Development and Assessment of an Artificial Intelligence–Based Tool for Skin Condition Diagnosis by Primary Care Physicians and Nurse Practitioners in Teledermatology Practices

Ayush Jain, MS; David Way, ME; Vishakha Gupta, MS; Yi Gao, PhD; Guilherme de Oliveira Marinho, BS; Jay Hartford, MS; Rory Sayres, PhD; Kimberly Kanada, MD; Clara Eng, PhD; Kunal Nagpal, MD; Karen B. DeSalvo, MD, MPH, MSc; Greg S. Corrado, PhD; Lily Peng, MD, PhD; Dale R. Webster, PhD; R. Carter Dunn, MS, MBA; David Coz, MS; Susan J. Huang, MD; Yun Liu, PhD; Peggy Bui, MD, MBA; Yuan Liu, PhD

Abstract

IMPORTANCE Most dermatologic cases are initially evaluated by nondermatologists such as primary care physicians (PCPs) or nurse practitioners (NPs).

OBJECTIVE To evaluate an artificial intelligence (AI)-based tool that assists with diagnoses of dermatologic conditions.

DESIGN, SETTING, AND PARTICIPANTS This multiple-reader, multiple-case diagnostic study developed an AI-based tool and evaluated its utility. Primary care physicians and NPs retrospectively reviewed an enriched set of cases representing 120 different skin conditions. Randomization was used to ensure each clinician reviewed each case either with or without AI assistance; each clinician alternated between batches of 50 cases in each modality. The reviews occurred from February 21 to April 28, 2020. Data were analyzed from May 26, 2020, to January 27, 2021.

EXPOSURES An AI-based assistive tool for interpreting clinical images and associated medical history.

MAIN OUTCOMES AND MEASURES The primary analysis evaluated agreement with reference diagnoses provided by a panel of 3 dermatologists for PCPs and NPs. Secondary analyses included diagnostic accuracy for biopsy-confirmed cases, biopsy and referral rates, review time, and diagnostic confidence.

RESULTS Forty board-certified clinicians, including 20 PCPs (14 women [70.0%]; mean experience, 11.3 [range, 2-32] years) and 20 NPs (18 women [90.0%]; mean experience, 13.1 [range, 2-34] years) reviewed 1048 retrospective cases (672 female [64.2%]; median age, 43 [interquartile range, 30-56] years; 41920 total reviews) from a teledermatology practice serving 11 sites and provided 0 to 5 differential diagnoses per case (mean [SD], 1.6 [0.7]). The PCPs were located across 12 states, and the NPs practiced in primary care without physician supervision across 9 states. Artificial intelligence assistance was significantly associated with higher agreement with reference diagnoses. For PCPs, the increase in diagnostic agreement was 10% (95% CI, 8%-11%; \( P < .001 \)), from 48% to 58%; for NPs, the increase was 12% (95% CI, 10%-14%; \( P < .001 \)), from 46% to 58%. In secondary analyses, agreement with biopsy-obtained diagnosis categories of malignant, precancerous, or benign increased by 3% (95% CI, 1%-5%) for PCPs and by 8% (95% CI, 3%-13%) for NPs. Rates of desire for biopsies decreased by 1% (95% CI, 0-3%) for PCPs and 2% (95% CI, 1%-3%) for NPs; the rate of desire for referrals decreased by 3% (95% CI, 1%-4%) for PCPs and NPs. Diagnostic agreement on cases not indicated for a dermatologist referral increased by 10% (95% CI, 8%-12%)

Key Points

Question Can artificial intelligence help primary care physicians and nurse practitioners diagnose skin conditions more accurately?

Findings In this diagnostic study of 20 primary care physicians and 20 nurse practitioners reviewing 1048 retrospective cases, artificial intelligence assistance was significantly associated with higher agreement with diagnoses made by a dermatologist panel, with an increase from 48% to 58% for primary care physicians and an increase from 46% to 58% for nurse practitioners. These outcomes correspond to a benefit for 1 in every 8 to 10 cases.

Meaning Artificial intelligence may help clinicians diagnose skin conditions more accurately in primary care practices, where most skin diseases are initially evaluated.

Open Access. This is an open access article distributed under the terms of the CC-BY-NC-ND License.


April 28, 2021 1/14

Downloaded From: https://jamanetwork.com/ by a Non-Human Traffic (NHT) User on 09/20/2021
Abstract (continued)

8%-12%) for PCPs and 12% (95% CI, 10%-14%) for NPs, and median review time increased slightly by
5 (95% CI, 0-8) seconds for PCPs and 7 (95% CI, 5-10) seconds for NPs per case.

CONCLUSIONS AND RELEVANCE Artificial intelligence assistance was associated with improved
diagnoses by PCPs and NPs for 1 in every 8 to 10 cases, indicating potential for improving the quality
dermatologic care.


Introduction

With 2 billion people affected globally, skin conditions are a leading cause of morbidity. The
examination of some skin conditions by dermatologists results in significantly higher diagnostic
accuracy and is associated with better clinical outcomes than non-dermatologist examination.
However, owing to lack of access to dermatologists, only 28% of skin cases are seen by a specialist;
therefore, nonspecialists play a pivotal role in the assessment of skin lesions and initiation of clinical
management and referrals. The diagnostic accuracy of nonspecialists is reportedly only 24% to
70%, suggesting that currently available resources, such as dermatology textbooks, medical
information portals, and online image search engines, remain insufficient to guide nonspecialists.

Several algorithms incorporating artificial intelligence (AI) have been developed to help
interpret both clinical and dermoscopic images for a variety of skin conditions, and the effect
of AI-based support on dermoscopic images has been studied. However, an open question
remains as to whether AI assistance can help primary care physicians (PCPs) and nurse practitioners
(NPs) diagnose skin conditions from clinical images (ie, taken without specialized equipment).

We developed an AI-based tool and conducted a multiple-reader, multiple-case diagnostic
study in which PCPs and independently practicing NPs retrospectively reviewed skin cases from a
teledermatology service, representing 120 different skin conditions. We used randomization to
ensure readers reviewed each case only once, either with or without AI assistance. Our primary
objective was to measure the AI assistance-associated changes in diagnostic accuracy of PCPs and
NPs without specialist training in dermatology.

Methods

This study was approved by the Quorum Institutional Review Board, Seattle, Washington, and
deemed exempt from informed consent because all data and images were de-identified. The
Standards for Reporting of Diagnostic Accuracy (STARD) reporting guideline was followed for
this study.

The AI Tool

Liu et al previously described an AI algorithm that provides a differential diagnosis given clinical
photographs of skin conditions and the medical history (Table 1 in the Supplement). Their AI model
was developed using 16114 cases and used a convolutional neural network to output prediction
scores across 419 skin conditions. In the present study, we created a web-based tool using the AI
model described by Liu et al by incorporating user experience insights (Figure 1).

The tool provides information about the case, including demographic information, history of
present illness, and other elements of the patient’s medical history. For each case, 1 to 6 images were
available for review (median, 4), and readers could toggle between or zoom in on images. Primary
care physicians and NPs reviewed these cases using a laptop and could consult additional resources
as they would in clinical practice.
The AI assistance component of the web-based tool was only available during the assisted mode of the study (described below). At the top of the panel, the interface displayed the skin conditions that were output by the AI, sorted in order of the AI’s predicted likelihood scores. Artificial intelligence predictions with low scores (<0.05) were removed, and the list was limited to 5 skin conditions to avoid presenting extraneous information. Each condition could be clicked on to display additional information (Figure 1 and eFigure 1 and the AI Tool Interface section in the eMethods in the Supplement).

Study Design
To evaluate whether this tool could assist primary care clinicians in diagnosing skin conditions, we conducted a multiple-reader, multiple-case diagnostic study with 20 PCPs and 20 NPs (Figure 1). The characteristics of the clinicians are described in the Reader Characteristics section of the eMethods and eFigures 2 and 3 in the Supplement. Before reviewing the study cases, each reader was presented with materials describing how to use the AI assistant and given the opportunity to practice using the AI assistant with 2 sample cases (independent of the study cases). Additional details of this training can be found in the Onboarding Process section in the eMethods in the Supplement.

The study used cases from 2 retrospective data sets from California and Hawaii previously used to validate the AI algorithm. Specifically, the prior study used a validation set A and a subset (validation set B) enriched for rarer conditions via random sampling stratified by condition. Validation set B (963 cases) was included in its entirety. From validation set A, all 85 cases for which biopsy results were available were also included to yield a total of 1048 cases (Table). None of the PCPs or NPs in this study previously reviewed these cases, and the AI algorithm used was identical to the one used in the previous study.

Each reader was randomly assigned to 1 of 2 reader cohorts. The 2 reader cohorts read the same cases but with the opposite assistance modalities (ie, unassisted vs AI assisted) for each case. To reduce effects associated with switching modalities, the 1048 cases were divided into batches of 50 cases (except the last 48 cases, which were divided into 2 batches of 24 cases), and the assistance modality switched after each batch of cases. For the first batch of 50 cases, reader cohort 1 reviewed 50 cases, and the assistance modality alternated every 50 cases. For every case, each clinician was instructed to rank as many as 3 differential diagnoses using a search-as-you-type interface and selecting matching skin conditions from a list of 3961 conditions. If their desired skin condition was not present, clinicians could provide free-text entries. All skin conditions were mapped to a list of 419 conditions. SCC indicates squamous cell carcinoma; SCCis, SCC in situ.

<table>
<thead>
<tr>
<th>Number of Cases</th>
<th>AI-assisted cases</th>
<th>Unassisted cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>1048 cases</td>
<td>50</td>
<td>24</td>
</tr>
<tr>
<td>1048 cases</td>
<td>50</td>
<td>24</td>
</tr>
<tr>
<td>1048 cases</td>
<td>50</td>
<td>24</td>
</tr>
</tbody>
</table>

The AI assistant shows as many as 5 top predictions of skin conditions, with the confidence in each prediction shown as colored dots and additional information (eg, sample images from an atlas) available with a click. More details are available in eFigure 1 in the Supplement. The study was designed as a multiple-reader, multiple-case (MRMC) study comprising 1048 cases. Two groups of clinicians (primary care physicians [PCPs] and nurse practitioners [NPs]) reviewed each case with or without AI assistance. The modality alternated every 50 cases. For every case, each clinician was instructed to rank as many as 3 differential diagnoses using a search-as-you-type interface and selecting matching skin conditions from a list of 3961 conditions. If their desired skin condition was not present, clinicians could provide free-text entries. All skin conditions were mapped to a list of 419 conditions. SCC indicates squamous cell carcinoma; SCCis, SCC in situ.
<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Data set</th>
<th>Cases with diagnoses from histologic findings (n = 152)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Full study (n = 1048)a</td>
<td>2017-2018</td>
</tr>
<tr>
<td></td>
<td>No. of years</td>
<td>2017-2018</td>
</tr>
<tr>
<td>No. of sites</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>No. of images included in study</td>
<td>3935</td>
<td>413</td>
</tr>
<tr>
<td>No. of patients included in study</td>
<td>1016</td>
<td>152</td>
</tr>
<tr>
<td>Age, median (IQR), ya</td>
<td>43 (30-56)</td>
<td>49 (35-59)</td>
</tr>
<tr>
<td>Sex, No. (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>672 (64.2)</td>
<td>99 (65.1)</td>
</tr>
<tr>
<td>Male</td>
<td>375 (35.8)</td>
<td>53 (34.9)</td>
</tr>
<tr>
<td>Race and ethnicity, No. (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>American Indian or Alaska Native</td>
<td>9 (0.9)</td>
<td>0</td>
</tr>
<tr>
<td>Asian</td>
<td>102 (9.7)</td>
<td>5 (3.3)</td>
</tr>
<tr>
<td>Black or African American</td>
<td>66 (6.3)</td>
<td>5 (3.3)</td>
</tr>
<tr>
<td>Hispanic or Latino</td>
<td>447 (42.7)</td>
<td>59 (38.8)</td>
</tr>
<tr>
<td>Native Hawaiian or Pacific Islander</td>
<td>20 (1.9)</td>
<td>2 (1.3)</td>
</tr>
<tr>
<td>White</td>
<td>365 (34.9)</td>
<td>80 (52.6)</td>
</tr>
<tr>
<td>Not specified</td>
<td>38 (3.6)</td>
<td>1 (0.7)</td>
</tr>
<tr>
<td>Fitzpatrick skin type (6 types), No. (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>2 (0.2)</td>
<td>2 (1.3)</td>
</tr>
<tr>
<td>II</td>
<td>109 (10.4)</td>
<td>17 (11.2)</td>
</tr>
<tr>
<td>III</td>
<td>668 (63.8)</td>
<td>111 (73.0)</td>
</tr>
<tr>
<td>IV</td>
<td>205 (19.6)</td>
<td>14 (9.2)</td>
</tr>
<tr>
<td>V</td>
<td>25 (2.4)</td>
<td>1 (0.7)</td>
</tr>
<tr>
<td>VI</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Unknown</td>
<td>38 (3.6)</td>
<td>7 (4.6)</td>
</tr>
<tr>
<td>Skin conditions based on primary diagnosis, No. (%)c</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acne</td>
<td>40 (3.8)</td>
<td>NA</td>
</tr>
<tr>
<td>Actinic keratosis</td>
<td>39 (3.7)</td>
<td>1 (0.7)</td>
</tr>
<tr>
<td>Allergic contact dermatitis</td>
<td>25 (2.4)</td>
<td>NA</td>
</tr>
<tr>
<td>Alopecia areata</td>
<td>37 (3.5)</td>
<td>NA</td>
</tr>
<tr>
<td>Androgenetic alopecia</td>
<td>32 (3.1)</td>
<td>NA</td>
</tr>
<tr>
<td>Basal cell carcinoma</td>
<td>36 (3.4)</td>
<td>32 (21.1)</td>
</tr>
<tr>
<td>Cyst</td>
<td>32 (3.1)</td>
<td>1 (0.7)</td>
</tr>
<tr>
<td>Eczema</td>
<td>53 (5.1)</td>
<td>NA</td>
</tr>
<tr>
<td>Folliculitis</td>
<td>32 (3.1)</td>
<td>3 (2.0)</td>
</tr>
<tr>
<td>Hidradenitis</td>
<td>34 (3.2)</td>
<td>NA</td>
</tr>
<tr>
<td>Lentigo</td>
<td>32 (3.1)</td>
<td>3 (2.0)</td>
</tr>
<tr>
<td>Melanocytic nevus</td>
<td>61 (5.8)</td>
<td>28 (18.4)</td>
</tr>
<tr>
<td>Melanoma</td>
<td>20 (1.9)</td>
<td>6 (3.9)</td>
</tr>
<tr>
<td>Postinflammatory hyperpigmentation</td>
<td>28 (2.7)</td>
<td>NA</td>
</tr>
<tr>
<td>Psoriasis</td>
<td>40 (3.8)</td>
<td>NA</td>
</tr>
<tr>
<td>SCC/SCCIS</td>
<td>34 (3.2)</td>
<td>14 (0.2)</td>
</tr>
<tr>
<td>SK/ISK</td>
<td>52 (5.0)</td>
<td>13 (8.6)</td>
</tr>
<tr>
<td>Scar condition</td>
<td>34 (3.2)</td>
<td>2 (1.3)</td>
</tr>
<tr>
<td>Seborrheic dermatitis</td>
<td>37 (3.5)</td>
<td>NA</td>
</tr>
<tr>
<td>Stasis dermatitis</td>
<td>36 (3.4)</td>
<td>3 (2.0)</td>
</tr>
<tr>
<td>Tinea</td>
<td>25 (2.4)</td>
<td>NA</td>
</tr>
<tr>
<td>Tinea versicolor</td>
<td>31 (3.0)</td>
<td>1 (0.7)</td>
</tr>
<tr>
<td>Urticaria</td>
<td>34 (3.2)</td>
<td>NA</td>
</tr>
<tr>
<td>Verruca vulgaris</td>
<td>37 (3.5)</td>
<td>8 (5.3)</td>
</tr>
<tr>
<td>Vitiligo</td>
<td>36 (3.4)</td>
<td>NA</td>
</tr>
<tr>
<td>Otherd</td>
<td>116 (11.1)</td>
<td>65 (42.8)</td>
</tr>
</tbody>
</table>

Abbreviations: IQR, interquartile range; NA, not applicable; SCC/SCCIS, squamous cell carcinoma/squamous cell carcinoma in situ; SK/ISK, seborrheic keratosis/irritated seborrheic keratosis.

a One case was removed from the study for logistical reasons.
b Of 165 cases, 13 had equivocal biopsy results and were excluded from the biopsy analysis. A total of 141 cases had growths and 53 were malignant.
c Enrichment was performed to avoid skew toward common conditions (eg, acne and eczema) as described previously and additionally to include all available cases with biopsy confirmation.
d Conditions with fewer than 10 cases each.
these cases with AI assistance, whereas reader cohort 2 reviewed the same cases unassisted. The next batch of cases were reviewed in the opposite modality (Figure 1). By ensuring each reader reviewed each case only once in either the assisted or unassisted modality, this design eliminated any memory effect associated with a crossover study (where memorable cases may inflate the diagnostic performance when reviewed a second time by the same readers).28,29

During the case reviews, the readers either provided their top differential diagnoses or indicated that they were unable to diagnose a case. They also answered a few questions on their intended clinical next steps for each case (see the Study End Points section below). Reviews were performed without time constraint. These reviews occurred from February 21 to April 28, 2020.

Reference Diagnoses
Reference diagnoses were provided by a panel of dermatologists.26 Briefly, 3 US board-certified dermatologists (from a pool of 12) independently reviewed each case. The dermatologists participated in the study via Advanced Clinical, Deerfield, Illinois; had 5 to 13 years of experience (mean [SD], 7.2 [2.7] years); and practiced in multiple states, including Colorado, Hawaii, Iowa, Maryland, New York, South Carolina, Tennessee, and Texas. Reference diagnoses were obtained using a previously described collective intelligence approach, which results in more reproducible diagnoses than diagnoses obtained by individual dermatologist review (eTable 2 in the Supplement).26,30 This approach assigns a vote to each diagnosis based on its ranking: the first diagnosis in a dermatologist's differential was given a weight of 1/1 = 1; the secondary diagnosis was given a weight of 1/2 = 0.5. The votes for each diagnosis were summed across the 3 dermatologists, and the top-voted diagnosis was considered the primary diagnosis of the panel.

Agreement was also assessed against biopsy-confirmed diagnoses when available. Diagnoses were extracted from pathology reports by the teledermatology service before transfer to study investigators. These diagnoses were then mapped to skin conditions by US board-certified dermatologists (including K.K. and S.J.H.). The case distribution across these diagnoses (both clinical and histologic) are presented in the Table; of 152 cases with available biopsy results, the diagnosis of 141 cases was growths.

Study End Points
Our study was designed to evaluate 2 prespecified primary end points: (1) the agreement rate of the primary differential diagnosis of the PCPs with the reference diagnosis and (2) the agreement rate of the primary differential diagnosis of the NPs with the reference diagnosis. Based on the relative frequencies of conditions in this data set, the chance agreement is 3.77%.

Several secondary analyses were planned. First, for cases with biopsy results, diagnoses were classified as malignant, precancerous, or benign and were evaluated against biopsy-determined diagnoses. Clinicians were also asked to report whether they would have recommended a biopsy or referred the case to a dermatologist. For the subset of reads in which clinicians reported they would not opt for a referral, we assessed the diagnostic agreement rate. We also analyzed the time taken to review cases and self-reported diagnostic confidence.

Finally, 2 additional metrics (top-3 agreement and average overlap)31 were used for more comprehensive evaluation of cases in which additional follow-up may be needed to arrive at a definitive diagnosis (Additional Evaluation Metrics section in the eMethods in the Supplement). An exploratory analysis also measured the effect of AI assistance on dermatologist agreement with reference diagnoses.

Statistical Analysis
Data were analyzed from May 26, 2020, to January 27, 2021. To compare clinicians reviewing cases with AI assistance and reviewing cases without, we used a permutation test32 with 1000 iterations. In each iteration, we permuted the assignment of whether reads were assisted or unassisted (ie, one-half of the full set of assisted and unassisted reads per case were selected to be assisted and the
other half unassisted). Sensitivity analysis using a permutation test that preserved the reader cohorts’ structure and another statistical analysis via a generalized linear mixed model produced similar results (see the Alternative Statistical Analyses section in the eMethods in the Supplement). Because this study had 2 prespecified primary end points (both 1-tailed superiority tests), we applied the Bonferroni correction, and \( P < .0125 \) was considered statistically significant (halved from \( \alpha = 0.05 \) owing to 1-tailed tests and halved again owing to having 2 primary end points). Confidence intervals were computed by bootstrapping across both cases and readers for each sampled case (1000 iterations; sampling both cases and reader with replacement in each iteration). Hypothesis tests were conducted in Python, version 3.6.7 (Python Software Foundation).

**Results**

This study involved the participation of 40 board-certified clinicians, including 20 PCPs (14 women [70.0%] and 6 men [30.0%]; mean experience, 11.3 [range, 2-32] years) who were located across 12 states and 20 NPs (18 women [90.0%] and 2 men [10.0%]; mean experience, 13.1 [range, 2-34] years) who practiced in primary care without physician supervision across 9 states. These clinicians reviewed 1048 teledermatology cases (672 women [64.2%] and 375 men [35.8%], with 1 missing; median age, 43 [interquartile range, 30-56] years) from 11 sites (Table) and provided 0 to 5 differential diagnoses per case (mean [SD], 1.6 [0.7]), for a total of 41 920 case reviews. Every PCP and NP reviewed each case only once, either with or without AI assistance (Figure 1).

Artificial intelligence assistance was associated with significantly higher top-1 agreement with the reference diagnosis (Figure 2A and eTable 3 in the Supplement). For PCPs, the increase in diagnostic agreement was 10% (95% CI, 8%-11%; \( P < .001 \)), from 48% to 58%; for NPs, the improvement was 12% (95% CI, 10%-14%; \( P < .001 \)), from 46% to 58%. Assistance was associated with improvements for all 40 readers, although the magnitude varied by reader (range, 2%-22%; median, 10%) (Figure 2B). Similar improvements were observed beyond the primary diagnosis based on the top-3 agreement, average overlap, per-condition sensitivity, and \( \kappa \) value (eFigures 4 and 5 and eTable 3 in the Supplement). In an exploratory analysis, 2 dermatologists’ agreement with the reference diagnosis remained largely unchanged with AI assistance, increasing by 2% (95% CI, −1% to 5%), from 63% to 66% (eFigures 4 and 5 in the Supplement).²⁶

For cases with available biopsy diagnoses (\( n = 141 \)), the readers’ accuracy at classifying lesions as malignant, precancerous, or benign trended upward by 3% for PCPs (95% CI, −1% to 7%) from 64% to 67% and by 8% for NPs (95% CI, 3%-13%) from 60% to 68% (Figure 2C-D). Subgroup analysis further found that sensitivity for malignant lesions, precancerous lesions, infectious skin diseases, and categories of hair loss trended upward or remained similar with assistance for both NPs and PCPs, with improvements ranging from −1% to 36% (eTable 4 in the Supplement).

On the subset of cases in which the top prediction of AI was accurate (63% of cases), the use of assistance was associated with an increased top-1 agreement with reference diagnosis of 18% (95% CI, 16%-20%) for PCPs and 21% (95% CI, 19%-23%) for NPs. On the contrary, when none of the AI tool’s predictions was correct (13% of cases), the agreement was 8% lower (95% CI, 5%-12%) for PCPs and 9% lower (95% CI, 6%-12%) for NPs. The effects were intermediate when the correct diagnosis was in the second or third position instead of the first (see the Impact of AI Accuracy on Assistance section of eMethods and eFigures 6 and 7 in the Supplement). An exploratory analysis also suggested that assistance was particularly beneficial for less ambiguous cases. For example, in the subset of cases in which the dermatologist panel had unanimous agreement, the use of AI assistance was associated with a top-1 agreement increase of 13% (95% CI, 10%-15%) for PCPs and of 16% (95% CI, 14%-19%) for NPs (eFigure 8 in the Supplement). Subanalyses also indicated that assistance-associated benefits were consistent during the study and across several skin types (eFigures 9 and 10 in the Supplement).

Artificial intelligence assistance was also associated with changes in several simulated clinical decisions (Figure 3A-B). The rates of indicating a need for biopsy were 1% lower (95% CI, 0%-3%) for...
Every clinician (primary care physicians [PCPs] or nurse practitioners [NPs]) provided their differential diagnosis (several rank-ordered conditions), which were then mapped to 419 skin conditions. Only agreement in the top differential diagnosis (how often the clinicians’ primary diagnosis agreed with the top diagnosis of a panel of dermatologists [top-1 agreement]) is considered, with additional details in eFigures 4 and 5 in the Supplement. Panels A and B cover all 1048 cases, whereas panels C and D cover 141 cases with growths and biopsy confirmation. A, Top-1 agreement increased with AI assistance ($P < .001$ for both PCPs and NPs). B, For top-1 agreement for unassisted vs assisted modalities for each individual clinician, a value above the diagonal indicates that the clinician had a higher agreement with dermatologists when assisted by AI. C and D, A similar analysis evaluated diagnostic accuracy for growths with biopsy confirmation on the 3-way classification of malignant, precancerous, and benign. Error bars represent 95% CIs. Additional analysis of assistance stratified by AI agreement with the reference diagnoses is presented in eFigures 6 and 7 in the Supplement.
Figure 3. Comparing Simulated Clinical Decisions by Clinicians When Assisted by Artificial Intelligence vs Unassisted

A) Biopsy rate

B) Referral rate

C) Agreement with reference diagnoses among nonreferred cases

D) Agreement with reference diagnoses among referred cases

A, Rate of biopsy for all cases. B, Rate of referrals for all cases. C, Diagnostic accuracy among nonreferred cases. D, Diagnostic accuracy among referred cases. Top-3 agreement rates for cases for whom the primary care physicians (PCPs) and nurse practitioners (NPs) did and did not indicate a referral are presented in eFigure 11 in the Supplement. Error bars represent 95% CIs.

Figure 4. Comparing Clinicians’ Confidence and Case Review Time When Assisted by Artificial Intelligence vs Unassisted

A) Distribution of confidence scores per read

B) Distribution of review times per case

A, Confidence of the primary care physicians (PCPs) and nurse practitioners (NPs) as a stacked bar plot. NA indicates cases for which the clinician could not provide a diagnosis. B, Comparison of the differences in case review time for the full set of 1048 cases as a box plot. The box edges represent quartiles, whereas the whiskers extend to the last observed points that fall within 1.5 times the interquartile range from the quartiles. Outliers beyond the whiskers are indicated with dots; a total of 182 (0.4% of the reads) outliers beyond 900 seconds are excluded from the 4 box plots for ease of visualization. The median time for diagnosis increased from 89 to 94 seconds for PCPs and from 77 to 84 seconds for NPs.
**Discussion**

In this study, 40 clinicians each reviewed 1048 teledermatology cases, with AI assistance for a random half of the cases and without AI assistance for the remaining half. Artificial intelligence assistance was associated with a higher agreement rate with dermatologists' reference diagnoses for both PCPs and NPs. The absolute effect size of 10% and 12% corresponds to an improved diagnosis for 1 in every 8 to 10 cases.

For both PCPs and NPs, AI assistance was also associated with lower rates of recommending a biopsy or specialist referral, marked increase in self-reported diagnostic confidence, and higher diagnostic agreement rates (with dermatologists) in nonreferred cases. These observations suggest that AI assistance improved skin condition diagnosis and diagnostic confidence of nonspecialists without incurring a reflexive increased use of referrals or biopsies. These improvements came at a modest cost of only a median of 5 to 7 additional seconds per case.

Our observations suggest that AI has the potential to augment the ability of PCPs and NPs independently practicing primary care to diagnose and triage skin conditions more effectively. Cutaneous disease is the chief complaint in 12% to 21% of primary care visits, and access to dermatologists is limited. Nonspecialists have suboptimal diagnostic accuracy and have been shown to perform more biopsies while diagnosing fewer malignant neoplasms than dermatologists. Therefore, improving the diagnostic accuracy of nonreferred cases while reducing unnecessary referrals and biopsies could have enormous implications for health care systems.

According to the American Academy of Dermatology, the estimated direct health care cost of skin disease in the US is $75 billion, including $46 billion in medical costs (office visits, procedures, and tests), with an additional $11 billion of indirect opportunity costs from missed work or decreased productivity for patients and their caregivers. Appropriate diagnosis of dermatologic conditions at the point of care in primary care settings could translate to fewer delays in diagnosis and management and increased capacity for dermatology offices. Artificial intelligence also has the potential to enhance triage by improving the quality of information in referrals and enable dermatology offices to better prioritize the urgency of referrals. The clinical impact of this tool would need to be determined in prospective studies.

This AI tool uses as input images of the skin condition as well as a structured medical history. These images were taken using consumer-grade point-and-shoot cameras and mobile devices without specialized hardware. The interface used in this study was designed for store-and-forward teledermatology; however, extension to live, interactive teledermatology is in principle straightforward. In either case, the telemedicine format could be particularly useful in the COVID-19 era for populations at high risk of complications in the event of infection due to in-person care. The AI tool could also be used in an in-person clinic setting because AI interpretation of images is feasible within seconds on modern smartphones. Such use could enable physicians to conduct follow-up tests (eg, potassium hydroxide test to confirm fungal infection), ask clarifying questions about the medical history, or conduct a closer physical examination to realize greater improvements in diagnostic ability.

More generally, and consistent with the consensus statements from both the American Medical Association and the American Academy of Dermatology, this tool was specifically designed to augment clinicians' diagnostic ability. To improve trust and empower readers to evaluate suggestion reliability, the tool provides a measure of its confidence and canonical examples of each suggested diagnosis. For skin conditions from which the AI algorithm had limited data to learn, suggestions are accompanied by a limited data warning. These features were designed to enable nonspecialists to diagnose cases more accurately and with greater confidence.

Other studies have explored the potential of AI-based dermatology tools. Han et al found a 7% increase in diagnostic accuracy when 2 dermatologists and 2 residents reviewed 2201 cases a second time with AI assistance. Assistance-associated improvements were also seen for 21 dermatologists and 26 residents on 240 images for detection of malignant neoplasms. Tschandl et al highlighted...
the importance of effective human/computer interaction for AI tools for interpreting dermoscopic images, with improvements in showing multiclass prediction probabilities by skin condition but not for binary predictions of malignant neoplasms or AI-based retrieval of similar images. Our study complements these prior works. First, we evaluated images from nonspecialized, widely available devices. Second, we specifically examined the effect of AI assistance on PCPs and NPs, who perform most skin condition assessments. In addition, we assessed 2 pivotal clinical decisions: biopsy and referral. Finally, our randomized study design avoids any potential memory effects of reviewing the same case more than once.

Limitations
This study has some limitations. First, these were teledermatology cases that were a mix of cases that were referred from primary care and other cases that were submitted at the patient’s request. The potentially increased case difficulty and case enrichment may have affected clinician diagnostic performance. Second, in terms of Fitzpatrick skin types (which categorize skin tone and propensity to tan), types I and V are underrepresented, and type VI is absent in this data set. Because disease can present differently across skin types, the further study of additional skin types is warranted. Third, AI-associated improvements for malignant neoplasms were lower than those across all cases, and future work is needed to further improve the AI tool for malignant neoplasms. Our randomized study design of 1 modality per case/reader pair precludes inferences about any specific case and reader. Alternative study designs such as sequential reading (unassisted followed by assisted) or fully crossed setups could be explored, although biases from anticipation of AI assistance or incomplete washout will need to be averted. Finally, the “store-and-forward” nature of these cases restricted the ability of the clinicians to ask follow-up questions and perform tests. As such, the insights here are more directly relevant to a store-and-forward setting than in-person clinics or live interactive telemedicine visits.

Conclusions
Our AI tool was significantly associated with improved PCP and NP diagnostic agreement with dermatologists on skin condition cases from a teledermatology service. Prospective studies are warranted to study the impact of its use in both telemedicine settings and in-person primary care visits.

ARTICLE INFORMATION
Accepted for Publication: February 1, 2021.
Open Access: This is an open access article distributed under the terms of the CC-BY-NC-ND License. © 2021 Jain A et al. JAMA Network Open.
Corresponding Author: Yun Liu, PhD, Google Health, 3400 Hillview Ave, Palo Alto, CA 94304 (liuyun@google.com).
Author Affiliations: Google Health, Palo Alto, California (Jain, Way, Gupta, Gao, de Oliveira Marinho, Hartford, Sayres, Eng, Napgal, DeSalvo, Corrado, Peng, Webster, Dunn, Coz, Yun Liu, Bui, Yuan Liu); Google Health via Advanced Clinical, Deerfield, Illinois (Kanada, Huang); Division of Hospital Medicine, University of California, San Francisco (Bui).
Author Contributions: Drs Bui and Yuan Liu contributed equally to the study. Mr. Jain and Dr Yun Liu had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: Jain, Way, de Oliveira Marinho, Sayres, Eng, Corrado, Peng, Webster, Dunn, Coz, Huang, Yun Liu, Bui, Yuan Liu.
Acquisition, analysis, or interpretation of data: Jain, Way, Gupta, Gao, Hartford, Sayres, Kanada, Nagpal, DeSalvo, Dunn, Coz, Huang, Yun Liu, Bui, Yuan Liu.

Drafting of the manuscript: Jain, Gupta, de Oliveira Marinho, Hartford, Sayres, Kang, Nagpal, Yun Liu, Bui, Yuan Liu.

Critical revision of the manuscript for important intellectual content: Jain, Way, Gao, Sayres, Kanada, Eng, Nagpal, DeSalvo, Corrado, Peng, Webster, Dunn, Coz, Huang, Yun Liu, Bui, Yuan Liu.

Statistical analysis: Jain, Way, Gupta, Gao, Yun Liu, Yuan Liu.

Obtained funding: Corrado, Peng, Webster, Dunn, Bui, Yuan Liu.

Administrative, technical, or material support: Way, Gupta, de Oliveira Marinho, Hartford, Kanada, Eng, DeSalvo, Peng, Dunn, Coz, Huang, Yun Liu, Bui, Yuan Liu.

Supervision: Corrado, Peng, Webster, Coz, Huang, Yun Liu, Bui, Yuan Liu.

Conflict of Interest Disclosures: Mr Jain reported a patent pending and ownership of Alphabet stock. Mr Way reported a patent pending and ownership of Alphabet stock. Mrs Gupta reported a patent pending and ownership of Alphabet stock. Dr Gao reported ownership of Alphabet stock. Mr de Oliveira Marinho reported ownership of Alphabet stock. Mr Hartford reported ownership of Alphabet stock. Dr Sayres reported a patent pending and ownership of Alphabet stock. Dr Kanada reported paid consulting for Google during the conduct of the study. Dr Eng reported a patent pending and ownership of Alphabet stock. Mr Nagpal reported ownership of Alphabet stock. Dr DeSalvo reported ownership of Alphabet stock. Dr Corrado reported ownership of Alphabet stock. Dr Peng reported ownership of Alphabet stock. Dr Webster reported ownership of Alphabet stock. Mr Dunn reported a patent pending and ownership of Alphabet stock. Mr Coz reported a patent pending and ownership of Alphabet stock. Dr Huang reported paid consulting for Google during the conduct of the study. Dr Yun Liu reported multiple patent pending ownership of Alphabet stock. Dr Bui reported a patent pending and ownership of Alphabet stock. Dr Yuan Liu reported a patent pending and ownership of Alphabet stock. No other disclosures were reported.

Funding/Support: This study was supported by Google LLC.

Role of the Funder/Sponsor: Google LLC was involved in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Additional Contributions: Sara Gabriele, BS, David Yen, BA, T. Saensukpaa, MHCI, ME, Carrie J. Cai, PhD, William Chen, BA, Quang Duong, PhD, Miles Hutson, BS, Dennis Ai, MS, Aaron Loh, MS, Bilson Campana, PhD, Jonathan Deaton, MS, Vivek Natarajan, MS, Ignacio Blanco, BS, Christopher Semturs, MS, Jessica Gallegos, MBA, Anita Misra, BTech, Roy Lee, BS, and all employees of Google LLC provided technical advice, discussion, and support. Peter Schalock, MD, and Sabina Bis, MD, both consultants for Google Health via Advanced Clinical, assisted with mapping free-text diagnoses to the structured list of 419 conditions. Justin Ko, MD, MBA, and Steven Lin, MD, Stanford Health Care, provided helpful discussions. No one was financially compensated for the stated contribution aside from their standard salary and associated compensation.

REFERENCES


SUPPLEMENT.

eMethods. Procedures and Metrics

eFigure 1. Detailed User Interface of the Artificial Intelligence (AI)-Based Assistive Tool

eFigure 2. States Where Readers Are Licensed to Practice

eFigure 3. Summary of Prior Clinical Experience of Primary Care Physician (PCP) and Nurse Practitioner (NP) Readers

eFigure 4. Diagnostic Accuracy of Primary Care Physicians (PCPs) and Nurse Practitioners (NPs)

eFigure 5. Sensitivity for the Top 26 Most Common Skin Conditions, With vs Without Artificial Intelligence (AI) Assistance

eFigure 6. Impact of Artificial Intelligence (AI) Assistance Stratified by the Position of the Correct Diagnosis in the Assistant's Interface

eFigure 7. Impact of Artificial Intelligence (AI) Assistance Stratified by the AI's Confidence Level (as Indicated by 5 Dots in the Assistant's Interface)

eFigure 8. Impact of Case Difficulty (as Measured by Interdermatologist Agreement Within the Panel Providing the Reference Diagnosis) on the Performance of Unassisted and Assisted Readers

eFigure 9. Changes During the Course of the Study for Top-1 Agreement, Top-3 Agreement, Average Overlap, and Median Diagnosis Time
eFigure 10. Impact of Artificial Intelligence (AI) Assistance Stratified by Estimated Fitzpatrick Skin Type for Top-1 Agreement, Top-3 Agreement, and Average Overlap

eFigure 11. Changes in Top-3 Diagnostic Accuracy by Primary Care Physicians (PCPs) and Nurse Practitioners (NPs) Without Artificial Intelligence (AI) Assistance vs When Using AI Assistance on Nonreferred Cases and Referred Cases

eFigure 12. Breakdown of Accuracy of Cases Based on Confidence of Primary Care Physicians (PCPs) and Nurse Practitioners (NPs)

eFigure 13. Sample Cases Evaluated Against Reference Diagnoses Provided by a Panel of 3 Dermatologists

eFigure 14. Sample Cases Evaluated Against Biopsy-Provided Diagnoses

eFigure 15. Reader Perceptions of the Value of Training Materials and Labeling Tool

eFigure 16. Reader Perceptions of the Value of Artificial Intelligence (AI) Assistance

eFigure 17. Perceived Usefulness of Different Features in the Artificial Intelligence (AI)-Based Assistance for Primary Care Physician (PCP) and Nurse Practitioner (NP) Readers

eTable 1. Clinical Metadata Used in This Study for Both AI Algorithm Input and Presented to Clinicians

eTable 2. Improvement in Top-1 Agreement and Average Overlap Between 2 Dermatologist Panels Using the Aggregated (“Voting”) Based Approach

eTable 3. Additional Metrics (Top-1 and Top-3 Agreement, Average Overlap, and Kappa) for Measuring Agreement Rates of PCPs and NPs, Respectively, With the Dermatologist Panel

eTable 4. Subgroup Analysis for 3 Categories of Skin Condition for Growths, Erythematous and Papulosquamous Skin Disease, and Hair Loss

eTable 5. Rates of Recommending a Biopsy and Referral for PCPs and NPs