Race- and Ethnicity-Stratified Analysis of an Artificial Intelligence–Based Tool for Skin Condition Diagnosis by Primary Care Physicians and Nurse Practitioners

Ayush Jain1, MS; David Way1, ME; Vishakha Gupta1, MS; Yi Gao1, PhD; Guilherme de Oliveira Marinho1, BS; Jay Hartford1, MS; Rory Sayres1, PhD; Kimberly Kanada2, MD; Clara Eng1, PhD; Kunal Nagpal1, MS; Karen B DeSalvo1, MSc, MPH, MD; Jay Hartford1, MS; Rory Sayres1, PhD; Kimberly Kanada2, MD; Clara Eng1, PhD; Kunal Nagpal1, MS; Karen B DeSalvo1, MSc, MPH, MD; Greg S Corrado1, PhD; Lily Peng1, MD, PhD; Dale R Webster1, PhD; R Carter Dunn1, MS, MBA; David Coz1, MS; Susan J Huang2, MD; Yun Liu1, PhD; Peggy Bui1,3, MD, MBA; Yuan Liu1, PhD

1Google Health, Palo Alto, CA, United States
2Work done at Google Health via Advanced Clinical, Deerfield, IL, United States
3University of California, San Francisco, San Francisco, CA, United States

Corresponding Author:
Yun Liu, PhD
Google Health
ATTN liuyun
3400 Hillview Ave
Palo Alto, CA, 94304
United States
Phone: 1 415 736 0823
Email: liuyun@google.com

Abstract

Background: Many dermatologic cases are first evaluated by primary care physicians or nurse practitioners.

Objective: This study aimed to evaluate an artificial intelligence (AI)-based tool that assists with interpreting dermatologic conditions.

Methods: We developed an AI-based tool and conducted a randomized multi-reader, multi-case study (20 primary care physicians, 20 nurse practitioners, and 1047 retrospective teledermatology cases) to evaluate its utility. Cases were enriched and comprised 120 skin conditions. Readers were recruited to optimize for geographical diversity; the primary care physicians practiced across 12 states (2–32 years of experience, mean 11.3 years), and the nurse practitioners practiced across 9 states (2–34 years of experience, mean 13.1 years). To avoid memory effects from incomplete washout, each case was read once by each clinician either with or without AI assistance, with the assignment randomized. The primary analyses evaluated the top-1 agreement, defined as the agreement rate of the clinicians’ primary diagnosis with the reference diagnoses provided by a panel of dermatologists (per case: 3 dermatologists from a pool of 12, practicing across 8 states, with 5–13 years of experience, mean 7.2 years of experience). We additionally conducted subgroup analyses stratified by cases’ self-reported race and ethnicity and measured the performance spread: the maximum performance subtracted by the minimum across subgroups.

Results: The AI’s standalone top-1 agreement was 63%, and AI assistance was significantly associated with higher agreement with reference diagnoses. For primary care physicians, the increase in diagnostic agreement was 10% (P<.001), from 48% to 58%; for nurse practitioners, the increase was 12% (P<.001), from 46% to 58%. When stratified by cases’ self-reported race or ethnicity, the AI’s performance was 59%–62% for Asian, Native Hawaiian, Pacific Islander, other, and Hispanic or Latinx individuals and 67% for both Black or African American and White subgroups. For the clinicians, AI assistance–associated improvements across subgroups were in the range of 8%–12% for primary care physicians and 8%–15% for nurse practitioners. The performance spread across subgroups was 5.3% unassisted vs 6.6% assisted for primary care physicians and 5.2% unassisted vs 6.0% assisted for nurse practitioners. In both unassisted and AI-assisted modalities, and for both primary care physicians and nurse practitioners, the subgroup with the highest performance on average was Black or African American individuals, though the differences with other subgroups were small and had overlapping 95% CIs.

Conclusions: AI assistance was associated with significantly improved diagnostic agreement with dermatologists. Across race and ethnicity subgroups, for both primary care physicians and nurse practitioners, the effect of AI assistance remained high at 8%–15%, and the performance spread was similar at 5%–7%.

https://www.iproc.org/2022/1/e36885
Acknowledgments:  This work was funded by Google LLC.

Conflicts of Interest:  AJ, DW, VG, YG, GOM, JH, RS, CE, KN, KBD, GSC, LP, DRW, RCD, DC, Yun Liu, PB, and Yuan Liu are/were employees at Google and own Alphabet stocks.

(iproc 2022;8(1):e36885)  doi: 10.2196/36885

KEYWORDS
depth learning; computer-assisted diagnosis; dermatology; clinical images

Multimedia Appendix 1

Results of randomized reader study comparing clinicians assisted by artificial intelligence (AI, in orange) and those without assistance (“unassisted”, in blue). Performance was measured using the top-1 agreement metric, which indicates the rate at which the clinicians’ primary diagnosis matched that of a panel of dermatologists. The leftmost column summarizes the overall results for all readers and cases, whereas the other columns represent subgroups based on race/ethnicity. The results for primary care physicians (PCPs, top) and nurse practitioners (NPs, bottom) were similar.

[PNG File, 94 KB-Multimedia Appendix 1]