Jordan Medical Journal, Supplement 1, 2024
DOI: https://doi.org/10.35516/jmj.v58i2.2960

Evaluation of Artificial Intelligence-Based Chatbot Responses to Common Dermatological Queries

Indrasish Podder[1], Neha Pipil[2], Arunima Dhabal[3], Shaikat Mondal[4], Vitsomenuo Judith Pienyii[5], Himel Mondal[6]✉

Abstract

Background and aim: Conversational artificial intelligence (AI) can streamline healthcare by offering instant and personalized patient interactions, answering queries, and providing general medical information. Its ability for early disease detection and treatment planning may improve patient outcomes. We aimed to investigate the utility of conversational AI models in addressing diagnostic challenges and treatment recommendations for common dermatological ailments.

Methods: A dataset comprising 22 case vignettes of dermatological conditions was compiled, each case accompanied by three specific queries. These case vignettes were presented to four distinct conversational AI models - ChatGPT 3.5, Google Gemini, Microsoft Copilot (GPT 4), and Perplexity.ai and responses were saved. To assess clinical appropriateness and accuracy, two expert dermatologists independently evaluated the responses of the AI systems using a 5-point Likert scale ranging from highly accurate (= 5) to inaccurate (= 1).

Results: The average score of ChatGPT was 4.1±0.61, Gemini was 3.86±0.88, Copilot was 4.51±0.33, and Perplexity was 4.14±0.64, P=0.01. The high difference in score was for Gemini vs. Copilot (Cohen's d = 0.98), ChatGPT vs. Copilot (Cohen's d = 0.83), and Copilot vs. Perplexity (Cohen's d = 0.75). All of the chat bot's scores were similar to 80% accuracy (one sample t-test with a hypothetical value of 4) except Copilot which showed an accuracy of nearly 90%.

Conclusion: This study highlights AI chatbots' potential in dermatological healthcare for patient education. However, findings underscore their limitations in accurate disease diagnosis. The programs may be used as a supplementary resource rather than primary diagnostic tools.

Keywords: Artificial Intelligence, Dermatologists, Search Engine, Delivery of Health Care, Intelligence.

(J Med J 2024; Supplement 1: 271–278)

Received	*Accepted*
April 3, 2024	*July 11, 2024*

INTRODUCTION

Dermatological diseases encompass a broad spectrum of conditions, ranging from mild irritations to severe disorders, affecting individuals across all age groups and demographics. Accurate diagnosis and timely treatment are crucial to prevent disease progression, alleviate discomfort, and mitigate cosmetic and psychological impacts. [1] However, dermatological diagnosis often poses challenges due to the visual nature of skin conditions, variations in presentations, and the shortage of dermatology specialists in certain regions. These challenges may deprive many patients in obtaining healthcare or relevant information for their diseases. [2,3]

Advancements in artificial intelligence (AI) have revolutionized various industries including healthcare. One intriguing application is the integration of AI-powered conversational agents into medical diagnosis and treatment. [4] Dermatological diseases, being among the most prevalent health concerns globally, present a unique opportunity for the implementation of such technology. [5] The utilization of conversational AI in diagnosing and solving cases of common

[1] Department of Dermatology and Veneriology, College of Medicine and Sagore Dutta Hospital, West Bengal, India.
[2] Department of Pharmacology, Rajshree Medical Research Institute, Bareilly, Uttar Pradesh, India.
[3] Department of Dermatology, Jagannath Gupta Institute of Medical Sciences, West Bengal, India.
[4]Department of Physiology, Raiganj Government Medical College and Hospital, West Bengal, India.
[5] Department of Physiology, Nagaland Institute of Medical Science and Research, Kohima, Nagaland, India.
[6]Department of Physiology, All India Institute of Medical Sciences, Deoghar, Jharkhand, India.
✉Corresponding author: himelmkcg@gmail.com

dermatological conditions has not been explored extensively. [6]

The ability of AI systems to comprehend and generate human language facilitates seamless interaction between technology and users, enabling the delivery of medical assistance through natural conversations. [7] Integrating those programs into dermatological diagnosis holds potential advantages, such as enabling patients to describe their symptoms in a familiar manner. Moreover, these systems have the capacity to analyze a vast database of medical literature that helps in aiding diagnosis and treatment recommendations. [8-10]

Various AI Chabot like ChatGPT-3.5, Google Gemini, Microsoft Copilot (GPT-4, customized), Perplexity, and others may provide different output to a same question due to their variations in underlying architecture, domain-specific fine-tuning, and intended application. [11-13] A recent article by Sallam et al. emphasized the need to evaluate the large language model (LLM) like ChatGPT in healthcare for better decision-making regarding the use of LLM in various domains of healthcare. [14]

With this background, this study aimed to explore the current landscape of AI Chatbot in solving cases of common dermatological diseases. By examining the capability, this paper seeks to highlight the potential benefits and challenges associated with integrating AI Chatbot into dermatological practice.

MATERIALS AND METHOD

Study Type and Settings

This study was a cross-sectional study. The research was conducted on the World Wide Web. The study's objective was to analyze the performance of AI chatbots in solving dermatological cases. The study is compliant to METRICS and the details are available in Table 1.[15]

Table 1: Compliance to METRICS guidelines

Attribute	Details
Model and settings	ChatGPT 3.5 (free research version), Google Gemini, Microsoft Copilot (Precise search; GPT4), and Perplexity. The chatbots were accessed from personal computer connected to home broadband internet connection. The data collection spanned from 5 to 10 December, 2023.
Evaluation approach	The responses of LLMs were rated by two expert raters on an objective scale as shown in Table 2.
Time of testing	The tests were conducted during 5 – 10 December, 2023.
Transparency of data source	The cases were prepared by two dermatologists and can be obtained from the corresponding author for research purpose.
Range of tested topics	The topics are related to common dermatological diseases encountered in tertiary care hospitals.
Randomization of queries	The queries were randomly asked to the LLMs with a freshly initiated chat.
Individual factor/interrater reliability	The raters were blinded regarding the name of the LLM they are rating and interrater reliability has been calculated by intraclass correlation coefficient
Count of queries	A total of 22 common dermatological cases were asked to each LLM
Specificity of prompts	The prompt was structured to define role of the LLM, specific task allocation, the details of the case, and the questions to answer as shown in Figure 1. The first response was considered final and we did not use "regeneration" function.
METRICS guidelines can be obtained from DOI: 10.2196/54704 [15]	

Chatbot selection

We have taken the large language model generative artificial intelligence chatbots for this study. Only free (accessible to any users) were selected. With reference to previously published articles on similar topics in other subjects, four freely available AI chatbots were tested – ChatGPT 3.5 (free research version), Google Gemini, Microsoft Copilot (Precise search; GPT4), and Perplexity. Henceforth in the manuscript, they are called ChatGPT, Gemini, Copilot, and Perplexity, respectively. The study was conducted in December 2023 on the World Wide Web accessed on a personal computer (ASUS VivoBook Max X541N) connected with a personal 150 Mbps broadband internet connection.

Dermatology cases

For comparing the score of the four groups by

ANOVA, with a significance level of 0.05, power 0.9, and effect size 0.5, the sample size is 16 in each group. However, we aimed to include more than this minimum sample size. A dataset comprising 22 common dermatology cases was compiled, each accompanied by three standardized questions pertaining to the patient's symptoms, medical history, and potential diagnosis. The cases spanned a range of dermatological conditions, including eczema, acne, psoriasis, and fungal infections, ensuring a diverse representation of cases encountered in clinical practice. The questions were validated and used in undergraduate and postgraduate medical examinations as case vignettes. A sample case and related prompts to the chatbots are presented in Figure 1.

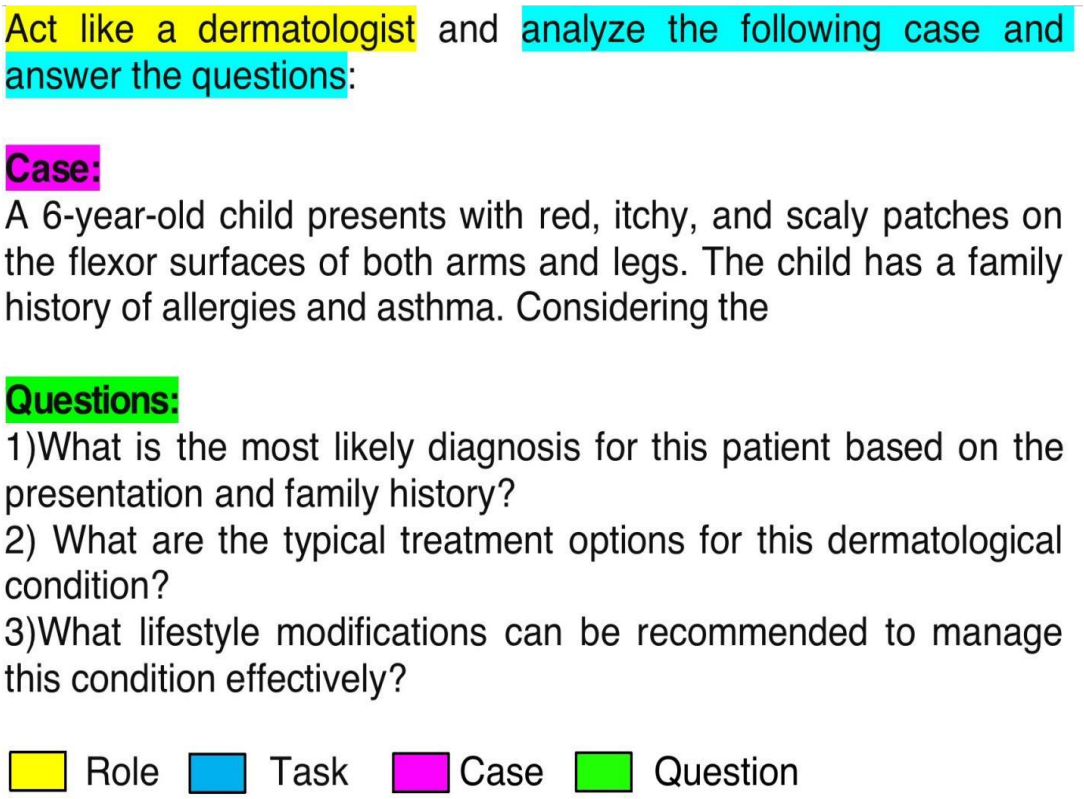

Figure 1: Example of a case and associated questions along with role definition and specification of task

Data Collection

The dermatology cases and associated questions were asked to each of the four AI chatbots. The outputs generated by the chatbots in response to the questions were captured and recorded for further analysis.

Clinical Accuracy Assessment

To evaluate the clinical accuracy of the AI-generated responses, a panel of two experienced dermatologists recruited according to convenience (having >5 years' experience after obtaining a post-graduation degree) independently reviewed and assessed the responses for each case. They used the rating scale presented in Table 2 to provide scores for each question of cases. [16]

Table 2: Scale for scoring artificial intelligence-generated contents for clinical and educational purposes

Score	Level	Analysis
5	Highly accurate	Thoroughly accurate, aligning perfectly with clinical knowledge and best practices
4	Moderately accurate	Mostly accurate, with only minor discrepancies that do not significantly impact its clinical reliability
3	Somewhat accurate	Several inaccuracies that may require clarification or verification by a medical professional
2	Slightly accurate	Noticeable inaccuracies and its clinical reliability is questionable without substantial correction
1	Inaccurate	Fundamentally incorrect and could pose serious risks to patient care if relied upon without thorough review and correction
Adapted from Kumari et al. [13]		

Statistical Analysis

Descriptive statistics were used to summarize the clinical accuracy ratings of the AI-generated responses. The mean score among four AI chatbots was compared by ANOVA with a post-hoc test. The result is presented with effect size. Additionally, the inter-rater reliability coefficient was computed to determine the agreement between the dermatologists' assessments. GraphPad Prism 9.5.0 was used to analyze the data statistically. A p-value <0.05 was considered statistically significant.

Ethical Considerations

Ethical approval for the study was not necessary according to prevalent guidelines. No patient's data were used in the study.

RESULT

The accuracy scores of four AI-based chatbot-generated answers across 22 case vignettes are presented in Table 3. Two raters evaluated the answers provided by ChatGPT, Gemini, Copilot, and Perplexity. The overall score suggests that the score across the program is significantly different (P = 0.01). In post hoc analysis, the score of ChatGPT vs. Copilot and Gemini vs. Copilot was found to be significantly different. The highest score was for Copilot and the lowest was for Gemini.

Table 3: Accuracy score of four artificial intelligence-based chatbot generated answers to 22 case vignettes

	ChatGPT	Gemini	Copilot	Perplexity	P, post hoc significant pair
Rater 1	3.91±0.86	3.58±0.97	4.15±0.66	3.76±0.85	0.08
Rater 2	4.29±0.55	4.14±0.91	4.86±0.22	4.52±0.56	0.006*, ChatGPT vs. Copilot, Gemini vs. Copilot
Overall	4.1±0.61	3.86±0.88	4.51±0.33	4.14±0.64	0.01*, ChatGPT vs. Copilot, Gemini vs. Copilot
*Statistically significant P value of repeated measure ANOVA					

The differences in accuracy between the content generated by the four chatbots were assessed using Cohen's d effect size and are shown in Table 4. The accuracy gap was highest between Gemini vs. Copilot followed by ChatGPT vs. Copilot. The lowest difference was for ChatGPT vs. Perplexity.

Table 4: Effect size of difference between the accuracy of content generated by four chatbots

Pair	Cohen's d
ChatGPT vs. Gemini	0.31
ChatGPT vs. Copilot	0.83
ChatGPT vs. Perplexity	0.06
Gemini vs. Copilot	0.98
Gemini vs. Perplexity	0.37
Copilot vs. Perplexity	0.75
Interpretation of effect size: small effect size, d = 0.2 to 0.51, medium effect size, d = 0.51 to 0.8, and large effect size, d ≥ 0.81 Adapted from Lakens [21]	

We conducted a one-sample t-test to check the difference from the highest achievable score of 5. All the programs showed significantly lower scores as shown in Table 5. When we tested the score again a hypothetical value of 4, we found all of them had a similar score of 4 except the score of Copilot which is similar to 4.5 (discrepancy = 0.007576, 95% CI = -0.1375 to 0.1527, p = 0.9146).

Table 5: Discrepancy of scores of four artificial intelligence-based chatbot generated answers with hypothetical values

		ChatGPT	Gemini	Copilot	Perplexity
Hypothetical 5	Discrepancy	-0.9015	-1.144	-0.4924	-0.8636
	95% CI	-1.174 to -0.6291	-1.532 to -0.7554	-0.6375 to -0.3473	-1.136 to -0.5913
	P	<0.0001*	<0.0001*	<0.0001*	<0.0001*
Hypothetical 4	Discrepancy	0.09848	-0.1439	0.5076	0.1364
	95% CI	-0.1739 to 0.3709	-0.5325 to 0.2446	0.3625 to 0.6527	-0.1360 to 0.4087
	P	0.46	0.45	<0.0001*	0.31
*Statistically significant P value of one-sample t-test					

On average, the Gemini showed the highest level of ICC in score among the questions, followed by Perplexity and ChatGPT. The Copilot showed the lowest level among all (Table 6).

Table 6: Relationship of score of three answers to the questions asked about the case

Rater	Statistics	ChatGPT	Gemini	Copilot	Perplexity
Rater 1	ICC	0.696	0.741	0.668	0.751
	95% CI	0.382 to 0.864	0.473 to 0.884	0.325 to 0.85	0.494 to 0.889
	P	0.001*	<0.0001*	0.001*	<0.0001*
Rater 2	ICC	0.424	0.76	0.468	0.638
	95% CI	-0.171 to 0.743	0.513 to 0.893	-0.281 to 0.779	0.263 to 0.838
	P	0.063	<0.0001*	0.078	0.003*
Overall	ICC	0.539	0.77	0.551	0.712
	95% CI	0.171 to 0.818	0.532 to 0.897	0.087to 0.799	0.414 to 0.871
	P	0.007	<0.0001	0.014	<0.0001
>0.5 = poor reliability, 0.5 to 0.75 = moderately reliability, 0.76 to 0.9 = good reliability, and >0.91 = excellent reliability Adapted from Bobak et al.[22]					

When we compared the raters, we found an average measure ICC = 0.707, 95% CI = 0.552 to 0.808, P <0.0001 that corresponds to a moderate level of reliability.

DISCUSSION

The varying levels of accuracy scores seem to highlight the supremacy of Copilot, which garnered the highest score, in contrast to Gemini, which received the lowest. The underlying reasons behind these discrepancies could stem from the specific algorithms, training data, and linguistic nuances employed by each chatbot, contributing to differential performance across the case vignettes.[17,18] The most substantial difference in accuracy was detected between Gemini and Copilot. Conversely, the smallest disparity in accuracy was observed between ChatGPT and Perplexity, suggesting a relatively more aligned performance between these two chatbots.

All the chatbot programs exhibited scores significantly below the maximum value. When the scores were compared against a hypothetical value of 4, all chatbots, except for Copilot, yielded similar scores of 4. Copilot's score was similar at around 4.5. This analysis underscores the variability in the accuracy levels of chatbot-generated responses. Hence, all of them had an accuracy level above 80% but below 100%. In the context of diagnosing dermatological diseases, this level of accuracy may be helpful for getting relatively accurate information. AI-powered chatbots have the potential to offer valuable assistance to patients in dermatological healthcare. They provide accessible

and prompt information, enabling individuals to gather preliminary insights into various skin conditions, symptoms, and potential treatment options. These chatbots can aid in educating patients about common dermatological issues, offering self-care tips, and providing guidance on when to seek professional medical advice. Moreover, they can help alleviate the burden on healthcare systems by addressing straightforward queries and offering general advice, allowing dermatologists to focus on more complex cases. [19]

However, we also found that the accuracy of chatbot-generated responses, particularly in terms of diagnosing dermatological diseases, might fall short of the desired 100% precision. Furthermore, patients may use the chatbot of their choice and may get different accuracy in different chatbot. Patients and healthcare providers should be mindful of the limitations of these chatbots and consider them as supplementary resources rather than definitive diagnostic tools. The need for continuous improvement and rigorous training of AI models becomes evident, as enhancing their accuracy can contribute to their greater utility in providing accurate and reliable information to aid both patients and medical professionals in the dermatological healthcare domain. [20-22]

The implications of the findings have significant relevance in the realm of dermatological healthcare. While these programs offer a convenient and accessible avenue for information, their inaccuracies may pose potential risks when it comes to accurate disease identification and treatment recommendations. [23] The finding of the study accentuates the necessity for continuous refinement and augmentation of these AI models to achieve higher diagnostic accuracy and reliability, thereby enhancing their potential as supportive tools within dermatological healthcare contexts.

Several limitations should be considered while interpreting the result. We used a set of 22 dermatological case vignettes and two raters rated the clinical accuracy. The cases were formulated with typical presentation of dermatological diseases and may not represent actual clinical cases. The study's cross-sectional evaluation might miss the evolving nature of AI models, which can be subject to frequent updates and improvements that might alter their performance over time.

CONCLUSION

This study sheds light on the potential of AI-powered chatbots in assisting patients with dermatological healthcare information. These tools offer accessible and prompt information, aiding in educating patients about various skin conditions and self-care practices. However, our findings emphasize the importance of cautious interpretation of chatbot-generated responses, particularly in the context of diagnosing dermatological diseases. The observed discrepancies in accuracy levels underscore the limitations of relying solely on these chatbots for precise diagnoses. As such, they should be considered supplementary resources rather than definitive diagnostic tools. Continuous improvement and rigorous training of AI models are essential to enhance their accuracy and reliability in providing valuable information to both patients and medical professionals in the dermatological healthcare field.

REFERENCES

1. Ferreira IG, Weber MB, Bonamigo RR. History of dermatology: the study of skin diseases over the centuries. An Bras Dermatol. 202;96(3):332-345. doi: 10.1016/j.abd.2020.09.006.
2. Prasad S, Bassett IV, Freeman EE. Dermatology on the global stage: The role of dermatologists in international health advocacy and COVID-19 research. Int J Womens Dermatol. 2021;7(5):653-659. doi: 10.1016/j.ijwd.2021.10.003.
3. Seth D, Cheldize K, Brown D, Freeman EF. Global Burden of Skin Disease: Inequities and Innovations. Curr Dermatol Rep. 2017;6(3):204-210. doi: 10.1007/s13671-017-0192-7.
4. Dave M, Patel N. Artificial intelligence in healthcare and education. Br Dent J. 2023;234(10):761-764. doi: 10.1038/s41415-023-5845-2.
5. Patel S, Wang JV, Motaparthi K, Lee JB. Artificial intelligence in dermatology for the clinician. Clin Dermatol. 2021;39(4):667-672. doi: 10.1016/j.clindermatol.2021.03.012.
6. Tudor Car L, Dhinagaran DA, Kyaw BM, Kowatsch T, Joty S, Theng YL, Atun R. Conversational Agents in Health Care: Scoping Review and Conceptual Analysis. J Med Internet Res. 2020;22(8):e17158. doi: 10.2196/17158.
7. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J. 2021;8(2):e188-

e194. doi: 10.7861/fhj.2021-0095.

8. Jin JQ, Dobry AS. ChatGPT for healthcare providers and patients: Practical implications within dermatology. J Am Acad Dermatol. 2023:S0190-9622(23)01106-4. doi: 10.1016/j.jaad.2023.05.081.
9. Kluger N. Potential applications of ChatGPT in dermatology. J Eur Acad Dermatol Venereol. 2023;37(7):e941-e942. doi: 10.1111/jdv.19152.
10. Mondal H, Mondal S, Podder I. Using ChatGPT for Writing Articles for Patients' Education for Dermatological Diseases: A Pilot Study. Indian Dermatol Online J. 2023;14(4):482-486. doi: 10.4103/idoj.idoj_72_23.
11. Agarwal M, Sharma P, Goswami A. Analysing the Applicability of ChatGPT, Gemini, and Copilot to Generate Reasoning-Based Multiple-Choice Questions in Medical Physiology. Cureus. 2023;15(6):e40977. doi: 10.7759/cureus.40977.
12. Meyer, J.G., Urbanowicz, R.J., Martin, P.C.N. et al. ChatGPT and large language models in academia: opportunities and challenges. BioData Mining 2023;16:20. doi: 10.1186/s13040-023-00339-9
13. Kumari A, Kumari A, Singh A, Singh SK, Juhi A, Dhanvijay AD, et al. Large Language Models in Hematology Case Solving: A Comparative Study of ChatGPT-3.5, Google Gemini, and Microsoft Copilot. Cureus 2023;15(8): e43861. doi:10.7759/cureus.43861.
14. Sallam M, Al-Farajat A, Egger J. Envisioning the Future of ChatGPT in Healthcare: Insights and Recommendations from a Systematic Identification of Influential Research and a Call for Papers. Jordan Med J 2024;58(1): 95 – 108. doi: 10.35516/jmj.v58i1.2285
15. Sallam M, Barakat M, Sallam M. A Preliminary Checklist (METRICS) to Standardize the Design and Reporting of Studies on Generative Artificial Intelligence-Based Models in Health Care Education and Practice: Development Study Involving a Literature Review. Interact J Med Res. 2024;13:e54704. doi: 10.2196/54704
16. Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Zadnik Sullivan PL, Cielo D, Oyelese AA, Doberstein CE, Telfeian AE, Gokaslan ZL, Asaad WF. Performance of ChatGPT, GPT-4, and Google Gemini on a Neurosurgery Oral Boards Preparation Question Bank. Neurosurgery. 2023. doi: 10.1227/neu.0000000000002551.
17. Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI Responds to Common Lung Cancer Questions: ChatGPT vs Google Gemini. Radiology. 2023;307(5):e230922. doi: 10.1148/radiol.230922.
18. Fan X, Chao D, Zhang Z, Wang D, Li X, Tian F. Utilization of Self-Diagnosis Health Chatbots in Real-World Settings: Case Study. J Med Internet Res. 2021;23(1):e19928. doi: 10.2196/19928.
19. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. N Engl J Med. 2023;388(13):1233-1239. doi: 10.1056/NEJMsr2214184.
20. Miles O, West R, Nadarzynski T. Health chatbots acceptability moderated by perceived stigma and severity: A cross-sectional survey. Digit Health. 2021;7:20552076211063012. doi: 10.1177/20552076211063012.
21. Nadarzynski T, Miles O, Cowie A, Ridge D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. Digit Health. 2019;5:2055207619871808. doi: 10.1177/2055207619871808.
22. Karabacak M, Margetis K. Embracing Large Language Models for Medical Applications: Opportunities and Challenges. Cureus. 2023;15(5):e39305. doi: 10.7759/cureus.39305.
23. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol. 2013;4:863.
24. Bobak C, Barr P, O'Malley A. Estimation of an inter-rater intra-class correlation coefficient that overcomes common assumption violations in the assessment of health measurement scales. BMC Med Res Methodol 2018;18:93. doi: 10.1186/s12874-018-0550-6.

تقييم استجابات برامج بوت المحادثة القائمة على الذكاء الاصطناعي للاستعلام عن الأمراض الجلدية شائعة الحدوث

إندراسيش بودر[1]، نيها بيبيل[2]، أرونيما دهبال[3]، شيكات موندال[4]، فيتسومينو جوديث بينيي[5]، هيميل موندال[6]

[1] قسم الأمراض الجلدية والتناسلية، كلية الطب ومستشفى ساجور دوتا، ولاية البنغال الغربية، الهند.

[2] قسم علم الأدوية، معهد راجشري للأبحاث الطبية، باريلي، أوتار براديش، الهند.

[3] قسم الأمراض الجلدية، معهد جاغانات غوبتا للعلوم الطبية، ولاية البنغال الغربية، الهند .

[4] قسم علم وظائف الأعضاء، كلية ومستشفى رايغانج الطبي الحكومي، ولاية البنغال الغربية، الهند .

[5] قسم علم وظائف الأعضاء، معهد ناجالاند للعلوم والأبحاث الطبية، كوهيما، ناجالاند، الهند.

[6] قسم علم وظائف الأعضاء، معهد عموم الهند للعلوم الطبية، ديوغار، جهارخاند، الهند.

الملخص

الخلفية والأهداف: يمكن لبرامج بوت المحادثة القائمة على الذكاء الاصطناعي تسهيل الرعاية الصحية من خلال تقديم تفاعلات فورية وشخصية للمرضى، بالإضافة للإجابة على الاستفسارات وتوفير معلومات طبية عامة. إن قدرة هذه البرامج على الكشف المبكر عن الأمراض واقتراح خطط علاجية قد يؤدي إلى تحسن في نتاج الرعاية الصحية للمرضى. هدفت هذه الدراسة إلى التحقق من جدوى استخدام لبرامج بوت المحادثة القائمة على الذكاء الاصطناعي في التعامل مع تحديات التشخيص وتوصيات العلاج للأمراض الجلدية الشائعة.

منهجية الدراسة: تم تجميع مجموعة بيانات تضم 22 حالة من الحالات الجلدية، وكانت كل حالة مصحوبة بثلاثة استفسارات محددة. تم تقديم هذه النماذج القصيرة للحالة إلى أربعة نماذج محادثة متميزة للذكاء الاصطناعي – ChatGPT 3.5، وGoogle Gemini، Microsoft Copilot (GPT 4)، Perplexity.ai، وتم حفظ الردود لتقييم الملاءمة والدقة السريرية. قام اثنان من أطباء الجلد الخبراء بشكل مستقل بتقييم استجابات أنظمة الذكاء الاصطناعي باستخدام مقياس ليكرت المكون من 5 نقاط يتراوح من الدقة العالية (= 5) إلى الأقل دقة (= 1).

النتائج: كان متوسط درجة ChatGPT 4.1 ± 0.61، وكان Gemini 3.86 ± 0.88، وكان Microsoft Copilot 4.51 ± 0.33، وكان Perplexity.ai 4.14 ± 0.64، P = 0.01. كان الفارق الكبير في النتيجة بين Gemini مقابل Microsoft Copilot (كوهين d = 0.98)، ChatGPT مقابل Microsoft Copilot (كوهين d = 0.83)، Microsoft Copilot مقابل Perplexity.ai (كوهين d = 0.75). كانت جميع نتائج برامج بوت المحادثة القائمة على الذكاء الاصطناعي مشابهة لدقة 80% (واحد عينة اختبار t بقيمة افتراضية 4) باستثناء Microsoft Copilot والذي أظهر دقة تقارب 90٪.

الاستنتاجات: تسلط هذه الدراسة الضوء على إمكانات برامج بوت المحادثة القائمة على الذكاء الاصطناعي في مجال الرعاية الصحية الجلدية من أجل تثقيف المرضى. وبالرغم من ذلك، فإن النتائج تؤكد محدوديتها في التشخيص الدقيق للأمراض. يمكن استخدام هذه البرامج كمصادر تكميلية بدلاً من أدوات التشخيص الأولية.

الكلمات الدالة: الذكاء الاصطناعي، أطباء الجلد، محرك البحث، تقديم الرعاية الصحية، الذكاء.