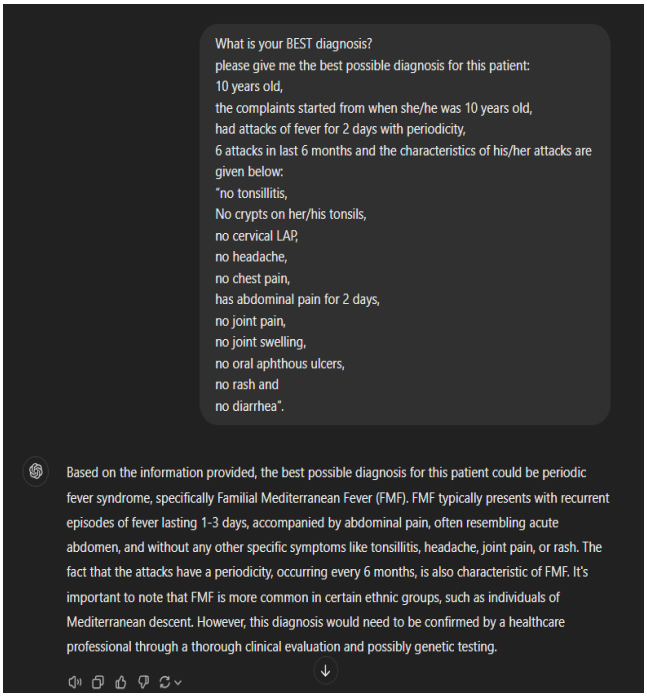

Figure 1: An example of a question format posed to ChatGPT

KÜ Tıp Fak Derg 2025;27(2):211-216
Doi: 10.24938/kutfd.1708478

Target Population and Sampling

The standardized patient histories were compiled into a single paragraph and presented to the ChatGPT-3.5 model and two rheumatology fellows. The diagnoses provided by these sources were then recorded in a database for comparison with the patients' confirmed diagnoses.

Statistical Analysis

Statistical analyses were performed using IBM SPSS software version 23. Continuous variables were expressed as mean ± standard deviation (SD) or median (interquartile range), depending on their distribution. Categorical variables were summarized as frequencies and percentages. Comparisons of baseline characteristics between groups were conducted using the Chi-square test for categorical variables. A p-value of <0.05 was considered statistically significant.

This study was conducted in accordance with the Helsinki Declaration. Ethics Committee approval was obtained from Gazi University (Decision number: E-77082166-604.01-905676 / 2024-389 dated 27.02.2024). Permission was granted to collect anonymized data without individualized consent, as the study exclusively utilized previously collected data.

RESULTS

A total of 50 patients were included in the study, with 60% (n=30) diagnosed with PFAPA and 40% (n=20) diagnosed with FMF. The mean age at symptom onset was 3.44±0.52 years, while the average age at diagnosis was 4.93±0.52. When compared between groups, PFAPA patients tend to have an earlier age at symptom onset and diagnosis, with symptoms arising at a mean age of 2.4±0.45 years, compared to FMF patients, whose symptoms typically begin at a mean age of 5.0±1.03 years ($p<0.05$). Additionally, PFAPA patients experience significantly longer febrile episodes (mean of 4.37 days) compared to FMF patients (mean of 1.9 days) ($p<0.001$). While the number of episodes in the past 6 months was similar in both groups, there were notable differences in accompanying symptoms. Tonsillitis was found in all PFAPA patients ($p<0.001$), but only 2 FMF patients; among these, one experienced 2 tonsillitis attacks and the other had 1 attack in the past 6 months. Meanwhile, abdominal pain, joint pain, and joint swelling were more frequent in FMF patients ($p<0.01$, $p<0.05$, $p<0.05$, respectively). Specifically, joint swelling was exclusive to FMF. Headaches were also more common in FMF patients ($p<0.05$). In contrast, oral aphthosis and rash were more frequently observed in PFAPA patients ($p<0.01$, $p<0.05$, respectively), with 17 PFAPA patients experiencing oral ulcers compared to just 3 FMF patients, and a rash occurring in 7 PFAPA patients but none in the FMF group. Although diarrhea was more common in FMF patients, the difference was not statistically significant ($p=0.170$). These findings suggest distinct clinical features that can aid in differentiating FMF from PFAPA, with tonsillitis, oral aphthosis, and rash being more indicative of PFAPA, while abdominal pain, joint symptoms, and shorter febrile episodes are more characteristic of FMF. The characteristics of the patients are presented in Table 1.

When asked to provide the most likely diagnosis based on the clinical history, the AI model suggested a rheumatologic diagnosis in 94% of cases (n=47). However, the correct diagnosis was made in only 50% of these cases (n=25). In comparison, clinician-1 accurately predicted the diagnosis in 76% of cases (n=38), while clinician-2 did so in 70% of cases (n=35). The final diagnoses of the patients for whom the AI model made incorrect diagnoses are presented in Figure 2. The diagnostic accuracy of clinicians and ChatGPT revealed that clinicians achieve higher accuracy in diagnosing FMF, whereas ChatGPT demonstrates comparable performance to clinicians in diagnosing PFAPA (Figure 3).

DISCUSSION

In our study, we aimed to evaluate the diagnostic accuracy of ChatGPT-3.5, an AI model that can be accessed free of charge by patients, caregivers, and healthcare professionals through internet access, in diagnosing autoinflammatory diseases. To the best of our knowledge, this is the first study comparing the diagnostic capabilities of clinicians and an AI model in the context of autoinflammatory disease. In this study, cases were presented to both the AI model and clinicians in a standard model, and their diagnostic accuracy rates were compared. The included patients had definitive diagnoses that fully met internationally accepted diagnostic and classification criteria, ensuring no uncertainty regarding their diagnoses.[17] The AI model correctly identified the diagnosis in only 50% of the cases, whereas clinicians achieved accurate diagnoses in approximately 75% of cases. However, a noteworthy finding was that the AI model recommended referral to a rheumatology centre for 94% of the patients.

When evaluating the clinical characteristics of the patients, it was observed that the longer febrile attack durations in PFAPA patients and the presence of abdominal pain, joint symptoms, and relatively shorter febrile attack durations in FMF patients were consistent with findings reported in the literature.[17,18] Although rash is not a classical feature of PFAPA, 7 patients in our cohort experienced it during some febrile episodes. All met internationally accepted diagnostic criteria and responded well to standard treatment strategies. Infectious causes were excluded through clinical and laboratory evaluations. While a comprehensive genetic

KÜ Tıp Fak Derg 2025;27(2):211-216
Doi: 10.24938/kutfd.1708478

panel was not performed, *MEFV* mutation analysis was available for all patients, and no additional clinical findings suggested monogenic autoinflammatory syndromes. Additionally, in the two FMF patients who experienced tonsillitis, clinical review revealed that the febrile episodes with pharyngeal symptoms were most likely triggered by intercurrent infections rather than reflecting a true overlap with PFAPA. These episodes were infrequent, not periodic in nature, and did not recur in a stereotyped pattern, which supports the conclusion that these findings were not suggestive of a coexisting PFAPA diagnosis. FMF and PFAPA are the most common autoinflammatory diseases and share many overlapping features.[17,18] The high prevalence of these diseases and the relatively well-established diagnostic and treatment algorithms were key features in patient selection for this study.

Table 1: Patient characteristics

		FMF (n=20)	PFAPA (n=30)	p-value
Age at symptom onset (mean±SD)		5.0±1.03	2.40±0.45	**p=0.015**
Age at diagnosis(mean±SD)		6.55±0.97	3.85±0.50	**p=0.006**
Clinical characteristics				
Number of febrile days (mean±SD)		1.9±0.143	4.37±0.242	**p<0.001**
Number of episodes in the past 6 months *(before treatment)* (mean±SD)		5.5±0.702	6.6±0.681	p=0.297
Accompanying symptoms related to the febrile episodes				
Tonsillitis	Yes	2	30	**p<0.001**
	No	18	0	
Cervical lymphadenopathy	Yes	11	16	p=0.569
	No	9	14	
Headache	Yes	13	10	**p=0.028**
	No	7	20	
Chest pain	Yes	2	2	p=0.528
	No	18	28	
Abdominal pain	Yes	15	10	**p=0.004**
	No	5	20	
Joint pain	Yes	12	9	**p=0.035**
	No	8	21	
Joint swelling	Yes	4	0	**p=0.021**
	No	16	30	
Oral aphthosis	Yes	3	17	**p=0.003**
	No	17	13	
Rash	Yes	0	7	**p=0.02**
	No	20	23	
Diarrhea	Yes	3	1	p=0.170
	No	17	29	

SD: Standart deviation

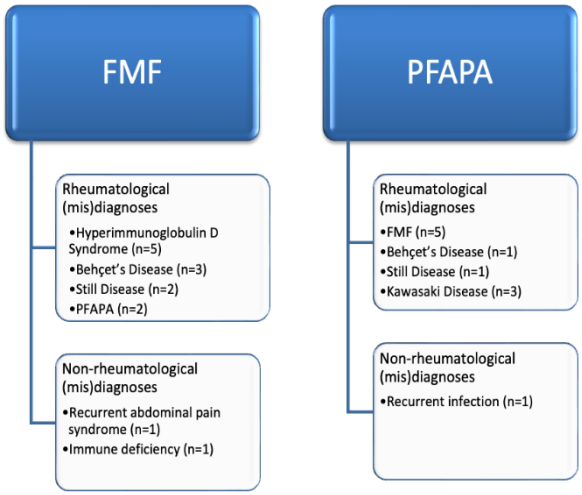

Figure 2: Confirmed diagnoses of patients misdiagnosed by Chat-GPT3.5

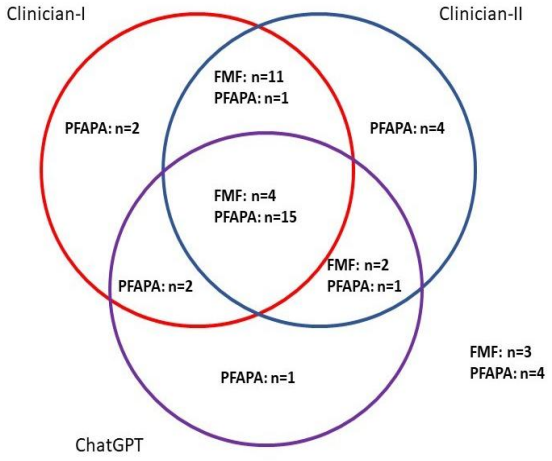

Figure 3: Distribution of diagnostic accuracy of ChatGPT and clinicians

KÜ Tıp Fak Derg 2025;27(2):211-216
Doi: 10.24938/kutfd.1708478

Artificial intelligence models have increasingly been utilized across various fields of medicine. However, the effectiveness of ChatGPT in providing accurate advice on rheumatic diseases and related queries remains uncertain.[19] In Spain, ChatGPT-3.5 correctly answered 63% of rheumatology-related questions from the national residency entrance exam, while ChatGPT-4 achieved an accuracy rate of 93%.[20] Furthermore, ChatGPT-4 has been reported to outperform rheumatologists in diagnostic accuracy for inflammatory rheumatic diseases based on medical history.[21] Additionally, ChatGPT-4 has demonstrated the ability to provide faster, higher quality, and even more empathetic responses to frequently asked questions from patients with systemic lupus erythematosus (SLE).[22] In their cross-sectional study, Ayers et al. found that responses provided by a chatbot to medical questions posed by the public on a social media platform were preferred over those given by doctors and were rated significantly higher in quality and empathy.[23] Another study involving a survey containing frequently asked questions related to SLE reported that ChatGPT-4 responses were more consistent and comprehensive than those provided by rheumatologists.[19] In our study, when a structured patient history was provided to the ChatGPT-3.5 model, it accurately diagnosed half of the cases, though it performed less effectively than clinicians. Nonetheless, it assigned a rheumatological diagnosis and recommended referral to a rheumatology clinic in 94% of cases. In contrast, clinicians achieved an accuracy rate of approximately 75% based solely on the written patient histories. Furthermore, clinicians were observed to outperform ChatGPT in diagnosing FMF. This difference may be attributed to FMF being considered a more localized and frequently encountered diagnosis among Turkish clinicians, whereas PFAPA is recognized as a more global condition, and ChatGPT's reliance on a globally sourced database. The error margin observed among clinicians highlights the critical importance of conventional processes in a clinical setting, such as obtaining a detailed history directly from the patient and their caregivers, performing a physical examination, synthesizing complex information, and systematically progressing toward a diagnosis.

The widespread adoption of generative AI in the future appears almost certain, particularly as current tools represent only the initial stages of these advancements.[24] However, in the context of using AI in medical decision-making, issues such as ethics, patient consent, and data privacy are also significant, necessitating critical guidelines for the applications of LLMs like ChatGPT.[21]

Our study has several limitations. Firstly, ChatGPT is a general AI model, and its knowledge is not entirely up-to-date, with its latest update dating back to October 2023. Secondly, ChatGPT-3.5 was selected as it is a model that patients can access freely and easily; however, the patient histories provided to the model were structured by two rheumatology fellows before being submitted. These histories were not directly taken from patients but were presented from a physician's perspective, adopting a more professional tone, which likely made the task easier for ChatGPT. In addition, the limited number of patients included in this study (n=50) restricts the generalizability of the findings. Future research should aim to include larger and more heterogeneous cohorts to strengthen the evidence base.

In conclusion, AI has already become an integral part of our lives, serving as a routine tool for both patients and healthcare professionals. However, it is crucial to recognize that AI can lead to incorrect outcomes, particularly in the healthcare field. In our study, the correct prediction rate of rheumatologic diseases was 94%, while this rate was reduced to 50% for accurate diagnoses within the spectrum of rheumatological diseases. Therefore, the results generated by AI must always be reviewed and supervised by an experienced "human" expert. We emphasize that AI is not intended to replace physicians but to enhance their capabilities. It can be considered a complementary rather than an opposing tool in medicine and can be adapted to our clinical practice in a controlled manner to improve our diagnostic ability.

Future research should focus on evaluating the performance of these AI models across various scientific disciplines and assessing their impact on critical decision-making. This would facilitate their safer use under standardized strategies.

Conflict of interest: There is no conflict of interest between the authors.

Researchers' Contribution Rate Statement:
Concept/Design: ÇY, SAB; Analysis/Interpretation: ÇY, BK;Data Collection: ÇY, BK, NB, MK, NK; Writer: ÇY, BK, DGY; Critical Review:PEŞ, DGY, SAB; Approver: SAB

Support and Acknowledgment: No financial support was received from any institution or person.

Ethical approval: Ethics Committee approval was obtained from Gazi University (Decision number: E-77082166-604.01-905676 / 2024-389 dated 27.02.2024).

REFERENCES

1. Garg S, Chauhan A. Chat GPT-4: Potentials, barriers, and future directions for newer medical researchers. *Am J Emerg Med.* 2024;367(6):406-408.
2. Ariyaratne S, Jenko N, Davies AM, Iyengar KP, Botchu R. Could ChatGPT pass the UK radiology fellowship examinations? *Acad Radiol.* 2024;31(5):2178-2182.

3. Scheschenja M, Viniol S, Bastian MB, Wessendorf J, König AM, Mahnken AH. Feasibility of GPT-3 and GPT-4 for in-depth patient education prior to interventional radiological procedures: A comparative analysis. *Cardiovasc Intervent Radiol.* 2024;47(2):245-250.
4. Tran CG, Chang J, Sherman SK, De Andrade JP. Performance of ChatGPT on American board of surgery in-training examination preparation questions. *J Surg Res.* 2024;299:329-335.
5. Palenzuela DL, Mullen JT, Phitayakorn R. AI Versus MD: Evaluating the surgical decision-making accuracy of ChatGPT-4. *Surgery.* 2024;176(2):241-245.
6. Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: A systematic review and meta-analysis. *J Biomed Inform.* 2024:104620.
7. Zaboli A, Brigo F, Sibilio S, Mian M, Turcato G. Human intelligence versus Chat-GPT: Who performs better in correctly classifying patients in triage? *Am J Emerg Med.* 2024;79:44-47.
8. Venerito V, Bilgin E, Iannone F, Kiraz S. AI am a rheumatologist: A practical primer to large language models for rheumatologists. *Rheumatology.* 2023;62(10):3256-3260.
9. Günay S, Öztürk A, Yiğit Y. The accuracy of Gemini, GPT-4, and GPT-4o in ECG analysis: A comparison with cardiologists and emergency medicine specialists. *Am J Emerg Med.* 2024;84:68-73.
10. Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol.* 2023;29(3):721.
11. Howard A, Hope W, Gerada A. ChatGPT and antimicrobial advice: The end of the consulting infection doctor? *Lancet Infect Dis.* 2023;23(4):405-406.
12. Wei Q, Wang Y, Yao Z, et al. Evaluation of ChatGPT's performance in providing treatment recommendations for pediatric diseases. *Pediatr Discov.* 2023;1(3):e42.
13. Nakhleh A, Spitzer S, Shehadeh N. ChatGPT's response to the diabetes knowledge questionnaire: Implications for diabetes education. *Diabetes Technol Ther.* 2023;25(8):571-573.
14. Cadamuro J, Cabitza F, Debeljak Z, et al. Potentials and pitfalls of ChatGPT and natural-language artificial intelligence models for the understanding of laboratory medicine test results. An assessment by the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Working Group on Artificial Intelligence (WG-AI). *Clin Chem Lab Med.* 2023;61(7):1158-1166.
15. Pawar VV, Farooqui S. Ethical consideration for implementing AI in healthcare: A chat GPT perspective. *Oral Oncol.* 2024;149:106682-106682.
16. La Bella S, Attanasi M, Porreca A, et al. Reliability of a generative artificial intelligence tool for pediatric familial Mediterranean fever: Insights from a multicentre expert survey. *Pediatr Rheumatol Online J..* 2024;22(1):78.
17. Gattorno M, Hofer M, Federici S, et al. Classification criteria for autoinflammatory recurrent fevers. *Ann Rheum Dis.* 2019;78(8):1025-1032.
18. Adrovic A, Sahin S, Barut K, Kasapcopur O. Familial Mediterranean fever and periodic fever, aphthous stomatitis, pharyngitis, and adenitis (PFAPA) syndrome: Shared features and main differences. *Rheumatol Int.* 2019;39(1):29-36.
19. Xu D, Zhao J, Liu R, et al. ChatGPT4's proficiency in addressing patients' questions on systemic lupus erythematosus: A blinded comparative study with specialists. *Rheumatology.* 2024;63(9):2450-2456.
20. Madrid-García A, Rosales-Rosado Z, Freites-Nuñez D, et al. Harnessing ChatGPT and GPT-4 for evaluating the rheumatology questions of the Spanish access exam to specialized medical training. *Sci Rep.* 2023;13(1):22129.
21. Krusche M, Callhoff J, Knitza J, Ruffer N. Diagnostic accuracy of a large language model in rheumatology: Comparison of physician and ChatGPT-4. *Rheumatol Int.* 2024;44(2):303-306.
22. Haase I, Xiong T, Rissmann A, Knitza J, Greenfield J, Krusche M. ChatSLE: Consulting ChatGPT-4 for 100 frequently asked lupus questions. *Lancet Rheumatol.* 2024;6(4):e196-e199.
23. 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. *AMA Intern Med.* 2023;183(6):589-596.
24. Jo A. The promise and peril of generative AI. *Nature.* 2023;614(1):214-216.