SKIN

RESEARCH LETTER

A Comparison of ChatGPT-4 Vision's Diagnostic Accuracy for Inpatient Skin Conditions in White and Skin of Color Patients

Joseph McGrath BS[1,4*], Evelyn Fagan BS[2,4*], Graham Grisedale BS[3], Jesse Hirner MD MS[4†], and Erin X. Wei MD[4†]

[1] University of Minnesota Medical School; Minneapolis, MN, USA
[2] Mercer University School of Medicine; Savannah, GA, USA
[3] Creighton University School of Medicine; Omaha, NE, USA
[4] University of Nebraska Medical Center Department of Dermatology; Omaha, NE, USA
[*] equal contribution
[†] equal contribution

INTRODUCTION

There is growing interest in the use of ChatGPT-4's vision (ChatGPT-4V) feature for dermatologic diagnoses.[1] ChatGPT-4 is an artificial intelligence tool that uses neural networks to synthesize human-like text in response to queries, while ChatGPT-4V has the added ability to interpret images.[2] Although not designed to make dermatologic diagnoses, it has shown an ability to do so in previous studies.[3] Given the shortage of inpatient dermatologists in the U.S., the present study focuses on ChatGPT-4V's ability to diagnose the most common inpatient skin conditions, recognizing that this could be an advantageous tool for inpatient providers.[4] It also aims to compare ChatGPT-4V's diagnostic accuracy among white and skin of color (SOC) individuals, as a previous study found it performed significantly worse on SOC images.[3]

METHODS

Five white and five SOC images were selected for the fifteen inpatient dermatologic conditions with the most admissions in the U.S.[5] SOC was defined as Fitzpatrick phototypes IV-VI. Due to the broad nature of some conditions, several sub-categories were introduced (e.g. bullous disease included bullous pemphigoid and pemphigus vulgaris, **Table 1**). The images were obtained from VisualDx, dermatology journals (*JAAD* and *JAMA Dermatology)*, and the textbook, *Dermatology for Skin of Color*. Each pair of white and SOC images were matched by type, location, and severity. Two board-certified dermatologists with 5+ years of clinical experience post-residency confirmed that each image was diagnostic for the condition. Images were placed into ChatGPT-4V with the prompt: "Provide the three most common differential diagnoses and one primary diagnosis for the image." The accuracy of ChatGPT-4V was assessed on its ability to correctly identify the condition as the primary diagnosis or within its top three differential diagnoses (ddx). ChatGPT-4V's diagnostic accuracy was compared between white and SOC images using Chi-Square tests with Yates correction in SPSS.

Table 1: ChatGPT Diagnostic Accuracy for Inpatient Skin Condition in White and SOC Patients

Condition	White		SOC	
	Primary Dx	Primary or DDx	Primary Dx	Primary or DDx
Bacterial infection (cellulitis/erysipelas, necrotizing fasciitis)	3/5	4/5	2/5	4/5
Ulcer	5/5	5/5	2/5	5/5
Connective tissue disease (lupus, dermatomyositis)	3/5	5/5	4/5	5/5
Viral infection (HSV 2, herpes zoster)	4/5	5/5	4/5	5/5
Drug eruption (fixed, morbilliform)	1/5	4/5	0/5	1/5
Benign lesion (cyst)	4/5	4/5	2/5	3/5
Non-melanoma skin cancer (basal cell carcinoma, squamous cell carcinoma)	2/5	3/5	2/5	4/5
Hidradenitis suppurativa	5/5	5/5	5/5	5/5
Bullous disease (bullous pemphigoid/ pemphigus vulgaris)	3/5	4/5	2/5	3/5
Melanoma	4/5	4/5	4/5	5/5
Contact dermatitis	2/5	5/5	0/5	2/5
Psoriasis	5/5	5/5	4/5	4/5
Cutaneous lymphoma (mycoses fungoides)	0/5	0/5	0/5	0/5
Urticaria	1/5	1/5	0/5	0/5
Fungal infections (sporothrix, candida, mucormycoses)	1/5	1/5	1/5	1/5
Total Accuracy	43/75 (57.3%)	55/75 (73.3%)	32/75 (42.7%)	47/75 (62.7%)

Dx= diagnosis; DDx= differential diagnosis

RESULTS

ChatGPT-4V analyzed 75 white and 75 SOC images. The primary diagnostic accuracy was greater for white images (43/75, 57.3%) than SOC images (32/75, 42.7%), though this difference was not statistically significant (P=0.103). ChatGPT-4V showed greater accuracy when examining the top three ddx in both white (55/75, 73.3%) and SOC images (47/75, 62.7%), and the disparity in

accuracy across skin types was not statistically significant (P=.221). Primary diagnostic and ddx accuracy varied across conditions, with poor performance noted for cutaneous lymphoma, urticaria, and fungal infections in both skin types (**Table 1**).

CONCLUSION

This study found ChatGPT-4V's diagnostic performance to be poor for white and SOC images, as total accuracy for primary diagnosis or ddx did not exceed 75%. This was markedly worse than the >89% ddx accuracy noted in a previous study assessing ChatGPT-4V's ability to diagnose the most prevalent dermatologic conditions worldwide.[3] The results suggest ChatGPT-4V cannot be used as a reliable diagnostic tool by inpatient providers. However, ChatGPT-4V may have limited utility for certain conditions where it displayed 100 percent accuracy in both white and SOC images. While the difference in ChatGPT-4V's accuracy between white and SOC images was not statistically significant, it did correctly identify a larger proportion of conditions on white skin. This may be due to the higher proportion of white individuals presented in training sets available to ChatGPT and erythema being harder to appreciate in SOC.[6] Future versions of ChatGPT-Vision will likely continue to improve, serving as increasingly valuable tools in dermatologic practice. Rather than replacing clinicians, such technologies have the potential to support diagnostic reasoning and efficiency, while medical providers remain essential for contextual interpretation and communication with the patient. Additional studies will be needed to assess the evolution of ChatGPT's diagnostic performance and how this performance may be affected by the presentation of dermoscopy or dermatopathology images alongside clinical images.

Conflict of Interest Disclosures: None

Funding: None

Corresponding Author:
Erin X. Wei, MD
985645 Nebraska Medical Center
Omaha, NE 68198-5645
Phone: (402) 559-3825
Email: ebarrett@unmc.edu

References:

1. Goktas P, Grzybowski A. Assessing the impact of ChatGPT in dermatology: a comprehensive rapid review. J Clin Med [Internet]. 2024 Oct 3 [cited 2025 Apr 28];13(19):5909. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11477344/
2. Ray PP. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*. 2023;3:121-154. doi:10.1016/j.iotcps.2023.04.003
3. Lau CB, Kwa M, Shen L, Smith GP. Assessing the diagnostic performance of ChatGPT in dermatology across Fitzpatrick phototypes and skin of color. Journal of the American Academy of Dermatology [Internet]. 2025 Mar [cited 2025 Apr 28];92(3):578–9. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0190962224030536
4. Hydol-Smith JA, Gallardo MA, Korman A, Madigan L, Shearer S, Nelson C, et al. The United States dermatology inpatient workforce between 2013 and 2019: a Medicare analysis reveals contraction of the workforce and vast access deserts—a cross-sectional analysis. Arch Dermatol Res [Internet]. 2024 [cited 2025 Apr 28];316(4):103. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10940353/
5. Arnold JD, Yoon S, Kirkorian AY. The national burden of inpatient dermatology in adults. Journal of the American Academy of Dermatology [Internet]. 2019 Oct [cited 2025 Apr 28];81(4):AB285. Available from:

https://linkinghub.elsevier.com/retrieve/pii/S0190962219322352

6. Perlman KL, Williams NM, Egbeto IA, Gao DX, Siddiquee N, Park JH. Skin of color lacks representation in medical student resources: A cross-sectional study. International Journal of Women's Dermatology [Internet]. 2021 Mar [cited 2025 Apr 28];7(2):195–6. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2352647521000010