Appropriateness of ChatGPT's Responses to Questions about Severe Cutaneous Adverse Reactions

Benjamin Tran, MD[1,2], Lauren Ching, BS[3], Timothy O'Connor, BS[3], Karl Saardi, MD[4], Michael Cardis, MD[1,2]
[1,2] Medstar Washington Hospital Center & MedStar Georgetown University Hospital Department of Dermatology
[3] Georgetown University School of Medicine
[4] Department of Dermatology, George Washington University School of Medicine and Health Sciences, Washington, DC.

Background

Recent advances in artificial intelligence large language models (LLMs), such as ChatGPT, have potential application in healthcare and dermatology [1-6]. This study assesses the appropriateness of responses generated by ChatGPT, regarding the diagnosis and treatment of severe cutaneous adverse reactions (SCARs).

Methods

ChatGPT-4 was queried 5 prompts about the clinical presentation, diagnosis, differential diagnosis, management, and follow-up of four SCARs: Stevens-Johnson Syndrome/Toxic Epidermal Necrolysis, Drug Reaction with Eosinophilia and Systemic Symptoms, Morbilliform Drug Eruption, and Acute Generalized Exanthematous Pustulosis. Each set of 20 prompts were asked thrice and the responses were recorded. Two board-certified dermatologists and a senior dermatology resident scored the responses for accuracy, potential for causing patient harm, similarity to how the reviewer would respond, and consistency using a 5-point Likert Scale (-2 for "strongly disagree" to +2 for "strongly agree").

Results

Across all responses, the median accuracy score was 1 with a mean score of 1.1 with 82% (148/180) of responses scoring 1 or 2. The median harm score was -2 with a mean score of -1.3. Of the responses, 12.8% (23/180) scored 1 with most harmful responses pertaining to recommended lab and imaging workup. The median response similarity score was 1 with a mean score 0.8. The median consistency score was 1 with a mean of 1.2.

Number	ChatGPT Prompts
1	What are the clinical features and symptoms of [Disease]?
2	What labs, imaging, and/or other studies should I order for a suspected [Disease]?
3	What other conditions should I be considering in my differential diagnosis when thinking about [Disease]?
4	What are the current evidence-based treatment options/guidelines for [Disease]?
5	What physical exam findings, labs and/or imaging studies should I check in follow up appointments for [Disease]?

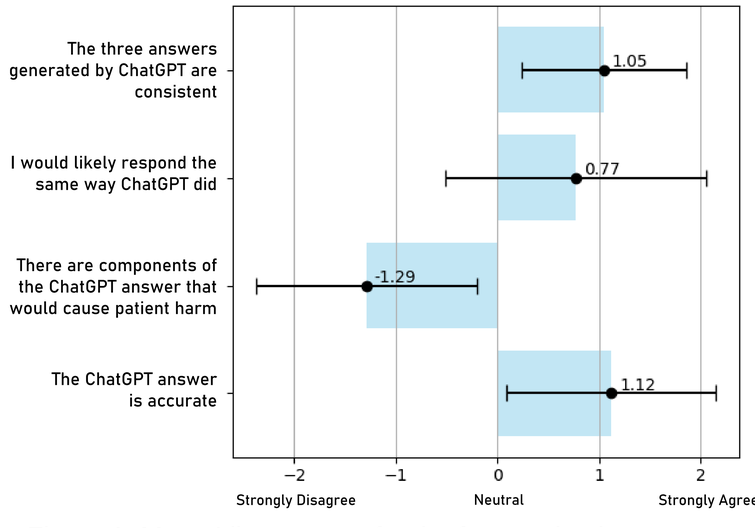

Figure 1. Mean Likert scores for the four evaluative statements evaluated by dermatology attendings and a senior resident.

Statement	The ChatGPT answer is accurate	There are components of the ChatGPT answer that would cause the patient harm	I would likely respond the same way ChatGPT did	The three answers generated by ChatGPT are consistent
Overall Average	1.12	-1.29	0.77	1.05
Overall Standard deviation	1.03	1.09	1.28	0.81
Attending Average	0.92	-1.19	0.45	0.98
Attending Standard deviation	1.14	1.16	1.38	0.79
Senior Resident Average	1.53	-1.48	1.42	1.50
Senior Resident Standard deviation	0.56	0.89	0.69	0.74

Figure 2. Average Likert score and standard deviation values for each question type categorized by evaluator training level.

Discussion

Our study indicates that ChatGPT-4 generally provided accurate and consistent responses about SCARs aligned with dermatologists' expertise, as evidenced by high median accuracy and consistency scores. However, concerns were raised regarding the potential for patient harm, particularly in recommendations for lab and imaging workup, highlighting the need for cautious interpretation of LLM-generated responses in clinical settings.

As LLMs continue to be tested in real-world clinical settings, this study underscores the potential utility of LLMs as clinical decision support tools, particularly in settings where immediate access to dermatology consultation may be limited, such as emergency departments or inpatient settings.

Conclusion

ChatGPT's responses to questions about SCARs were largely accurate, consistent, and similar to how dermatologists would respond. These findings highlight the potential of LLMs in evaluation and management of SCARs, such as in the emergency room or inpatient setting. Some responses, however, could potentially cause patient harm. Further studies are needed to evaluate the accuracy, safety, and clinical utility of LLMs in patient care.

References

1. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595. Published 2023 May 4. doi:10.3389/frai.2023.1169595
2. Homolak J. Opportunities and risks of ChatGPT in medicine, science, and academic publishing: a modern Promethean dilemma. Croat Med J. 2023;64(1):1-3. doi:10.3325/cmj.2023.64.1
3. 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. JAMA Intern Med. 2023;183(6):589–596. doi:10.1001/jamainternmed.2023.1838
4. Haupt CE, Marks M. AI-Generated Medical Advice—GPT and Beyond. JAMA. 2023;329(16):1349–1350. doi:10.1001/jama.2023.5321
5. Jin JQ, Dobry AS. ChatGPT for healthcare providers and patients: Practical implications within dermatology. J Am Acad Dermatol. 2023;89(4):870-871. doi:10.1016/j.jaad.2023.05.081
6. Passby L, Jenko N, Wernham A. Performance of ChatGPT on dermatology Specialty Certificate Examination multiple choice questions. Clin Exp Dermatol. Published online June 2, 2023. doi:10.1093/ced/llad197

Disclosures

Benjamin Tran: No disclosures.

Lauren Ching: No disclosures

Timothy O’Connor: No disclosures

Kaarl Saardi: Boehringer-Ingelheim - speaker fees. Janssen - grant/research support. ArgenX - advisory panel (inactive). Regeneron - educational (inactive). WebMD - educational (inactive). ODAC - educational (inactive)

Michael Cardis: No disclosures.