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 Nekrolysis, 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.

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

Disclosures

Benjamin Tran: No disclosures.
Lauren Ching: No disclosures
Timothy O'Connor: No disclosures
Michael Cardis: No disclosures.