

ChatGPT Exhibits Reduced Sensitivity in Identifying Concern for Malignancy in Darker Skin Tones

Nicholas Schell MCIT,^a Simona A. Alomary BA,^{b,c} Nicole Baker BS,^{c,d} Jina Chung MD,^c
Temitayo Ogunleye MD,^c Susan C. Taylor MD^c

^aPerelman School of Medicine, University of Pennsylvania, Philadelphia, PA

^bRutgers New Jersey Medical School, Newark, NJ

^cDepartment of Dermatology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA

^dSidney Kimmel Medical College, Philadelphia, PA

To the Editor,

As artificial intelligence (AI) increases in popularity, understanding the nature of health information it provides to patients at home is paramount. ChatGPT's image upload feature lends itself in particular to dermatologic questions, given the specialty's reliance on visual assessment of skin lesions. Understanding the performance of AI tools like ChatGPT in answering common dermatologic questions is important for evaluating how patients conceptualize and make decisions about their skin health at home. This study investigates the ability of ChatGPT, a common AI tool, to identify concern for dermatologic malignancy in lesions across various skin tones.

656 images of biopsy-confirmed lesions were collected from the Stanford Diverse Dermatology Images (DDI) dataset, a pathologically confirmed and expertly curated clinical image dataset described by Daneshjou et al.¹ Use of the dataset for this study was conducted with permission from the DDI study team. The images were characterized by skin tone: "light" (Fitzpatrick I/II), "medium" (Fitzpatrick III/IV), or "dark" (Fitzpatrick V/VI), and malignancy: "benign" or "malignant." For each image, ChatGPT was interrogated with the question "Is this image concerning for malignancy?" Response accuracy was evaluated, and statistics were calculated using chi-squared.

TABLE 1.

Dataset Characteristics by Fitzpatrick Skin Type				
	Fitzpatrick I/II	Fitzpatrick III/IV	Fitzpatrick V/VI	Total
Benign	159 (76.4%)	167 (69.3%)	159 (76.8%)	485
Malignant	49 (23.6%)	74 (30.7%)	48 (23.2%)	171
Total	208	241	207	656

ChatGPT's overall response accuracy was similar across all skin tones (58.7% correct, 37.0% incorrect, 4.3% unknown). For patients with confirmed malignancy, sensitivity was stratified by skin tone. ChatGPT correctly identified malignancy in 47.9% light, 65.3% medium, and 28.3% dark skin lesions, however sensitivity differed across skin tones ($P=0.0004$). On the other hand, specificity was similar across skin tones (63.3% in light, 79.8% in medium, 69.7% in dark skin; $P=0.38$).

These findings demonstrate that ChatGPT performs poorly at identifying concern for malignancy in images across all skin tones, but exhibits significantly reduced sensitivity on darker skin. Several factors may play a role in these results, including potential difficulty recognizing common dermatologic findings associated with malignancy, such as erythema, vascular changes, or variations in pigmentation. Prior work has

TABLE 2.

Dataset Characteristics by Fitz Diagnostic Performance of ChatGPT by Fitzpatrick Skin Type					
		Fitzpatrick I/II	Fitzpatrick III/IV	Fitzpatrick V/VI	Total
Benign	True Negative	93 (44.7%)	103 (42.7%)	106 (51.2%)	302 (46.0%)
	False Positive	54 (26.0%)	60 (28.9%)	46 (22.2%)	160 (24.4%)
Malignant	True Positive	23 (11.0%)	47 (19.5%)	13 (6.3%)	83 (12.7%)
	False Negative	25 (12.0%)	25 (10.4%)	33 (15.9%)	83 (12.7%)
Sensitivity	--	47.9%	65.3%	28.3%	50.0%
Specificity	--	63.3%	79.8%	69.7%	65.4%

demonstrated reduced diagnostic accuracy of ChatGPT and other AI systems in skin of color when compared to lighter skin tones.^{2,3} This suggests that there may be broader limitations, such as limited representation of darker skin tones in the images used to develop these systems.

Understanding AI performance remains critical to effective patient counseling, particularly as patients increasingly rely on these tools to self-diagnose or determine whether to seek medical care.^{4,5} Given variability in performance, AI-generated assessments should not replace clinical evaluation, and individuals with lesions of concern should always seek timely evaluation by a board-certified dermatologist. This study is limited by the number of images available (656 total images) and the distribution of lesion type (485 benign (73.9%) and 171 malignant (26.1%)). Future studies should aim to evaluate AI performance for the evaluation of malignant versus benign lesions using larger and more diverse datasets.

DISCLOSURES

Dr Taylor has served as a consultant, advisory board member, and/or speaker for AbbVie, Arcutis, Armis Scientific, Avita, Beiersdorf, Biorez, Bristol-Myers Squibb, Cara Therapeutics, Dior, Eli Lilly, EPI Health, Evolus, Galderma, GloGetter, Hugel America, Incyte, Johnson & Johnson, L'Oreal USA, MedScape, MJH LifeSciences, Pfizer, Piction Health, Sanofi, Scientis US, UCB, and Vichy Laboratories. Dr. Ogunleye has served as a consultant, advisory board member, and/or speaker for Beiersdorf, MJH LifeSciences, Dermatology Times, and Health Central. All other authors have no conflicts of interest to disclose.

Use of Artificial Intelligence in Manuscript: By the nature of the methods of this study, AI was used to generate data (this study assesses the accuracy of AI in detecting malignant lesions across various skin tones). AI was not utilized in the writing of the manuscript.=

REFERENCES

1. Daneshjou R, Vodrahalli K, Novoa RA, et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Sci Adv*. 2022;8(32):eabq6147. doi: 10.1126/sciadv.abq6147
2. Lau CB, Kwa M, Shen L, et al. Assessing the diagnostic performance of ChatGPT in dermatology across Fitzpatrick phototypes and skin of color. *J Am Acad Dermatol*. 2025;92(3):578-579. doi: 10.1016/j.jaad.2024.10.038
3. Dowie T. Exploring the diagnostic capability of artificial intelligence in dermatology for darker skin tones: a narrative review. *Cureus*. 2025;17(10):e94909. doi: 10.7759/cureus.94909
4. Geracitano J, Anderson B, Coffel M, et al. The accuracy of ChatGPT in answering FAQs, making clinical recommendations, and categorizing patient symptoms: a literature review. *VXUL J*. 2025;VXUL2925. doi:10.63116/VXUL2925
5. Montazeri M, Galavi Z, Ahmadian L. What are the applications of ChatGPT in healthcare: gain or loss? *Health Sci Rep*. 2024;7(2):e1878. doi: 10.1002/hsr2.1878

AUTHOR CORRESPONDENCE

Susan C. Taylor MD

E-mail:..... Susan.Taylor@pennmedicine.upenn.edu