SHORT COMMUNICATION

Impact of Large Language Model–assisted Differential Diagnosis on Clinical Decision-making in Dermatology: A Feasibility Study Using ChatGPT-5

Yuto YAMAMURA, Kazuyasu FUJII*, Chisa NAKASHIMA and Atsushi OTSUKA
*Department of Dermatology, Faculty of Medicine, Kindai University Hospital, Osaka-Sayama, Japan. *Email: kazuyasu.fujii@med.kindai.ac.jp*
Submitted Nov 13, 2025. Accepted after revision Apr 16, 2026
Published May 11, 2026. DOI: 10.2340/actadv.v106.adv-2025-0180 Acta Derm Venereol 2026; 106: adv-2025-0180.

Recent advances in artificial intelligence (AI) have led to substantial progress in diagnostic support research in dermatology. In particular, deep learning-based image analysis has achieved dermatologist-level performance for various skin diseases, including cutaneous tumours, and has been shown to improve diagnostic accuracy among less experienced clinicians (1–3).

More recently, large language models (LLMs) have attracted attention as diagnostic support tools capable of generating clinically relevant differential diagnoses from textual case descriptions(4). In our previous study, ChatGPT-4o demonstrated diagnostic accuracy comparable to that of board-certified dermatologists when provided with structured dermatological case information (5), suggesting its potential utility as a supportive aid in clinical decision-making.

Despite these advances, how junior dermatology residents interact with AI-generated diagnostic suggestions in real-world settings remains insufficiently understood. Trainee physicians may either over-rely on or underutilize AI assistance, and incorrect AI outputs may negatively influence clinical decisions (6–8). We therefore conducted a feasibility study to evaluate the impact of LLM-assisted differential diagnosis on diagnostic accuracy among junior dermatology residents.

MATERIALS AND METHODS

A total of 30 clinical cases (15 neoplastic and 15 inflammatory) were selected from the peer-reviewed Japanese dermatology journal *Hifu no Kagaku*. All cases had established final diagnoses and were selected as in our previous study (5). Because these cases were derived from published literature, prior exposure of the language model cannot be excluded.Twenty-three junior dermatology residents who were actively engaged in clinical dermatology practice participated in the study. None were board-certified dermatologists at the time of participation.

The study employed a 2-phase diagnostic design. Participants first evaluated each case independently based on the provided clinical information and recorded an initial diagnosis without AI support. They then reassessed the same cases with reference to 3 differential diagnoses generated by ChatGPT-5, following the identical case order (**Fig. 1**A).

For each case, only information available at an initial dermatology consultation – including patient demographics and clinical features of the lesions – was provided to ChatGPT-5. The model generated three differential diagnoses with brief comments. Details of the standardized prompting procedure are provided in the Appendix S1.

Statistical analyses were performed using R (version 4.3.0; R Foundation for Statistical Computing, Vienna, Austria). The Wilcoxon signed-rank test was used to compare diagnostic accuracy with and without AI support, with $p<0.05$ considered statistically significant.

RESULTS

ChatGPT-5 included the correct diagnosis within its top 3 differential suggestions in 56.7% of cases. Among the 23 junior dermatology residents, the median diagnostic accuracy increased from 43.3% without AI support to 50.0% with AI support, showing a significant improvement with AI assistance (Wilcoxon signed-rank test, $p<0.0001$; Fig. 1B). For reference, the median diagnostic accuracy of 11 board-certified dermatologists in our previous study was 66.7%.

At the case level, diagnostic changes with and without AI support were analyzed across all 690 individual responses (30 cases × 23 residents). Of these, 380 (55.1%) diagnoses remained unchanged, whereas 310 (44.9%) differed between conditions (**Fig. 2**). Among the changed diagnoses, 267 (86.1%) adopted one of the three differential diagnoses suggested by ChatGPT-5, while 43 (13.9%) selected diagnoses not included in the AI-generated list.

Finally, analysis stratified by AI performance showed contrasting effects. In the 17 cases where the correct diagnosis was included among the AI-generated differentials (AI-hit cases), median diagnostic accuracy increased from 0.647 without AI support to 0.824 with AI support ($p<0.0001$, **Fig. 3**A). In contrast, in the 13 cases where the correct diagnosis was not included (AI-miss cases), median diagnostic accuracy decreased from 0.154 to 0.077 with AI assistance ($p=0.018$, Fig. 3B). Detailed analyses of baseline accuracy, interindividual variability and outcome distributions are provided in Figs S1–S3.

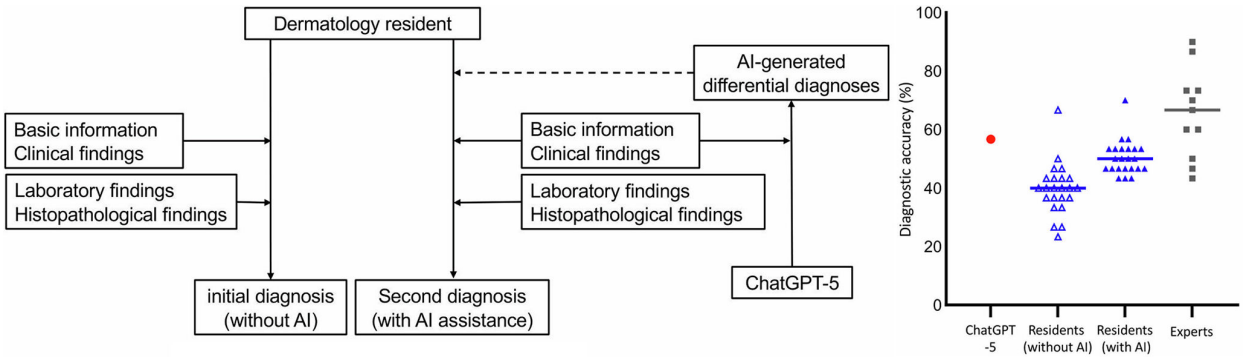

Fig. 1. Study design and overall diagnostic accuracy. A. study design of the two-phase diagnostic assessment. junior dermatology residents first evaluated each case independently without AI support and subsequently reassessed the same cases with reference to three differential diagnoses generated by ChatGPT-5. B. Overall diagnostic accuracy with and without AI assistance. blue triangles indicate diagnostic accuracy of junior dermatology residents without and with AI support. the red dot represents the proportion of cases in which ChatGPT-5 included the correct diagnosis among its top three differential suggestions. Grey squares indicate the diagnostic accuracy of board-certified dermatologists reported in a previous study (5).

DISCUSSION

In this study, we examined the utility and limitations of diagnostic support using a generative large language model (ChatGPT-5) among junior dermatology residents. Our findings demonstrated that the model was able to present plausible differential diagnoses based solely on initial clinical information, and that AI support was associated with an improvement in diagnostic accuracy. We also observed that participants tended to rely less on the AI output in cases where they felt confident in their diagnosis, whereas they were more likely to refer to the AI suggestions when encountering more challenging cases. These results are consistent with previous studies reporting dermatologist-level performance of image-based AI models in skin tumour diagnosis (1, 2) and improvements in diagnostic accuracy among less experienced clinicians when assisted by AI systems (3). Together, the findings suggest that language-based AI assistance may support clinical decision-making in dermatology.

However, the effect of AI support varied across individuals, possibly reflecting differences in diagnostic confidence and trust in AI output (6, 7). Furthermore, diagnostic accuracy improved when the AI suggestions included the correct diagnosis, whereas misleading AI outputs sometimes resulted in erroneous responses. This highlights the potential risk that the accuracy of AI output can directly influence clinical decisions. This study has several limitations. First, the AI was provided with structured clinical information and images, which

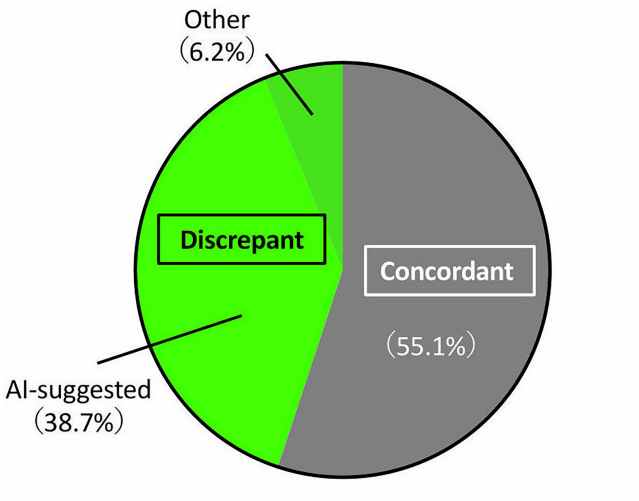

Fig. 2. Diagnostic changes with and without AI assistance. Proportion of concordant and discrepant diagnoses between AI-unassisted and AI-assisted conditions. Among 690 total responses, 380 (55.1%) diagnoses were concordant and 310 (44.9%) differed between conditions.

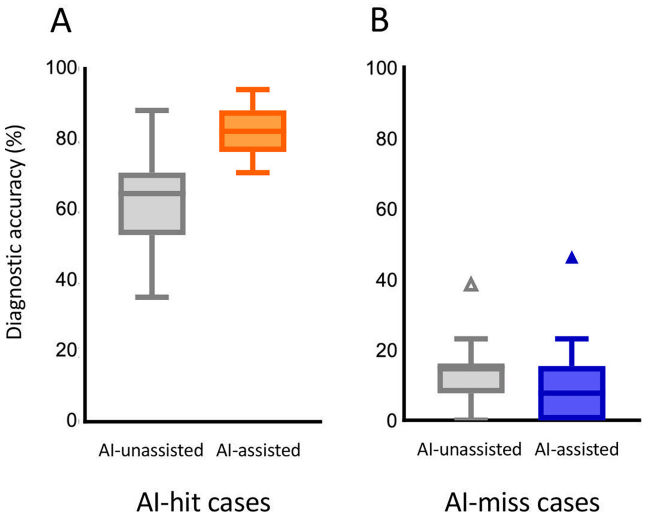

Fig. 3. Impact of AI performance on diagnostic accuracy. A. Diagnostic accuracy in AI-hit cases. In cases where ChatGPT-5 included the correct diagnosis among its top three differential suggestions (AI-hit cases, *n*=17), median diagnostic accuracy increased with AI assistance. B. Diagnostic accuracy in AI-miss cases. In cases where ChatGPT-5 did not include the correct diagnosis among its top three differential suggestions (AI-miss cases, n=13), median diagnostic accuracy decreased with AI assistance.

does not fully replicate real-world variability in clinical communication. The effectiveness of generative AI-based support may depend on the quality and clarity of input prompts. Second, while we quantitatively evaluated changes in diagnostic accuracy, cognitive factors such as diagnostic confidence or trust in AI were not assessed. Therefore, it remains unclear whether improvements reflected appropriate clinical reasoning or coincidental alignment with AI output. Third, the definition of "correctness" of AI differential suggestions was simplified, and the clinical appropriateness of each AI output was not qualitatively assessed in depth. Additionally, all participants reviewed the same cases twice in a fixed order (first without AI, then with AI). Therefore, learning or order effects cannot be excluded, and part of the observed improvement may reflect repeated exposure rather than the influence of AI assistance itself. Future studies should use counterbalanced or crossover designs with washout intervals to more rigorously isolate the causal effects of AI support. In addition, the number of AI-generated differential diagnoses was limited to 3, as presenting a single candidate may promote anchoring (9), whereas excessive options can increase cognitive burden (10, 11).

In conclusion, this study demonstrated that ChatGPT-5 can provide clinically relevant differential diagnoses based on initial clinical information, and that its suggestions can improve diagnostic accuracy among junior dermatology residents. Notably, in some cases residents selected diagnoses not included among the AI-generated differentials, suggesting that AI assistance may also stimulate diagnostic reasoning rather than merely constrain it (4). At the same time, inaccurate AI suggestions may lead to diagnostic errors, underscoring the need for critical appraisal when integrating AI support into clinical workflows. The findings suggest that generative AI may serve as a realistic support tool in early clinical encounters. Future work should explore conditions under which AI support is most beneficial and evaluate its educational applications in dynamic clinical settings. Rather than replacing clinical judgement, generative AI should be positioned as a cognitive partner that supports and enhances clinicians' diagnostic reasoning.

ACKNOWLEDGEMENTS

Data availability statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

The authors have no conflicts of interest to declare.

REFERENCES

1. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017; 542: 115–118. https://doi.org/10.1038/nature21056
2. Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. J Invest Dermatol 2018; 138: 1529–1538. https://doi.org/10.1016/j.jid.2018.01.028
3. Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human-computer collaboration for skin cancer recognition. Nat Med 2020; 26: 1229–1234. https://doi.org/10.1038/s41591-020-0942-0
4. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Discov 2019; 9: e1312. https://doi.org/10.1002/widm.1312
5. Yamamura Y, Fujii K, Nakashima C, Otsuka A. Evaluation of the accuracy of artificial intelligence (AI) models in dermatological diagnosis and comparison with dermatology specialists. Cureus 2025; 17: e77067. https://doi.org/10.7759/cureus.77067
6. Patil SV, Myers CG, Lu-Myers Y. Calibrating AI reliance – a physician's superhuman dilemma. JAMA Health Forum 2025; 6: e250106. https://doi.org/10.1001/jamahealthforum.2025.0106
7. Küper A, Lodde GC, Livingstone E, Schadendorf D, Krämer N. Psychological factors influencing appropriate reliance on AI-enabled clinical decision support systems: experimental web-based study among dermatologists. J Med Internet Res 2025; 27: e58660. https://doi.org/10.2196/58660
8. Esmaeilzadeh P. Ethical implications of using general-purpose LLMs in clinical settings: a comparative analysis of prompt engineering strategies and their impact on patient safety. BMC Med Inform Decis Mak 2025; 25: 342. https://doi.org/10.1186/s12911-025-03182-6
9. Hasanzadeh F, Josephson CB, Waters G, Adedinsewo D, Azizi Z, White JA. Bias recognition and mitigation strategies in artificial intelligence healthcare applications. NPJ Digit Med 2025; 8: 154. https://doi.org/10.1038/s41746-025-01503-7
10. Saposnik G, Redelmeier D, Ruff CC, Tobler PN. Cognitive biases associated with medical decisions: a systematic review. BMC Med Inform Decis Mak 2016; 16: 138. https://doi.org/10.1186/s12911-016-0377-1
11. Ferreira IBB, Menezes RC, Correia LCL, Andrade BB. Medicine beyond machines: viewpoint on the art of thinking in the age of AI. JMIR Form Res 2025; 9: e76669. https://doi.org/10.2196/76669