doi:10.36849/JDD.10000

Assessing Accuracy of ChatGPT in Identifying Contact Allergens in Commonly Prescribed Topical Therapies for Psoriasis

Sabrina Fong BA,[a] Alexandra Flamm MD,[b] John G. Zampella MD[a]

[a]Department of Dermatology, NYU Langone Health, Preston Robert Tisch Center for Men's Health, New York, NY

[b]Department of Dermatology, NYU Langone Health, NYU Langone Ambulatory Care Center, New York, NY

To the Editor,

Eight percent (8%) of adults are affected by contact allergies.[1] Patch testing is the gold standard for diagnosis of allergic contact dermatitis (ACD). Results are difficult to remember and identify due to scientific nomenclature, risking re-exposure and exacerbation of ACD.[2] Current practices utilize printouts and ingredient lists, which are cumbersome for patients, while limited capacity of providers to assist with allergen identification further complicates management. Calls for methods to improve allergen identification highlight an opportunity to improve outcomes.[2] As large language models (LLMs) like ChatGPT become increasingly utilized in medicine, these platforms may be useful in identifying allergen-containing products.[3] Studies assessing their accuracy have not been performed.

Herein, 90 allergens from the American Contact Dermatitis Society Core Allergen Series were assessed across 10 topical psoriasis medications, totalling 22 distinct treatments across formulations and manufacturers. Using freely available ChatGPT-4o, each treatment was queried for every allergen as follows: "Does [topical medication] from [manufacturer] contain [allergen]?" Responses were cross-referenced with manufacturer ingredient lists.

Across 1,980 queries, responses were analysed: true positives and true negatives represented correct identifications of presence or absence, respectively. False positives occurred when absent allergens were incorrectly identified as present, and false negatives when present allergens were missed. There were 39 true positives, 1929 true negatives, 3 false positives, and 9 false negatives, yielding a sensitivity of 0.813, specificity of 0.998, and overall accuracy of 99.3%. Negative predictive value was 0.995 and negative likelihood ratio was 0.187. Strong precision and recall were reflected in an F1 score of 0.867, with a *P*-value of 0.864, indicating substantial interrater agreement with referenced ingredient lists.

We report a method whereby allergens are accurately identified in products in 99.3% of cases. The only currently published metric for assessing successful identification and avoidance of patch test-positive allergens utilizes time-consuming calls to determine patch-test-positive patients' resolved/improved ratio, reported at 85%.[4] In contrast, LLMs provide answers in seconds.[5]

Of 12 erroneous responses, six were errors where allergens were either missed or falsely detected. Two errors resulted from ChatGPT's utilization of incorrect manufacturer labels. Fougera's hydrocortisone 2.5% cream was incorrectly detected to contain propylene glycol and sorbic acid. However, we identified that ChatGPT referenced the ingredient list of Crown Laboratories hydrocortisone, which contains propylene glycol and sorbic acid.

ChatGPT failed to detect cetyl steryl alcohol in five products. In manufacturer ingredient lists, cetyl steryl alcohol appeared under cetyl alcohol and stearyl alcohol, cetostearyl alcohol, and cetearyl alcohol. Similarly, D/L-α-tocopherol appears as all-rac-alpha-tocopherol in Mayne Pharma's calcipotriene foam. ChatGPT flagged the presence of D/L-α-tocopherol in MC2 Therapeutics' calcipotriene-betamethasone cream despite containing α-tocopherol. These errors highlight inconsistencies in nomenclature between patch-testing resources and manufacturer ingredient lists.

Novel methods for identifying allergens in products are needed.[2] Our findings suggest that ChatGPT may serve as a reliable preliminary tool in assessing allergen presence in topical medications. Key limitations are attributable to nomenclature differences between manufacturer ingredient lists and patch testing kits. Additional studies must validate these findings across other product categories. This preliminary study suggests that LLMs may be a useful tool for clinicians providing care to patients with contact allergies.

DISCLOSURES

Dr. Zampella has received honoraria from Veradermics, Dermavant, Janssen, Merck, Arcutis, and Ferndale pharmaceuticals, all of which are unrelated to the present submission. Sabrina and Dr. Flamm have no conflicts of interest to disclose.

REFERENCES

1. InformedHealth.org. Overview: allergic contact dermatitis. Accessed March 20, 2026. https://www.ncbi.nlm.nih.gov/books/NBK447113/
2. Özkaya E, Keskinkaya Z. Need for new strategies to improve the recall and avoidance rates of contact allergens: A retrospective cohort study from Turkey. *Contact Dermatitis*. 2024;90(1):41-50. doi:10.1111/cod.14427
3. Chen SF, Alyakin A, Seas A, et al. LLM-assisted systematic review of large language models in clinical medicine. *Nature Medicine*. doi:https://doi.org/10.1038/s41591-026-04229-5
4. Gallo R, Baldari M, Fausti V, et al. Measurement of a possible patch-testing outcome indicator. *Contact Dermatitis*. 2010;62(3):150-156. doi:10.1111/j.1600-0536.2009.01655.x
5. Croxford E, Gao Y, First E, et al. Evaluating clinical AI summaries with large language models as judges. *NPJ Digit Med*. 2025;8(1):640. doi:10.1038/s41746-025-02005-2

AUTHOR CORRESPONDENCE

John G. Zampella MD
E-mail:..... john.zampella@nyulangone.org