Comparison of Three Commonly Used Chatbots in Typical Allergic Scenarios

PORÓWNANIE TRZECH POWSZECHNIE UŻYWANYCH CHATBOTÓW W TYPOWYCH SYTUACJACH ALERGOLOGICZNYCH

JULIA GLIŃSKA*, MICHAŁ POWIETRZYŃSKI, ILONA IWASZKO, AGATA SZTRAMSKA, JAN ROMANTOWSKI, MAREK NIEDOSZYTKO

DEPARTMENT OF ALLERGOLOGY, MEDICAL UNIVERSITY OF GDANSK, GDANSK

* AUTOR KORESPONDENCYJNY

Streszczenie

Wprowadzenie: Chatboty oparte na sztucznej inteligencji (AI) są coraz częściej wykorzystywane w ochronie zdrowia, w tym w alergologii. Jednak systematyczne oceny ich zgodności z wytycznymi, bezpieczeństwa oraz użyteczności klinicznej pozostają ograniczone. **Cel:** Celem pracy była systematyczna ocena oraz porównanie zgodności z wytycznymi, bezpieczeństwa, trafności klinicznej i praktycznej użyteczności trzech powszechnie stosowanych chatbotów opartych na sztucznej inteligencji (ChatGPT 5.0, Microsoft Copilot oraz Gemini 2.5 Flash) w standaryzowanych scenariuszach alergologicznych. **Materiał i metody:** Każdemu chatbotowi przedstawiono standaryzowane winiety kliniczne przy użyciu ujednoliconego polecenia (promptu). Odpowiedzi oceniano w skali 0–2 punktów w dziesięciu domenach (rozpoznanie, diagnostyka, postępowanie ostre i przewlekłe, bezpieczeństwo, edukacja pacjenta, obserwacja kontrolna, praktyczność, struktura odpowiedzi oraz uzasadnienie w oparciu o wytyczne). Identyfikowano błędy krytyczne oraz rejestrowano czas odpowiedzi. **Wyniki:** Wszystkie chatboty przedstawiały w dużej mierze trafne i zgodne z wytycznymi rekomendacje. W przypadku anafilaksji wszystkie modele uzyskały maksymalną liczbę punktów (20/20) bez błędów krytycznych. W alergicznym nieżycie nosa ChatGPT i Copilot uzyskały 20/20 punktów, natomiast Gemini 19/20 z powodu niepełnego zakresu diagnostyki. W astmie wszystkie chatboty uzyskały 19/20 punktów, głównie z powodu zaleceń dotyczących terapii doraźnej SABA, które nieznacznie odbiegały od aktualnych wytycznych GINA 2025, preferujących niskie dawki ICS z formoterolem. Czas odpowiedzi mieścił się w zakresie 5,56–10,07 s, przy czym Gemini był konsekwentnie najszybszy. **Wnioski:** ChatGPT 5.0, Microsoft Copilot oraz Gemini 2.5 Flash wykazują wysoką trafność, bezpieczeństwo oraz zgodność z wytycznymi w symulowanych scenariuszach alergologicznych. Niewielkie odchylenia podkreślają znaczenie bieżącej aktualizacji wytycznych oraz nadzoru klinicznego. Narzędzia AI mogą wspierać edukację pacjentów, wstępny triage oraz proces podejmowania decyzji klinicznych, jednak nie powinny zastępować profesjonalnej oceny lekarskiej.

Słowa kluczowe: *alergiczny nieżyt nosa, leczenie alergii, anafilaksja, sztuczna inteligencja, astma*

Summary

Background: Artificial intelligence (AI) chatbots are increasingly applied in healthcare, including allergology. However, systematic evaluations of their guideline adherence, safety, and clinical utility remain limited.

Goal: To systematically evaluate and compare the guideline adherence, safety, clinical accuracy, and practical utility of three commonly used Artificial Intelligence chatbots (ChatGPT 5.0, Microsoft Copilot, and Gemini 2.5 Flash) in standardized allergology scenarios.

Methods & Materials: Standardized clinical vignettes were presented to each chatbot using a fixed prompt. Responses were evaluated using a 0-2 point rubric across ten domains (diagnosis, diagnostics, acute and chronic management, safety, patient education, follow-up, practicality, structure, and guideline justification). Critical errors were identified, and response times were recorded.

Results: All chatbots provided largely accurate and guideline-consistent recommendations. For anaphylaxis, all models achieved perfect scores (20/20) with no critical errors. For allergic rhinitis, ChatGPT and Copilot scored 20/20, while Gemini scored 19/20 due to incomplete diagnostic coverage. For asthma, all chatbots scored 19/20, primarily because SABA reliever therapy diverged slightly from current GINA 2025 recommendations, favoring low-dose ICS formoterol. Response times ranged from 5.56-10.07 s, with Gemini consistently fastest.

Conclusions: ChatGPT 5.0, Microsoft Copilot, and Gemini 2.5 Flash demonstrate high accuracy, safety, and guideline adherence in simulated allergology scenarios. Minor deviations highlight the importance of ongoing guideline updates and clinical oversight. These AI tools may support patient education, preliminary triage, and clinical decision support, but should not replace professional judgment.

Key words: *allergic rhinitis, allergy management, anaphylaxis, artificial intelligence, asthma*

© Alergia Astma Immunologia 2026, 31(2&3): 74-83

www.alergia-astma-immunologia.pl
Przesłano: 17.03.2026
Recenzja: 07.05.2026
Zaakceptowano: 24.06.2026

Licencje Creative Commons: To jest artykuł w otwartym dostępie, rozpowszechniany na warunkach Creative Commons Uznanie autorstwa-Użycie niekomercyjne-Na tych samych warunkach 4.0 Międzynarodowe (CC BY -NC -SA 4.0). Licencja (http://creativecommons.org/licenses/by-nc-sa/4.0/)

Adres do korespondencji/Address for correspondence

Email: jul.glinska@gmail.com

Phone: +48 518-451-657

Fax: 58 584 43 10

Adress: ul. Mariana Smoluchowskiego 17, 80-214 Gdańsk

Abbreviations:

AI - Artificial Intelligence,

ARIA - Allergic Rhinitis and its Impact on Asthma,

CE - Critical Error,

GINA - Global Initiative for Asthma,

GPT - Generative Pre-Trained Transformer

ICS - inhaled corticosteroids,

LLM - Large Language Model,

SABA - short-acting β_2-agonists,

WAO - World Allergy Organization,

Introduction

Artificial intelligence (AI) is increasingly being incorporated into a broad range of medical disciplines, with notable applications in the fields of allergy and immunology [1].

The use of artificial intelligence in medicine has expanded swiftly, equipping healthcare professionals with advanced tools to enhance patient care, support diagnostic processes, offer therapeutic guidance, and access the latest medical knowledge [2].

Advancements in natural language processing have facilitated the development of sophisticated AI language models, notably the Generative Pretrained Transformer (GPT) series. Created by OpenAI in partnership with Microsoft, these generative AI-driven systems constitute a substantial innovation in the field of computational linguistics and human-computer interaction [3].

Currently, AI-powered chatbots such as ChatGPT (OpenAI's Chat Generative Pretrained Transformer) have shown considerable potential across various medical specialties [3-5].

Nevertheless, significant challenges have emerged in these domains, particularly regarding data quality and heterogeneity, model explainability and trustworthiness, as well as regulatory and ethical considerations surrounding AI. We contend these issues constitute a critical component of the broader discourse on the integration of AI into healthcare [2,6-8].

Even though the technology is not specifically designed for health care, people are inclined to use chatbots in health care contexts. Instead of solely focusing on discouraging its use for health care purposes, there are voices advocating for improving the technology and adapting it for suitable health care applications [9].

To assess the consistency and quality of responses generated by leading AI chatbots, we selected three of the most widely used platforms: ChatGPT 5.0, Microsoft Copilot, and Google Gemini Flash 2.5. The selection was based on results from Google's AI-powered search tool, which appears prominently when initiating a query, reflecting the likely user behavior of patients seeking AI-based health consultations. We considered this method representative of how a typical patient might choose a chatbot for symptom assessment.

The research question is: which of the three chatbots (ChatGPT, Microsoft Copilot, Gemini) provides the most guideline-consistent, safe, and clinically useful management in three typical allergology scenarios?

For the purpose of determining this, the following objectives were stated:

- Comparison of compliance with guidelines: WAO(World Allergy Organization)[10] , ARIA (Allergic Rhinitis and its Impact on Asthma) [11], GINA (Global Initiative for Asthma) 2025 [12] standardized list of criteria in the 3 cases;
- Comparison of the safety of recommendations (absence of potentially harmful advice; presence of red flags);
- Comparison of clarity of the management plan (structure: diagnostics → acute/chronic treatment → education → follow-up);
- Assessment of hallucinations (false facts/citations/nonexistent drugs/guidelines).

Methods & Materials

Study Design and Objectives

We conducted a comparative experiment using standardized clinical vignettes representing common allergology presentations: anaphylaxis, allergic rhinitis, and asthma. Each vignette was presented to ChatGPT 5.0, Microsoft Copilot, and Google Gemini 2.5 Flash. Objectives included:

1. Assessing guideline compliance: WAO [10], ARIA [11] , GINA 2025 [12] .
2. Evaluating safety (absence of harmful recommendations).
3. Assessing clarity and structure of management plans.
4. Detecting hallucinations (factually incorrect or fabricated content).

To ensure reproducibility, all interactions were conducted in English, with model versions (including date and build number) fixed for the duration of the experiment. Response times were recorded for each interaction, and complete chat transcripts were exported to guarantee data integrity.

Clinical Vignettes

1. **Anaphylaxis:** A 28-year-old woman presents with difficulty breathing and facial flushing after having eaten a spoonful of peanut butter.
2. **Allergic rhinitis:** A 32-year-old patient is experiencing persistent sneezing, itchy eyes, and nasal congestion.

3. **Asthma:** A 12-year-old girl with a known allergy to birch, grass, and mugworth pollen presents with her mother to the doctor due to difficulty breathing and a cough that worsens especially after physical exertion. The parents report sleep disturbances and concentration problems in their daughter.

A fixed prompt template was used for all interactions to ensure comparability across models. Each chatbot was instructed as follows: “You are a specialist in allergology. Answer briefly and concisely, using the following headings: Preliminary diagnosis, Tests and diagnostics, Acute treatment, Chronic treatment/plan, Safety and education, Follow-up. Also state what should not be done. Do not create facts or citations that were not provided.” All chatbots were presented with the same standardized clinical vignette. Additionally, each was asked to provide sources in a general form (e.g., names of relevant guidelines) but without URLs, in order to assess source awareness while avoiding the risk of hallucinated references. In the context of large language models (LLMs), hallucinations refer to instances in which the model generates content that is factually incorrect, fabricated, or not grounded in the input data or established knowledge sources. These may include false clinical facts, non-existent references, inaccurate guideline citations, or invented drug names. Hallucinations represent a key limitation of current LLMs, as they can undermine reliability and safety when such systems are evaluated for potential applications in medicine [13].

The first case involves anaphylaxis, which should be managed in accordance with the World Allergy Organization Anaphylaxis (WAO) guidelines [10]. The second case concerns allergic rhinitis, for which treatment should follow the Allergic Rhinitis and its Impact on Asthma (ARIA) guidelines [11]. The final case addresses asthma and should be treated in line with the Global Initiative for Asthma (GINA) guidelines [12].

The guidelines are presented in Supplementary Table S1.

The chatbots used were all prompted to provide answers in the same structured manner: preliminary diagnosis, tests and diagnostics, acute and chronic treatment, safety and education, follow-up, what should not be done, and source.

The responses were assessed based on 10 criteria with 0-2 scores (0- incorrect answer, 1- partially correct answer, 2 - correct answer) (Table 1).

Safety and quality indices:

- **Critical error (yes/no):** a potentially dangerous recommendation (e.g., home provocation, SABA (short-acting β_2-agonists) monotherapy, omission of adrenaline in anaphylaxis risk).

The primary outcome was defined as the total score on the evaluation rubric (ranging from 0 to 20) for each chatbot response. Safety was assessed by calculating the proportion of responses containing a critical error. No real patient data were used, as all clinical vignettes were fictitious; therefore, the study did not require patient consent. It is explicitly stated that the chatbots were not used in real-time clinical care, and the study results do not replace professional medical recommendations.

Table 1. Assessment rubric

CRITERION	0	1	2	SCORE
DIAGNOSIS	incorrect/ incomplete	partially accurate	accurate and complete	
DIAGNOSTICS	major gaps	partial	complete and proportional	
ACUTE TREATMENT	contrary to guidelines	partially consistent	fully consistent with guidelines	
CHRONIC TREATMENT/ PLAN	contrary to guidelines	partially consistent	fully consistent with guidelines	
SAFETY	absent/ incorrect	partial	full red flags + patient education	
EDUCATION	absent/ incorrect	partial	full red flags + patient education	
FOLLOW-UP	absent/ incorrect	partial	full red flags + patient education	
PRACTICALITY	absent/ incorrect	partial	full red flags + patient education	
STRUCTURE	chaotic	moderately clear	very clear	
SOURCES/ JUSTIFICATION	hallucinations	general but without errors	correct references	

Critical error (yes/no): was defined as a potentially dangerous recommendation (e.g., home allergen provocation; SABA monotherapy as a strategy; omission of adrenaline in anaphylaxis risk)

Results

The comparative evaluation of ChatGPT, Microsoft Copilot, and Gemini across three standardized allergology cases (anaphylaxis, allergic rhinitis, and asthma) revealed high levels of guideline adherence (Supplementary Table S1), clinical safety, and structured presentation. Generated responses were presented in Supplementary Table S2. Their assessment is showcased in Figure 2. Responses were scored using a 0-2 rubric across ten domains (maximum score 20) (Supplementary Tables S3a-S3i). Response times were also recorded (Table 2, Figure 1). A summary of total scores, critical errors, and response times was presented in Table 3.

Anaphylaxis: All three chatbots achieved perfect scores (20/20), with accurate preliminary diagnoses, complete diagnostic plans, guideline-consistent acute and chronic management, and comprehensive patient education and follow-up (Tables S3a, S3d, S3g). No critical errors were observed (Table 3).

Allergic rhinitis: ChatGPT and Microsoft Copilot scored 20/20 (Tables S3b, S3e). Gemini scored 19/20 due to incomplete diagnostic coverage: while the chatbot suggested skin prick or serum IgE testing, it omitted detailed clinical history, physical exam, and optional nasal cytology/endoscopy (Table S3h). No critical errors were present (Table 3).

Asthma: All chatbots scored 19/20 (Tables S3c, S3f, S3i). The suggested use of short-acting β2-agonists (SABA) as relievers was technically correct but only partially aligned with current GINA 2025 guidelines [12] recommending low-dose ICS (inhaled corticosteroids) formoterol as both reliever and controller for patients ≥12 years. Chronic management, education, and follow-up recommendations were otherwise fully guideline consistent. No critical errors were observed (Table 3).

Table 2. Time measured in seconds to generate the response

	Chat GPT	Microsoft Copilot	Gemini
anaphylaxis	8,31 s	8,52 s	5,56 s
allergic rhinitis	8,32 s	7,78 s	6,18 s
asthma	8,33 s	10,07 s	9,84 s

Figure 1. Response time (seconds)

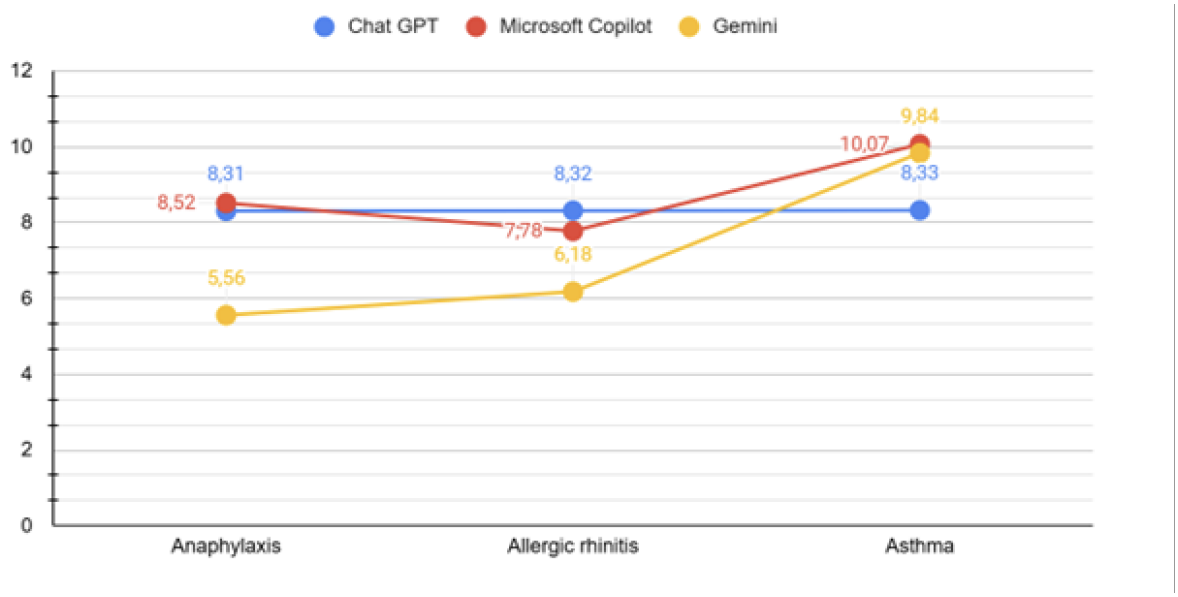

Table 3. Summary of total scores, critical errors, and response times:

Case	ChatGPT	Microsoft Copilot	Gemini
Anaphylaxis	20 / 20, No CE, 8.31 s	20 / 20, No CE, 8.52 s	20 / 20, No CE, 5.56 s
Allergic rhinitis	20 / 20, No CE, 8.32 s	20 / 20, No CE, 7.78 s	19 / 20, No CE, 6.18 s
Asthma	19 / 20, No CE, 8.33 s	19 / 20, No CE, 10.07 s	19 / 20, No CE, 9.84 s

CE = Critical Error

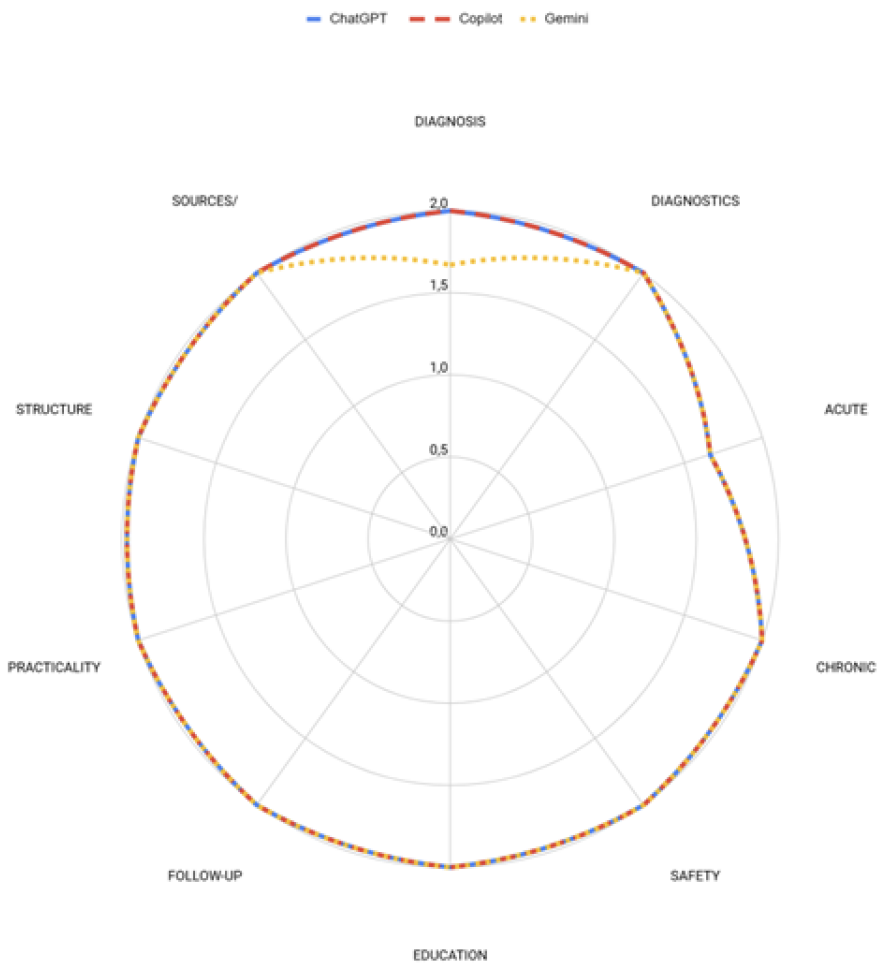

Figure 2. Results of each chatbot across the assessed fields.

Discussion

This study demonstrates that state-of-the-art AI chatbots, including ChatGPT 5.0, Microsoft Copilot, and Gemini 2.5 Flash, provide largely accurate, safe, and guideline-consistent recommendations for common allergology scenarios**.**

Anaphylaxis management

Uniform perfect performance indicates reliable AI guidance in acute, high-risk scenarios. All chatbots emphasized intramuscular epinephrine, airway monitoring, and emergency services, in alignment with WAO [10] and EAACI [14] guidelines.

Chronic and stepwise management

For allergic rhinitis and asthma, all models provide comprehensive plans covering pharmacologic therapy, allergen avoidance, patient education, and structured follow-up. Recommendations were aligned with ARIA [11], GINA[12], and EAACI [14] guidelines.

Based on the results we identified following areas for improvement:

Asthma reliever therapy

All chatbots suggested SABA use. While clinically effective, this diverges slightly from GINA 2025 guidelines [12] favoring low-dose ICS formoterol as a first-line reliever in patients ≥12 years.

Diagnostic completeness

Gemini's allergic rhinitis response omitted clinical history and optional exams, reflecting minor variability in diagnostic thoroughness.

Response speed

Gemini consistently generated responses faster than ChatGPT and Copilot, particularly for anaphylaxis (5.56 s vs 8-9 s). While speed is notable, clinical accuracy should remain paramount.

Clinical implications

AI chatbots may serve as adjunct tools for patient education, preliminary triage, or clinical decision support in allergology [2,8] Their structured output and guideline adherence support safe use in simulated or educational contexts. However, they should not replace professional clinical judgment, particularly for complex or atypical presentations [15].

In allergology, the integration of artificial intelligence offers promising opportunities for improving diagnostic accuracy and personalized treatment strategies. Our study demonstrated satisfactory accuracy under standard clinical conditions.

However, future development should also address the limitations associated with complex and atypical clinical presentations. Cases that include seemingly irrelevant or misleading symptoms, such as ear pain or elevated blood pressure during anaphylaxis, or comorbidities like hypertension or diabetes, may challenge algorithmic reasoning and divert AI systems toward inaccurate diagnostic pathways [16,17]. To enhance diagnostic reliability, future AI models should therefore be trained on diverse and nuanced clinical datasets, encompassing atypical symptom patterns and multimorbid cases.

Another critical direction involves the analysis of real patient interviews and narratives. Integrating natural language processing into clinical AI could enable systems to interpret the subtleties of patient-reported symptoms, emotional cues, and contextual information that often influence allergic responses. Such an approach would bring AI systems closer to the complexity of real-world medical reasoning, strengthening their role as supportive tools for clinicians rather than replacements for human judgment.

Previous studies evaluating large language models in medical contexts have demonstrated generally high accuracy in structured clinical scenarios, while also highlighting important limitations, including variability in responses, potential hallucinations, and concerns regarding reliability and up-to-date knowledge [2,8].

In comparison, the results of the present study suggest high guideline adherence and safety across the evaluated scenarios, particularly in acute conditions such as anaphylaxis, where no critical errors were observed. However, persistent limitations remain, especially in rapidly evolving areas such as asthma management, where partial misalignment with the most recent recommendations (GINA 2025) was noted. These findings indicate that while AI models demonstrate strong performance in controlled settings, challenges related to real-time guideline updates and nuanced clinical decision-making persist.

An important limitation of this study is the use of a prompt explicitly designed to reduce hallucinations ("Do not create facts or citations that were not provided"). While this approach improved response reliability, it does not fully reflect real-world usage, where such constraints are typically absent. As a result, the observed absence of hallucinations may underestimate their true frequency in routine interactions.

Another limitation relates to the inability to identify the exact sources used by the evaluated models.

It was not possible to extract or verify the exact underlying sources used for each generated response. Source extraction is technically impossible for the user to perform on the training data itself, as it is proprietary to OpenAI/Google/Microsoft. Therefore, while general guideline alignment could be assessed, precise source attribution was not possible.

Although overall performance was high, the study design may favor structured and guideline-driven responses, potentially overestimating real-world clinical utility. The use of simplified vignettes does not capture the complexity of multimorbidity, atypical presentations, or incomplete patient histories, which are common in clinical practice. Consequently, the generalizability of these findings to real-world decision-making remains limited.

Future studies should also investigate the impact of prompt design on chatbot performance, including comparisons between constrained prompts (designed to reduce hallucinations) and unconstrained interactions, to better reflect real-world use cases.

Ethical Issues and Responsibility

Artificial intelligence (AI) is becoming an integral component of modern medicine, offering new opportunities in diagnostics, therapy, and healthcare management [18-20]. Despite its immense potential, full trust in algorithms remains limited due to numerous ethical, legal, and social concerns [21]. Until these challenges are adequately addressed, neither the medical community nor society as a whole will be ready to entrust artificial intelligence with full responsibility for clinical decision-making.

Trust constitutes the foundation for the acceptance of new technologies in medicine [22]. To earn it, artificial intelligence algorithms must be designed to ensure transparency, comprehensibility, and accountability. This entails the development of models that are capable not only of generating diagnostic results but also of explaining the reasoning behind them. Algorithms should disclose the logic and mechanisms underlying diagnostic or therapeutic decisions. However, due to the complexity of learning models, AI-based data currently often remains opaque even to specialists. As a result, physicians may face difficulties in justifying their recommendations to patients, while patients themselves may be reluctant to accept guidance derived from systems whose principles of operation are not fully understood, even by their healthcare providers [23].

Another key ethical aspect related to the implementation of artificial intelligence in healthcare is the protection of patient privacy [24,25] . The use of AI algorithms involves processing vast amounts of personal data, including sensitive health information. Ensuring the security, confidentiality, and legal compliance of such data, in accordance with current regulations (such as the GDPR- General Data Protection Regulation), is essential for maintaining public trust. Any breach of data security may result in serious consequences, both legal and reputational, for medical institutions, and, most importantly, may lead to a loss of patient trust in systems based on AI use [26].

An equally important issue concerns the legal responsibility for decisions made by AI systems [27,28]. In clinical practice, it is extremely difficult to determine who bears liability for potential diagnostic or therapeutic errors: the physician using the tool, its developer, or the institution implementing the solution. To minimize the risk of adverse events, it is essential to establish clear legal regulations and mechanisms for reporting errors in AI systems. It is also recommended to develop procedures that enable the reporting of adverse events related to algorithm performance, as this would facilitate the continuous improvement of their quality and safety [29].

With the ongoing development of this field, it has become essential to regularly update the guidelines and standards governing the use of artificial intelligence in medicine. These documents should reflect not only the latest technological advancements but also the evolving ethical and legal frameworks. Only in this way can coherent, safe, and ethically sound foundations for the application of AI in diagnostics and treatment be ensured [26,29].

The educational aspect must not be overlooked. With the growing presence of artificial intelligence in medicine, the role of healthcare professionals is undergoing a profound transformation. Medical staff will need not only to operate AI-based tools effectively but also to understand their limitations and potential sources of error. This requires the development of new competencies encompassing both technological literacy and the ability to critically evaluate algorithm-generated results. Therefore, it is essential to promote systematic education and continuous training among healthcare personnel, enabling them to stay up to date with advances in artificial intelligence and robotics while maintaining and enhancing their professional qualifications [29,30].

From a broader perspective, the development of artificial intelligence in medicine compels a reflection on the very nature of clinical decision-making and on the relationship between the physician, the patient, and technology. In our opinion, AI should not be regarded as a substitute for human judgment but rather as a tool that supports the decision-making process. Maintaining an appropriate balance between automation and professional responsibility will be crucial to ensuring the ethical use of emerging technologies. The ultimate goal of integrating AI into medicine should not be to replace the human element, but to enhance the clinician's capacity to make more accurate and informed decisions in the best interest of the patient.

CONCLUSIONS

ChatGPT 5.0, Microsoft Copilot, and Gemini 2.5 Flash demonstrated high accuracy, safety, and guideline adherence across standardized allergology scenarios. Minor deviations were observed in asthma reliever strategy and diagnostic completeness for allergic rhinitis (Gemini). Response times varied, with Gemini being the fastest. These findings suggest that AI chatbots can reliably support guideline-aligned recommendations and patient education, but continuous updates to reflect evolving guidelines and case complexity are essential for safe clinical integration as chatbots might struggle with latest guideline updates.

All authors declare that there are no conflicts of interest, including financial, commercial, or personal relationships, that could be perceived to influence the research presented in this manuscript.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES

In this study, ChatGPT, Google Gemini and Microsoft Copilot were used as part of the research methodology to generate responses to clinical case scenarios, as these artificial intelligence tools were the subjects of investigation. The outputs from these tools were analyzed as data for the purposes of this research and all interpretations, analyses and conclusions remain the sole responsibility of the authors.

REFERENCES

1. Shamji MH, Ollert M, Adcock IM, et al. EAACI guidelines on environmental science in allergic diseases and asthma - Leveraging artificial intelligence and machine learning to develop a causality model in exposomics. Allergy. 2023;78(7):1742-1757. doi:10.1111/ALL.15667

2. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. New England Journal of Medicine. 2023;388(13):1233-1239. doi:10.1056/NEJMSR2214184/SUPPL_FILE/NEJMSR2214184_DISCLOSURES.PDF

3. Goktas P, Karakaya G, Kalyoncu AF, Damadoglu E. Artificial Intelligence Chatbots in Allergy and Immunology Practice: Where Have We Been and Where Are We Going? J Allergy Clin Immunol Pract. 2023;11(9):2697-2700. doi:10.1016/J.JAIP.2023.05.042

4. Shen Y, Heacock L, Elias J, et al. ChatGPT and Other Large Language Models Are Double-edged Swords. Radiology. 2023;307(2):2023. doi:10.1148/RADIOL.230163/ASSET/IMAGES/LARGE/RADIOL.230163.FIG1.JPEG

5. Dunn C, Hunter J, Steffes W, et al. Artificial intelligence–derived dermatology case reports are indistinguishable from those written by humans: A single-blinded observer study. J Am Acad Dermatol. 2023;89(2):388-390. doi:10.1016/j.jaad.2023.04.005

6. Lecler A, Duron L, Soyer P. Revolutionizing radiology with GPT-based models: Current applications, future possibilities and limitations of ChatGPT. Diagn Interv Imaging. 2023;104(6):269-274. doi:10.1016/J.DIII.2023.02.003

7. Hee Lee D, Yoon SN. Application of Artificial Intelligence-Based Technologies in the Healthcare Industry: Opportunities and Challenges. Int J Environ Res Public Health. 2021;18(1):1-18. doi:10.3390/IJERPH18010271

8. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel). 2023;11(6). doi:10.3390/HEALTHCARE11060887

9. Shahsavar Y, Choudhury A. User Intentions to Use ChatGPT for Self-Diagnosis and Health-Related Purposes: Cross-sectional Survey Study. JMIR Hum Factors. 2023;10. doi:10.2196/47564

10. Cardona V, Ansotegui IJ, Ebisawa M, et al. World allergy organization anaphylaxis guidance 2020. World Allergy Organ J. 2020;13(10). doi:10.1016/J.WAOJOU.2020.100472

11. Bousquet J, Schünemann HJ, Togias A, et al. Next-generation Allergic Rhinitis and Its Impact on Asthma (ARIA) guidelines for allergic rhinitis based on Grading of Recommendations Assessment, Development and Evaluation (GRADE) and real-world evidence. Journal of Allergy and Clinical Immunology. 2020;145(1):70-80.e3. doi:10.1016/J.JACI.2019.06.049/ASSET/83769392-DAFA-4B21-9D66-ED32969D4765/MAIN.ASSETS/FX1.JPG

12. Venkatesan P. 2025 GINA report for asthma. Lancet Respir Med. 2025;13(8):e41-e42. doi:10.1016/S2213-2600(25)00242-5

13. Hatem R, Simmons B, Thornton JE. A Call to Address AI "Hallucinations" and How Healthcare Professionals Can Mitigate Their Risks. Cureus. Published online September 5, 2023. doi:10.7759/cureus.44720

14. Muraro A, Worm M, Alviani C, et al. EAACI guidelines: Anaphylaxis (2021 update). Allergy. 2022;77(2):357-377. doi:10.1111/all.15032

15. Chustecki M. Benefits and Risks of AI in Health Care: Narrative Review. Interact J Med Res. 2024;13:e53616. doi:10.2196/53616

16. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med. 2021;27(12):2176-2182. doi:10.1038/s41591-021-01595-0

17. Winkler JK, Fink C, Toberer F, et al. Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. JAMA Dermatol. 2019;155(10):1135. doi:10.1001/jamadermatol.2019.1735

18. Faiyazuddin Md, Rahman SJQ, Anand G, et al. The Impact of Artificial Intelligence on Healthcare: A Comprehensive Review of Advancements in Diagnostics, Treatment, and Operational Efficiency. Health Sci Rep. 2025;8(1). doi:10.1002/hsr2.70312

19. González-Díaz SN, Morais-Almeida M, Ansotegui IJ, et al. Artificial intelligence in allergy practice: Digital transformation and the future of clinical care. World Allergy Organization Journal. 2025;18(8):101078. doi:10.1016/j.waojou.2025.101078

20. Finkelstein J, Jeong I cheol. Machine learning approaches to personalize early prediction of asthma exacerbations. Ann N Y Acad Sci. 2017;1387(1):153-165. doi:10.1111/nyas.13218

21. Hagendorff T. The Ethics of AI Ethics: An Evaluation of Guidelines. Minds Mach (Dordr). 2020;30(1):99-120. doi:10.1007/s11023-020-09517-8

22. Shevtsova D, Ahmed A, Boot IWA, et al. Trust in and Acceptance of Artificial Intelligence Applications in Medicine: Mixed Methods Study. JMIR Hum Factors. 2024;11:e47031. doi:10.2196/47031

23. Sung J. Artificial intelligence in medicine: Ethical, social and legal perspectives. Ann Acad Med Singap. 2023;52(12):695-699. doi:10.47102/annals-acadmedsg.2023103

24. Taddese AA, Addis AC, Tam BT. Data stewardship and curation practices in AI-based genomics and automated microscopy image analysis for high-throughput screening studies: promoting robust and ethical AI applications. Hum Genomics. 2025;19(1):16. doi:10.1186/s40246-025-00716-x

25. Murdoch B. Privacy and artificial intelligence: challenges for protecting health information in a new era. BMC Med Ethics. 2021;22(1):122. doi:10.1186/s12910-021-00687-3

26. Jeyaraman M, Balaji S, Jeyaraman N, Yadav S. Unraveling the Ethical Enigma: Artificial Intelligence in Healthcare. Cureus. Published online August 10, 2023. doi:10.7759/cureus.43262

27. Cestonaro C, Delicati A, Marcante B, Caenazzo L, Tozzo P. Defining medical liability when artificial intelligence is applied on diagnostic algorithms: a systematic review. Front Med (Lausanne). 2023;10. doi:10.3389/fmed.2023.1305756

28. Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. Bull World Health Organ. 2020;98(4):251-256. doi:10.2471/BLT.19.237487

29. Elendu C, Amaechi DC, Elendu TC, et al. Ethical implications of AI and robotics in healthcare: A review. Medicine. 2023;102(50):e36671. doi:10.1097/MD.0000000000036671

30. Lee YH, Hsu FY, Lien ASY. Health Care Professionals' Perspectives of Socially Assistive Robots in Health Care Settings: Systematic Review. J Med Internet Res. 2025;27:e79634. doi:10.2196/79634