Clinical, Cosmetic and Investigational Dermatology Dovepress

Open Access Full Text Article

SHORT REPORT

AI in Hidradenitis Suppurativa: Expert Evaluation of Patient-Facing Information

Anne-Cécile Ezanno[1], Anne-Claire Fougerousse[2], Christelle Pruvost-Balland[3], François Maccari[4], Charlotte Fite[5] On behalf of ResoVerneuil

[1]Department of Digestive, Surgery, Begin Military Teaching Hospital, Saint Mandé, France; [2]Department of Dermatology, Begin Military Teaching Hospital, Saint Mandé, France; [3]Department of Dermatology, University Hospital Pontchaillou, Rennes, France; [4]Department of Dermatology, Begin Military Teaching Hospital, Saint Mandé and Medical Center, La Varenne Saint-Hilaire, France; [5]Department of Dermatology, Saint Joseph Hospital, Paris, France

Correspondence: Anne-Cécile Ezanno, Department of Digestive, Surgery, Begin Military Teaching Hospital, Service de Chirurgie Viscérale – HIA BEGIN, 69 Avenue de Paris, St Mandé, 94160, France, Tel +33 (0)133985250, Fax: +33 (0)143985922, Email ezanno.annececile@gmail.com

Purpose: This study investigates the accuracy of Artificial Intelligence (AI) chatbots, ChatGPT and Bard, in providing information on Hidradenitis Suppurativa (HS), aiming to explore their potential in assisting HS patients by offering insights into symptoms, thus possibly reducing the diagnostic and treatment time gap.

Patients and Methods: Using questions formulated with the help of HS patient associations, both ChatGPT and Bard were assessed. Responses to these questions were evaluated by 18 hS experts.

Results: ChatGPT's responses were considered accurate in 86% of cases, significantly outperforming Bard, which only achieved 14% accuracy. Despite the general efficacy of ChatGPT in providing relevant information across a range of HS-related queries, both AI systems showed limitations in offering adequate advice on treatments. The study identifies a significant difference in the performance of the two AIs, emphasizing the need for improvement in AI-driven medical advice, particularly regarding treatment options.

Conclusion: The study highlights the potential of AI chatbots, particularly ChatGPT, in supporting HS patients by improving symptom understanding and potentially reducing the time to diagnosis and treatment. AI chatbots, while promising, cannot yet substitute for professional medical diagnosis and treatment, indicating the importance of enhancing AI capabilities for more accurate and reliable medical information dissemination.

Keywords: dermatology, hidradenitis suppurativa, acne inversa, artificial intelligence

Introduction

Hidradenitis suppurativa (HS) is a chronic and inflammatory skin disease characterized by painful abscesses, nodules, and sinus tracts, primarily affecting the axillae, groin, and perianal areas. The global prevalence of HS is estimated to be around 1%, with higher rates in females and individuals with a family history of the condition. HS can lead to significant physical and psychological morbidity, including pain, scarring, and social isolation. Many patients hesitate to share their symptoms with healthcare professionals due to the intimate nature of the affected areas, feelings of embarrassment, and prior negative healthcare experiences, which can result in delayed diagnosis and treatment.[1] In recent years, there has been a growing trend among patients to use Artificial intelligence (AI) to search for health information online. A recent survey found that approximately 30–40% of patients use AI systems like virtual assistants or chatbots for health-related inquiries,[2] highlighting the increasing role of AI in patient self-management and health information gathering. AI emerges as a potential ally for patients, offering a means for them to gain better insights into their symptoms and potentially reducing the considerable time gap between initial symptoms, medical diagnosis, and treatment.[1] AI systems (AIs), such as ChatGPT (Chat Generative Pre-Trained Transformer) developed by OpenAI and released on November 30, 2022, and Bard developed by Google and released on March 21, 2023, are advanced language models capable of

Received: 14 May 2024
Accepted: 26 October 2024
Published: 2 November 2024

providing conversational responses similar to interacting with a human.[3] The deployment of AI chatbots holds considerable promise for elevating patient care in dermatology and contributing to broader public health initiatives.[4]

We report on an innovative study evaluating the accuracy of AIs—ChatGPT and —in providing information on HS. Our study aims to explore the potential of AI chatbots in supporting HS patients by offering insights into symptoms, thus possibly reducing the time to diagnosis and treatment.

Material and Methods

The two most widely used AI systems were selected and studied: Chat GPT (version 3.5) and Bard.[5] Seven questions related to HS were developed with the help of HS patient associations:

Q1: I have abscesses in my armpits and/or perianal folds and/or pubic area all the time. What could it be?

Q2: I have several abscesses at the same time in the groin crease, armpits, perianal folds. What could it be?

Q3: What is HS? And is it a serious disease?

Q4: Is it hereditary?

Q5: Which doctor do I need to see?

Q6: What treatments are available?

Q7: In the meantime, what advice can you give me to improve my situation?

These questions were administered in an individual chat by an investigator (AC E). All questions were asked to both AI systems on the same day (December 20, 2023) in French for practicality.

The ChatGPT and Bard responses were independently assessed by 18 hS experts. All experts independently evaluated all responses using a 5-point Likert scale (1: strongly agree; 2: agree; 3: neutral; 4: disagree; 5: strongly disagree).

Statistical Analysis

We developed a web-based questionnaire Google Form to allow the experts to evaluate the AI responses. All participant responses were exported as an Excel file (Microsoft Corporation) from the Google Form website. An AI-generated response was deemed appropriate if the majority of experts strongly agreed or agreed with it; otherwise, it was classified as inappropriate.

The average points obtained for each question according to the Likert scale were calculated and presented as mean ± standard deviation. A reminder that the Likert scale used ranged from 1 (strongly agree) to 5 (strongly disagree). For comparisons where the data followed a normal distribution, an independent *t*-test was used. However, for comparisons where the data did not meet the normality assumption, the Mann–Whitney *U*-test was applied. Differences in scores between ChatGPT and Bard were specifically analyzed using the Mann–Whitney *U*-test, as the data did not conform to normal distribution. A statistical significance threshold was set at a p-value of < 0.05 for all analyses.

Results

The complete list of questions and answers can be found in Appendix.

Both AI systems responded to all seven questions. ChatGPT provided relatively shorter answers with a mean word count of 228±48 (range 148–292), compared to 254±77 (range 96–354) for Bard.

Responses by ChatGPT to 6 of the 7 queries were deemed "appropriate" (86%) by the experts, whereas the response to one question (14%) was rated inappropriate (Q6 "What treatments are available?"). Responses by Bard to 1 of the 7 queries were deemed "appropriate" (14%) by the experts (Q5 "Which doctor do I need to see?"), whereas responses to 86% (n=6), were rated inappropriate.

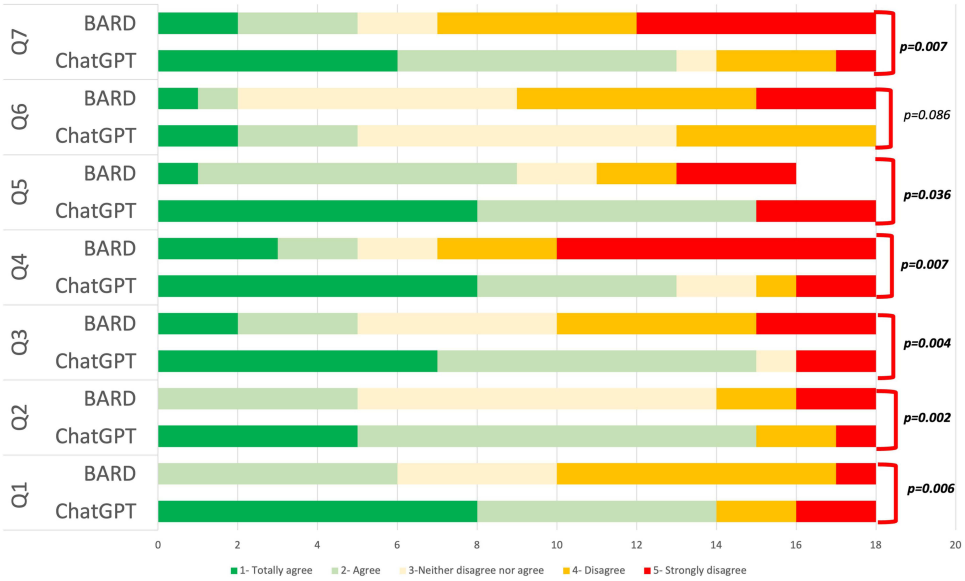

Figure 1 List of the seven queries submitted to ChatGPT and Bard along with their experts' evaluations.

Figure 1 presents the descriptive statistics for the scores given by the 18 participants. The correlation between the scores given by all evaluators for the answers provided by the two AI systems, is presented in Table 1.

A significant difference in mean scores between the two AI was found by all participants, who indicated that ChatGPT was the most accurate. The response to question 6 regarding treatment options was classified as "completely incorrect" when the majority of experts rated it with the lowest possible score on the Likert scale, indicating that they found the answer to be inaccurate or inappropriate. There was no significant difference between the two AIs (p=0.086) in the evaluation of question N°6. However, ChatGPT offered better advice to improve patient quality of life.

Discussion

This study marks a pioneering attempt to assess the capability of AI in HS self-diagnosis. ChatGPT outperformed Bard, offering more precise answers, particularly on symptoms, although both struggled with treatment-related questions.

Our findings suggest a significant difference in the performance of the 2 AIs. ChatGPT was deemed accurate (appropriate responses) in 86% of cases, overshadowing Bard's 14% accuracy rate. Notably, all experts highlighted

Table 1 Comparative Analysis of Scores by Evaluators for Responses from Two AIs

	Score for ChatGPT 3.5, Mean (SD)	Score for Bard, Mean (SD)	P value
Q1	2.1± 1.4	3.2±1	0.006
Q2	2.1±1.1	3.1±0.9	0.002
Q3	2±1.2	3.2±1.2	0.004
Q4	2.1±1.3	3.6±1.5	0.007
Q5	2.1±1.4	2.8±1.3	0.036
Q6	2.0±0.9	3.5±1	0.086
Q7	2.2±1?2	3.6±1.4	0.007

Abbreviation: SD, standard deviation.

ChatGPT's superior performance, marking a stark contrast in reliability and the potential utility in patient support. However, both AI systems fell short in offering satisfactory advice on HS treatment.

AI's ability to provide both accurate and inaccurate HS information has far-reaching implications for patients, healthcare professionals and the healthcare system as a whole. Accurate AI-generated information could empower patients by helping them understand their symptoms earlier and prompting them to consult healthcare professionals in a timely manner, potentially reducing delays in diagnosis. This could have a significant positive impact on patient outcomes, particularly for diseases such as HS, for which early intervention is essential. For healthcare professionals, AI systems could serve as additional tools to improve patient education and help with initial screening, reducing the workload of overburdened clinicians. However, there are serious risks if AI provides inaccurate or misleading information, particularly in the medical context. Inaccurate treatment advice or diagnostic information could delay appropriate care, leading to a deterioration in patients' conditions and increased healthcare costs due to mismanagement or unnecessary interventions.

From a systemic perspective, if AI systems such as ChatGPT or Bard can reliably improve symptom understanding, they could potentially ease pressure on healthcare systems by reducing the need for unnecessary consultations or diagnostic procedures. However, if AI systems frequently provide incorrect information, the resulting delays in diagnosis or treatment could add to the burden on healthcare services, as patients may present with more advanced or complicated cases. It is therefore essential to improve the accuracy of AI responses, so that AI becomes a valuable asset rather than a liability in healthcare.

For Question 6, related to treatment options, 61.1% of the experts rated ChatGPT's response as either "disagree" or "strongly disagree", indicating their dissatisfaction with the accuracy of the information provided. This is also the question where Bard's responses received the lowest ratings: 88.9% of the experts rated Bard's response as either "disagree" or "strongly disagree", indicating a high level of dissatisfaction with the accuracy of the information provided. This point is aligned with findings from a recent systematic review on ChatGPT by Sallam et al.[2] The tendency for incorrect responses may be attributed to the model's primary training on diverse internet content, including articles, books, Wikipedia, news, and websites, rather than sourcing information directly from scientific societies. The responses to treatment-related questions were rated poorly because they often lacked the necessary specificity and accuracy expected from current medical guidelines. These topics tend to be more factual and within the AI systems' capability to provide correct and consistent information, whereas treatment recommendations are more complex and require individualized, evidence-based guidance. We advocate for enhancing the accuracy of responses to medical queries by mandating that AIs consult specific scientific societies instead of relying solely on websites. The challenge stems from the potential variations in recommendations among different societies and countries.

In the initial four questions, ChatGPT outperformed Bard. Notably, in the field of diagnostics, ChatGPT has exhibited its ability in producing satisfactory results in previous studies.[6–9] It is essential to highlight that only one study has demonstrated superiority of ChatGPT over other AIs.[10]

However, the question arises: why is there disparity between the two AI systems? The emergence of AI technology has ushered in dialogue models that are gradually superseding traditional search engines such as Google or Bing. These models, rooted in large language models, employ deep learning techniques to generate natural language texts in various formats, including question and answer, summaries, and narratives, based on user-provided data.[11] The disparity between the two AI systems, ChatGPT and Bard, likely stems from differences in their training data and underlying algorithms. ChatGPT, which is designed to generate creative and coherent responses, excels in producing conversational answers that align well with patient queries about symptoms and general disease information.[12] Bard, on the other hand, focuses on providing comprehensive and informative responses but struggles with the level of precision needed for accurate medical information, particularly in complex areas like treatment recommendations. These differences in approach and data sources explain why ChatGPT performed better in this study, especially in areas requiring nuanced, human-like communication.

Despite these distinctions, when it comes to HS-related questions, this limitation did not seem to hinder ChatGPT's performance, as it consistently outperformed Bard.

Interestingly, neither of the two AIs suggested that patients refer to reputable medical society websites, which offer scientifically validated content overseen by expert committees. Nevertheless, it is essential to underscore that while AI serves as a helpful resource, it cannot substitute a comprehensive medical assessment. Given the complex and varied nature of HS symptoms, only a qualified healthcare provider can ensure an accurate diagnosis. Self-diagnosis based solely on online information, even with AI, carries the risk of misinterpretation.

Limitation

This study has several limitations that should be considered when interpreting the results. First, the sample size of 18 hS experts, while adequate for an exploratory study, may not be fully representative of the broader dermatology community. A larger cohort would provide a more comprehensive range of expert opinions and enhance the validity of the findings. Replicability is a key consideration for future research Second, the use of only 7 predefined questions limits the scope of the AI evaluation, as these questions may not encompass the full range of HS-related patient queries. Future studies should expand the question set to assess AI performance across diverse medical contexts. Third, the reliance on expert evaluations introduces a degree of subjectivity, despite their value. Incorporating more objective measures, such as patient outcomes or alignment with established clinical guidelines, would strengthen the conclusions. Lastly, the cross-sectional design limits our ability to observe changes in AI performance over time, particularly as these systems undergo regular updates and improvements.

A longitudinal approach could also provide valuable insights into how AI performance changes over time and its potential role in medical information dissemination.

Additionally, using AI chatbots in clinical contexts is not yet advisable due to the lack of comprehensive data on their medical expertise. The ethical implications of relying on AI for medical information also warrant caution, as misinformation could have significant consequences for patient care. Finally, these AI systems do not offer clinical imaging capabilities to support diagnostic hypotheses, limiting their utility in more complex medical scenarios.

Conclusion

In conclusion, the comparison between ChatGPT and Google Bard in responding to 7 specific questions about HS suggests that AI chatbots may have the potential to assist in providing information and hopefully reducing diagnostic delays. ChatGPT demonstrated a higher accuracy rate compared to Bard, particularly in providing symptom-related information, but both systems showed significant limitations in addressing treatment-related questions. While these findings are encouraging, they are limited to HS and cannot be generalized to other medical conditions or broader healthcare contexts. Additionally, AIs tools should not be considered replacements for professional medical consultation. Further research is needed to evaluate AI's performance across a wider range of medical conditions and questions before drawing broader conclusions about its role in healthcare. An ethical and legal framework for AIs should be collaboratively established, as the dissemination of misinformation can pose detrimental consequences for wandering patients.

Acknowledgments

All the authors would like to thank Hélène Raynal, the RESO association and all experts: Aude Nassif, Pierre-André Bécherel, Clement Zimmermann, Philippe Guillem, Jeremy Gottlieb, Caroline Jacobzone, Jean-Luc Perrot, Ziad Riguai, Germaine Gabison and Marie Lamiaux.

Disclosure

Dr Anne-Claire Fougerousse reports personal fees, non-financial support from AbbVie, Lilly, Janssen, Leo Pharma, Ucb pharma and Sanofi; personal fees from Novartis, Ammirall, Pfizer, Boehringer Ingelheim and BMS, outside the submitted work. The authors report no conflicts of interest in this work.

References

1. Loget J, Saint-Martin C, Guillem P, et al. Misdiagnosis of hidradenitis suppurativa continues to be a major issue. The R-ENS Verneuil study. *Ann Dermatol Venereol*. 2018;145(5):331–338. doi:10.1016/j.annder.2018.01.043

2. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare*. 2023;11(6):887. doi:10.3390/healthcare11060887
3. Nobles AL, Leas EC, Caputi TL, Zhu SH, Strathdee SA, Ayers JW. Responses to addiction help-seeking from Alexa, Siri, Google Assistant, Cortana, and Bixby intelligent virtual assistants. *NPJ Digit Med*. 2020;3(1):11. doi:10.1038/s41746-019-0215-9
4. Li J. Security implications of AI chatbots in health care. *J Med Internet Res*. 2023;25:e47551. doi:10.2196/47551
5. AI Industry Analysis: 50 most visited AI tools and their 24B+ traffic behavior - writerbuddy [internet]. Available from: https://writerbuddy.ai/blog/ai-industry-analysis. Accessed January 7, 2024.
6. Berg HT, van Bakel B, van de Wouw L, et al. ChatGPT and generating a differential diagnosis early in an emergency department presentation. *Ann Emerg Med*. 2024;83(1):83–86. doi:10.1016/j.annemergmed.2023.08.003
7. Kuroiwa T, Sarcon A, Ibara T, et al. The potential of ChatGPT as a self-diagnostic tool in common orthopedic diseases: exploratory study. *J Med Internet Res [Internet]*. 2023;25(1):e47621. doi:10.2196/47621
8. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. *Int J Environ Res Public Health*. 2023;20(4):3378. doi:10.3390/ijerph20043378
9. Johnson D, Goodman R, Patrinely J, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. *Res Sq*. 2023;rs.3.rs–2566942.
10. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing Chat in supporting evidence-based dentistry: comparative mixed methods study. *J Med Internet Res*. 2023;25:e51580. doi:10.2196/51580
11. Watkins R. Guidance for researchers and peer-reviewers on the ethical use of Large Language Models (LLMs) in scientific research workflows. *AI Ethics [Internet]*. 2023. [cited 2024 Jan 11]. doi:10.1007/s43681-023-00294-5.
12. Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI responds to common lung cancer questions: chatGPT vs Google Bard. *Radiology*. 2023;307(5):e230922. doi:10.1148/radiol.230922