Physician–machine partnerships boost diagnostic accuracy, but bias persists

In a large-scale digital experiment on dermatology diagnosis, we found that specialists and generalists achieved diagnostic accuracy of 38% and 19%, respectively. With decision support from a fair deep learning system, the diagnostic accuracy of physicians improved by more than 33%, but the gap in accuracy of generalists widened across skin tones.

The question

Artificial intelligence (AI) has the potential to augment medical diagnosis based on patient images. However, physician–machine partnerships — the use of AI systems as decision support for physicians — are not guaranteed to be superior to either physicians or machines alone in diagnosing disease, as both are susceptible to systematic errors, especially for diagnosis of under-represented populations[1–3]. The effectiveness of physician–machine partnerships depends on the physicians' ability to correctly incorporate or ignore AI suggestions, which further depends on physician expertise, AI system performance, and physician understanding of when an AI system is prone to errant suggestions. In a store-and-forward teledermatology setting, we examined how diagnostic decisions made by physician–machine partnerships compared with those made by physicians alone.

The discovery

We recruited 1,118 physicians, including dermatologists, residents, primary care physicians and other physicians, and asked them to diagnose skin diseases on the basis of images displayed on an interactive website (Fig. 1a). In total, we collected 14,261 differential diagnoses on 364 clinical images of skin disease. The collection of many observations in a controlled setting gave us the statistical power to evaluate differences in diagnostic accuracy across several important dimensions: patient skin tone, physician expertise, skin disease type, AI assistance accuracy, and interface design. The AI assistance is based on the training of a deep neural network on clinical images of skin diseases, and the assistance interface asks physicians to either incorporate or ignore the suggestion offered by the AI system. This integrative experimental design is straightforward to replicate and promotes commensurability by revealing the diagnostic accuracy of physicians under a variety of conditions that can inform clinical practice[4].

We found that the diagnosis of inflammatory-appearing skin disease on the basis of a single image and up to three free-response answers (instead of multiple-choice answers) was challenging for specialists and generalists alike but was significantly improved with access to AI assistance. Leading diagnoses from specialists and generalists were correct in 27% and 13% of observations, respectively, and the accuracy of their top three differential diagnoses was 38% and 19%, respectively. With access to AI assistance and the opportunity for physicians to swap their leading diagnosis with the AI suggestion, physician performance increased significantly from 27% to 36% for specialists and from 13% to 22% for generalists. When the AI system made an incorrect suggestion, we found that physician accuracy decreased by 1.2 percentage points, which was not statistically significant. These results reveal that AI assistance has the potential to significantly increase the diagnostic accuracy of physicians with minimal misdiagnoses. However, we found that physicians were 4 percentage points less accurate on dark skin tones than on light skin tones, and AI assistance exacerbated the accuracy disparities by generalists by 5 percentage points, which is statistically significant. These results (Fig. 1b) illustrate that success in improving overall diagnostic accuracy does not necessarily address bias in accuracy across skin tones.

The implications

Our results reveal systematic disparities across skin tones in the diagnostic accuracy of skin diseases by physicians. Specialists and generalists alike should seek additional training in diagnosing skin disease in dark skin. Furthermore, these results raise several policy questions for the future of physician–machine partnerships, including how accuracy gains should be weighed against fairness concerns; how clinical practice should address the accuracy trade-offs between specialists and generalists with AI assistance; and how AI systems designed for clinical applications should be evaluated (such as on their own performance on a traditional holdout set, on how they change the performance of physicians on a test set, or on something else).

Although this digital experiment resembles a store-and-forward teledermatology setting, it does not fully match a clinical evaluation in which a physician would have access to additional information such as adjustments in light and angle of view, as well as patient symptoms, clinical history and behavioral information.

Future work should consider diagnostic accuracy in clinical settings, enable interactions that allow the physician to appropriately calibrate their confidence on particular diagnostic suggestions, and examine how decision support from AI assistance compares with collective human intelligence.

Matthew Groh
Kellogg School of Management, Northwestern University, Evanston, IL, USA.

This is a summary of:
Groh, M. et al. Deep learning-aided decision support for diagnosis of skin disease across skin tones. *Nat. Med*. https://doi.org/10.1038/s41591-023-02728-3 (2024).

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Published online: 05 February 2024

EXPERT OPINION

"This paper addresses a hotly debated issue in medical AI and offers an empirically grounded novel contribution to such debate. In particular, I praise the authors for adopting a practical approach to discussing AI bias, which could greatly illuminate not only the technical but also the ethical and policy discussion of the issue." **Alessandro Blasimme, ETH Zürich, Zürich, Switzerland.**

FIGURE

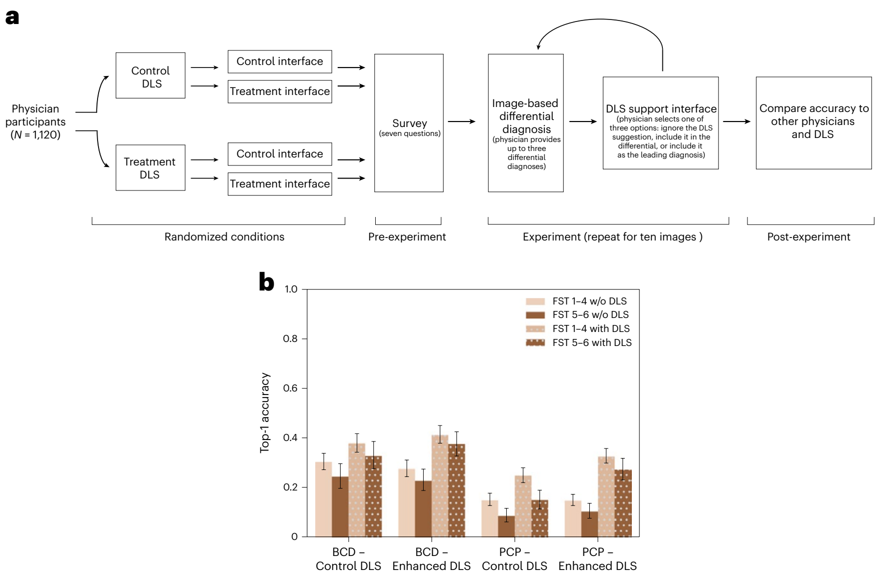

Fig. 1 | Experimental design and key results. a, A flowchart of the design underscoring our large-scale digital experiment on dermatology diagnosis, in which participants were asked to provide up to three differential diagnoses for a particular skin disease on the basis of a single image. **b**, Top-1 diagnostic accuracy of board-certified dermatologists (BCD) and primary care physicians (PCP) with or without (w/o) support from the deep learning systems (DLS) across Fitzpatrick skin types (FST). Control DLS and Treatment DLS refer to deep learning systems that are 47% and 84% accurate, respectively; the control DLS is trained on 31,219 images, and the treatment DLS is a 'Wizard of Oz' classifier that is included to represent a future, more accurate AI system. © 2024, Groh, M. et al., CCBY 4.0.

BEHIND THE PAPER

The initial formulation of this research question emerged in recognition of the promise and perils of AI in real-world applications. Given the propensity for systematic bias and error in humans and AI systems alike, we sought to design and conduct an experiment to reveal how human–AI collaboration for medical diagnosis unfolds in a clinical setting with diverse patients. Our research involved many moving parts, including building a multidisciplinary team with three board-certified dermatologists, annotating tens of thousands of images, taking a tangent to publish multiple papers along the way, curating appropriate images, developing a website for the digital experiment, and recruiting participants[5].

The most difficult aspect of the research proved to be participant recruitment; after years of preparation, we launched the digital experiment, but only a small number of physicians participated. Then we found Sermo — a social media network for physicians — which enabled us to recruit over 1,000 physician participants. **M.G.**

REFERENCES

1. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. AI in health and medicine. *Nat. Med.* **28**, 31–38 (2022).
   This review article covers advances in medical image analysis, problem formulation in human–AI collaboration, and common challenges, such as data scarcity and racial bias.
2. Liu, Y. et al. A deep learning system for differential diagnosis of skin disease. *Nat. Med.* **26**, 900–908 (2020).
   This paper demonstrates the potential of AI assistance in supporting general practitioners and nurse practitioners in diagnosing common skin diseases.
3. Daneshjou, R. et al. Disparities in dermatology AI performance on a diverse curated clinical image set. *Sci. Adv.* **8**, eabq6147 (2022).
   This paper reports that state-of-the-art dermatology AI models are less accurate on dark skin tones than on light skin tones.
4. Almaatouq, A. et al. Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral science. *Behav. Brain. Sci.* https://doi.org/10.1017/S0140525X22002874 (2022).
   This paper proposes an integrative experimental design whereby researchers map the design space of possible experiments and test these experiments together to promote commensurability in behavioral science.
5. Groh, M. et al. Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset. *Proc. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* 1820–1828 (2021).
   This paper presents a large dataset of clinical images annotated with the Fitzpatrick skin type scale and demonstrates that deep learning classifiers are most accurate on skin tones similar to those it was trained on.

FROM THE EDITOR

"As AI-enabled decision support becomes increasingly more capable, it will become crucial to understand the ways in which such decision support will be used by physicians and how this type of physician–machine partnership will affect diagnostic biases. This work presents a large-scale study of this topic in the context of skin disease diagnosis, involving 389 dermatologists and 459 primary care physicians from 39 countries." **Editorial Team, *Nature Medicine*.**