JAMA Dermatology | Original Investigation

Development and Clinical Evaluation of an Artificial Intelligence Support Tool for Improving Telemedicine Photo Quality

Kailas Vodrahalli, BS; Justin Ko, MD, MBA; Albert S. Chiou, MD, MBA; Roberto Novoa, MD; Abubakar Abid, PhD; Michelle Phung, MS; Kiana Yekrang, BS; Paige Petrone, MS; James Zou, PhD; Roxana Daneshjou, MD, PhD

IMPORTANCE Telemedicine use accelerated during the COVID-19 pandemic, and skin conditions were a common use case. However, many images submitted may be of insufficient quality for making a clinical determination.

OBJECTIVE To determine whether an artificial intelligence (AI) decision support tool, a machine learning algorithm, could improve the quality of images submitted for telemedicine by providing real-time feedback and explanations to patients.

DESIGN, SETTING, AND PARTICIPANTS This quality improvement study with an AI performance component and single-arm clinical pilot study component was conducted from March 2020 to October 2021. After training, the AI decision support tool was tested on 357 retrospectively collected telemedicine images from Stanford telemedicine from March 2020 to June 2021. Subsequently, a single-arm clinical pilot study was conducted to assess feasibility with 98 patients in the Stanford Department of Dermatology across 2 clinical sites from July 2021 to October 2021. For the clinical pilot study, inclusion criteria for patients included being adults (aged ≥18 years), presenting to clinic for a skin condition, and being able to photograph their own skin with a smartphone.

INTERVENTIONS During the clinical pilot study, patients were given a handheld smartphone device with a machine learning algorithm interface loaded and were asked to take images of any lesions of concern. Patients were able to review and retake photos prior to submitting, so each submitted photo met the patient's assumed standard of clinical acceptability. A machine learning algorithm then gave the patient feedback on whether the image was acceptable. If the image was rejected, the patient was provided a reason by the AI decision support tool and allowed to retake the photos.

MAIN OUTCOMES AND MEASURES The main outcome of the retrospective image analysis was the receiver operator curve area under the curve (ROC-AUC). The main outcome of the clinical pilot study was the image quality difference between the baseline images and the images approved by AI decision support.

RESULTS Of the 98 patients included, the mean (SD) age was 49.8 (17.6) years, and 50 (51%) of the patients were male. On retrospective telemedicine images, the machine learning algorithm effectively identified poor-quality images (ROC-AUC of 0.78) and the reason for poor quality (blurry ROC-AUC of 0.84; lighting issues ROC-AUC of 0.70). The performance was consistent across age and sex. In the clinical pilot study, patient use of the machine learning algorithm was associated with improved image quality. An AI algorithm was associated with reduction in the number of patients with a poor-quality image by 68.0%.

CONCLUSIONS AND RELEVANCE In this quality improvement study, patients use of the AI decision support with a machine learning algorithm was associated with improved quality of skin disease photographs submitted for telemedicine use.

JAMA Dermatol. doi:10.1001/jamadermatol.2023.0091
Published online March 15, 2023.

Multimedia

Supplemental content

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Authors: James Zou, PhD, Stanford School of Medicine, Stanford, CA (jamesz@stanford.edu); Roxana Daneshjou, MD, PhD, Department of Dermatology, Stanford School of Medicine, 450 Broadway, Pavillion C, Redwood City, CA 94063 (roxanad@stanford.edu).

Remote clinical care (telemedicine) uses digital means to facilitate a clinical visit. Visits can happen in real time over video calls or asynchronously, with patients submitting images to be reviewed later. Skin conditions are prevalent with an estimated 1 in 3 Americans experiencing skin disease at any given time.[1] Both primary care physicians and dermatologists use telemedicine, specifically teledermatology, to assess skin conditions due to the visibility of the condition.

As video quality is not sufficient for assessing skin disease, patients are often asked to submit photos (eg, of their lesion or rash).[2] Most clinical photo-taking applications, including those used at Stanford Healthcare, primarily rely on the patient's judgment for submitting adequate quality photos. However, even when given instructions, patients frequently take photos of insufficient quality for clinical use.[3-5] This is partially due to a lack of experience with what features clinicians care most about.[3] Common quality issues include blurriness, poor lighting conditions, cropping of the area of interest, and too little or too much zoom.[3,6]

At Stanford Healthcare, images prior to the appointment are manually reviewed by clinical staff, requiring significant care-team time. If images are sent right before the virtual appointment, there is no time for human review. In some instances of persistent poor-quality images, patients are asked to be evaluated in-person when a video visit or asynchronous encounter may have been clinically appropriate.

Previous work[3,7] has proposed a user interface paired with an artificial intelligence (AI) algorithm that guides patients to take high-quality clinical photos using their smartphones. Such a user interface could provide real-time actionable feedback for patients, helping correct common image problems before they reach the clinician. However, clinical validation of these proposed tools has not been conducted.

In this study, we developed a clinical photo quality assessment algorithm for skin disease, improving on previous work using classic machine learning methods[3] by using additional training data sources and deep learning methods. Our machine learning AI algorithm, termed TrueImage, is an ensemble of deep learning-based models and classic computer vision algorithms meant to be used as an AI decision support tool (eFigure 1 in Supplement 1).

After training and validating the AI decision support tool on retrospective images, we conducted a single-arm self-controlled clinical pilot study using a user interface that allowed patients to receive real-time feedback as they took and submitted photos of their skin disease using a smartphone. For evaluation, we compared patients' initial image submissions with the image submissions after using the AI algorithm; clinicians rated each image based on their ability to make a clinical decision using the image. We assessed whether patient use of the AI algorithm was associated with improved photo quality of images taken by patients for telemedical use.

Key Points

Question Can an artificial intelligence support tool help patients with taking better skin lesion images for telemedicine use?

Findings In this quality improvement study including 98 patients and 357 images, a machine learning algorithm trained on retrospective telemedicine images was found to identify poor-quality images and the reason for poor quality. In the clinical pilot study, patients using a machine learning algorithm had a 68% reduction in the number of poor-quality images compared with baseline.

Meaning Results of this study suggest that artificial intelligence support tools could assist patients in taking photos for telemedicine use and lead to a higher percentage of sufficient-quality photos submitted.

Table 1. Retrospectively Collected Patient-Taken Image Quality Data Set

Variable	No. (%)	
	Entire data set	Test split
No. of patients	650	136
Age, mean (SD)	46.3 (18.2)	44.8 (17.6)
Sex		
Female	356 (54.8)	75 (55.2)
Male	294 (45.2)	61 (44.8)
Skin tone[a]		
FST I-III	544 (83.7)	112 (82.3)
FST IV-VI	106 (16.3)	24 (17.7)
No. of images	1700	357
Quality score, mean (SD)	1.3 (0.9)	1.2 (0.9)
Good quality	1060 (62.4)	228 (64.8)
Blurry	170 (10.0)	25 (7.1)
Lighting issues	209 (12.3)	46 (13.1)
Zoom/crop issues	154 (9.1)	26 (7.4)
Other issues	107 (6.3)	27 (7.7)

Abbreviation: FST, Fitzpatrick skin type.

[a] Fitzpatrick classification assesses phototypes I (pale white skin, blue or green eyes, and blonde or red hair) and II (fair skin and blue eyes); phototypes III (darker white skin) and IV (light brown skin); and phototypes V (brown skin) and VI (dark brown or black skin).

Methods

This quality improvement study with an AI performance component and single-arm clinical pilot study component was conducted from March 2020 to October 2021. The AI decision support tool was developed using annotated telemedicine clinical images from Stanford Healthcare. The algorithm is an ensemble of multiple deep learning and classic computer vision algorithms. It can classify dermatology images as poor quality and give a reason for the poor quality from 3 options: blur, poor lighting, and other. Blurriness and poor lighting are empirically observed to be the most common reasons attributed to poor quality (**Table 1**).

Retrospective Image Data Set

To train the algorithm, we collected 1700 images of skin disease from 650 patients who had Stanford Healthcare telemedicine visits from March 2020 to June 2021 (Table 1). These images were submitted by the patient as part of the patient's

dermatology telemedicine visit. This study was conducted under the Stanford institutional review board, which granted a waiver of consent for the retrospective use of deidentified patient images for the development of AI algorithms. Images were stored on a Health Insurance Portability and Accountability Act-compliant server.

We annotated each image for photo quality using a Likert-like scale (eTable 1 in Supplement 1) developed in consultation with 4 board-certified dermatologists. The focus of this scale is the ability to make a clinical determination. Images are scored from 0 to 4, with each number corresponding to a letter grade given by the clinicians (A, B, C, D, and F). Images rated 0 to 1 (A-B) are considered good quality. Images rated 2 to 4 (C-F) are considered poor quality or inadequate for clinical decisions. In our analysis, we referenced image quality by their numerical grade. Images labeled as poor quality were also annotated with the reasons for poor quality: (1) blurriness, (2) lighting condition, (3) inadequate or excessive zoom and/or cropping of area of interest, or (4) other. Images were annotated by Stanford dermatology residents, matching the current Stanford Healthcare workflow in which residents are also tasked with manual image quality assessments.

Note that image quality is defined relative to whether a clinical assessment can be made from it. This has several ramifications. First, poor quality in background regions is generally acceptable. Quality is relative to the type of lesion or rash (eg, in assessing quality, we are implicitly classifying disease subgroup). Moreover, quality is subjective, as it is relative to a clinicians' comfort in making an assessment.

Splitting Our Retrospective Data Set

The image data set was split into 3 subsets—train (54% of images), validation (25%), and test (21%)—prior to training our algorithm (Table 1). The data split was conducted at a patient level, so all images of any individual patient were contained in a single split. The test split was used only for final evaluation.

Model Design

The AI decision support tool is built using an ensemble of deep learning models and classic computer vision algorithms (eFigure 2 in Supplement 1). The ensemble is a weighted sum across the individual model's predictions, with the weighting fitted on validation data.

The final output of the model is an overall classification of good or poor quality and if poor quality, an explanation for the poor quality. There are 4 possible explanations our algorithm can give for poor image quality: blur (the image is too blurry), lighting (issues with poor lighting), zoom or crop (image is too zoomed in or out), or other. The "other" explanation resulted when an image identified as a poor-quality image did not have other detectable issues. Note that each image may have multiple reasons for poor quality.

Deep Learning Models

Each deep learning-based model in our ensemble is an instance of the ResNet-18 model architecture.[8] These models were trained independently with different random seeds and with slight variations in hyperparameters. The final linear layer of each model is replaced by 4 separate linear classifiers to predict quality and reason for poor quality. Each classifier is responsible for a single binary prediction–(1) good/poor quality overall, and (2-4) good/poor blur, lighting, or zoom/crop. During evaluation, the first classifier serves to gate the predictions of the remaining 3. Detailed information on the model design and training are included in the eMethods of Supplement 1.

Classic Vision Models

The classic vision algorithms include logistic classifiers, support vector machine classifiers, and random forest classifiers. The algorithms and model hyperparameters were chosen through cross-validation. A separate model was trained for each of 4 binary classification decisions: (1) good/poor quality overall, and (2-4) good/poor blur, lighting, or zoom/crop. These are the same classification decisions made by the deep learning models.

Models are input hand-selected features designed to differentiate poor-quality images. There are 2 sets of features we used. The first set is primarily based around using local binary patterns[9] on the skin regions of the image; it also focuses only on the center region of the image. The second set is based around featurizing each region on a 5 × 5 grid in the image.

The featurization methods were chosen through validation studies, with the most informative features being kept. A more detailed description of the featurization methods used is described in the eMethods and eTable 5 in Supplement 1.

Model Ensemble and Threshold Parameters

The AI decision support tool consists of 4 deep learning models and 6 classic vision algorithms in an ensemble. The number of models used was chosen through validation studies.

The ensemble model takes each of the 10 model outputs and computes a weighted sum of their final predictions. This weighting is fit on the train subset of the retrospective data set. It is fit separately for each of the 4 classification decisions.

We also determine decision threshold parameters using a held-out validation set at this stage. For each classification decision by the ensemble, a single scalar value is selected to convert the continuous model output into a binary decision. The threshold was chosen to maximize the correctly identified poor-quality images (true-positive rate) while preventing the false-negative rate from rising too high. That is, we primarily focused on clinician benefit but did not want to place an undue burden on patients. The operating point was then manually adjusted using data from the clinical setting prior to the start of the study to counteract distribution shift. No changes to the algorithm or operating point were made during the clinical pilot study.

Clinical Pilot Study Design

To assess whether the AI decision support tool could aid patients to take better images for clinical use, we performed a single-arm self-controlled clinical pilot study. The aim of this study was to assess the feasibility of using the AI algorithm in the clinical setting.

To calculate the number of patients required to show a significant effect, we used a 2-sided *t* test power calculation (α = .05, power = .8). We took the retrospective patient data set as a representative sample of patient-taken images and assumed a 1-point quality improvement in at least 60% of the photos. Using the *t* test calculation, we found we needed at least 11 samples of poorly taken photos. In the retrospective data set, 37.7% of photos were poor quality, so we arrived at requiring at least 30 patients in our pilot study.

We exceeded this number, further enabling our analysis. In total, we collected data on 98 patients over a 4-month period (July-October 2021). Data were collected in a clinic using a handheld smartphone device (iPhone 12 [Apple Inc]). We report the results using the DECIDE-AI checklist.

User Interface

We created a web interface with a simple user interface for the clinical pilot study using Gradio.[10] The interface provides patients with the ability to take and select photos, submit the photos to our server for quality assessment, and receive textual feedback (eFigure 1 in Supplement 1). We logged the submitted images and the model's output. These logs were used to conduct the clinical pilot study analysis.

This interface was intended to be a minimal working implementation to test the real-time performance of the AI algorithm and, as such, no user studies were conducted. In the clinical pilot study, we relied on a clinical coordinator to ensure the interface usage was adequately understood by patients.

Patient Recruitment

Patients were recruited from Stanford Dermatology Clinics at 2 separate clinical sites and prospectively gave consent to the study team. Data on Fitzpatrick skin type (FST) but not race and ethnicity were collected. Inclusion criteria for participants included being adults (aged ≥18 years), presenting to clinic for a skin condition, and being able to photograph their own skin with a smartphone. Patients who could not read or write in English or provide their own informed consent were excluded. Investigators asked patients who met the inclusion criteria if they would be interested in participating. Interested patients gave consent to the study team and were given a handheld smartphone with the AI algorithm interface loaded (eFigure 1 in Supplement 1). Each patient received a unique sign-in for the interface. To simulate the situation in which patients take images for telemedicine, patients were asked to take an image of the skin condition that brought them to the clinic that day using the AI algorithm interface. Patients were able to review and retake photos prior to submitting, so each submitted photo met the patient's assumed standard of clinical acceptability. The AI algorithm would then give the patient feedback on whether the image was acceptable. If the image was rejected, the patient was provided a reason by the AI algorithm and allowed to retake the photo. Patients who did not produce an acceptable photo after 4 attempts were not asked to take additional photos, and instead the AI algorithm selected the best photo among those submitted.

Data Set Labeling

The data set of clinical pilot study images was labeled for quality annotations by 3 of the authors (A.C., J.K., and R.D.) using the same annotation procedure as was used for the retrospective data (eTable 1 in Supplement 1). All 3 authors are board-certified dermatologists (A.C. with 5 years of postresidency experience, J.K. with 10, and R.D. with 1).

The labels are generally concordant across the 3 labelers (eTable 2 in Supplement 1). When labelers disagreed, the disagreements were typically by 1 point on the quality scale; moreover, the disagreements typically did not cross the good or bad quality threshold, with labelers agreeing on 70% to 85% of binary label quality. In our presented analysis, we selected the median label as the ground-truth assessment.

Statistical Analysis

The AI algorithm performance was assessed using receiver operator curve area under the curve (ROC-AUC). In the retrospective image analysis, we assessed differences in performance as measured by ROC-AUC in subgroups defined by skin tone, age, and sex using the DeLong test,[11] with an implementation of the algorithm described by Sun and Xu[12,13] and $P < .05$ considered significant. To assess power, we performed a post hoc power calculation using simulated data, where we assumed a gaussian distribution for each binary response. Our analysis showed that for sex and age, we were powered ($\beta > 80\%$ and $\alpha < .05$) to detect a 0.15 decrease in AUC but were underpowered for skin tone.

For the clinical pilot study, we assessed the AI algorithm performance by applying the Wilcoxon signed rank test to image quality data. In particular, for each individual we looked at the paired difference between initial and final image quality, where the final image was selected after the individual used the AI algorithm. We considered $P < .05$ as significant.

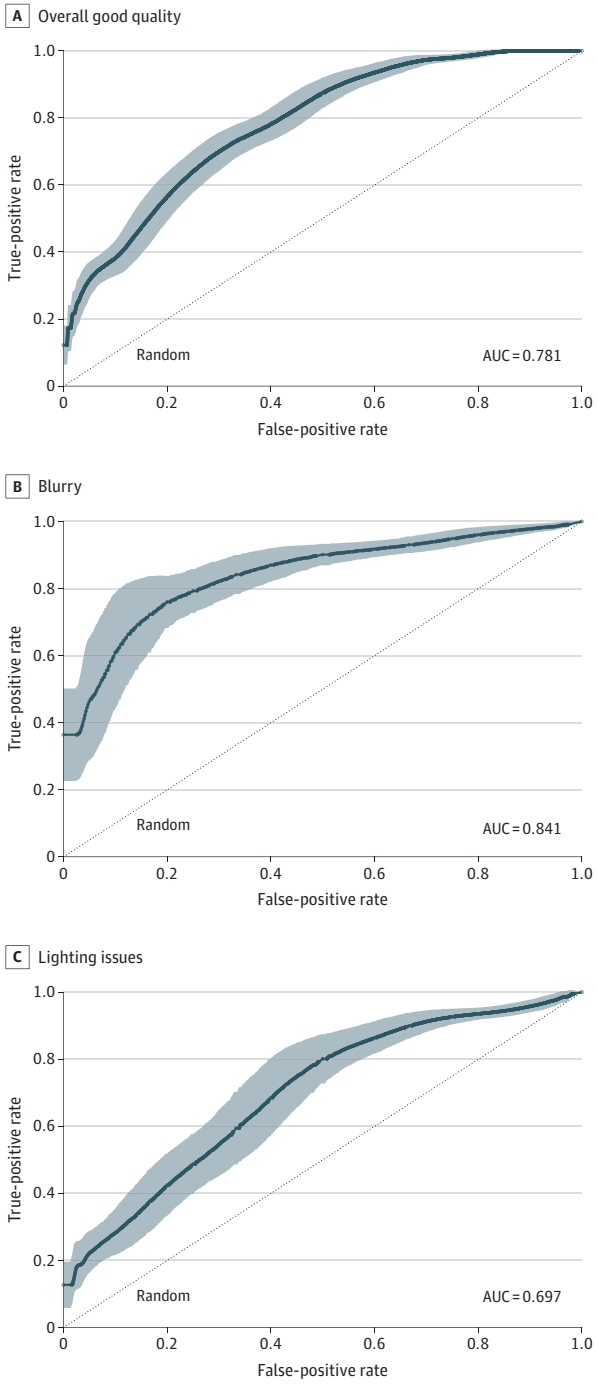
Results

Retrospective Data Analysis

This quality improvement study included 98 patients and 357 images. The AI algorithm can distinguish between poor-quality and good-quality images on the retrospectively collected telemedicine images. On the retrospective data test set, we observed AUC values of 0.781 (overall quality), 0.841 (blurry), and 0.697 (lighting issues) (**Figure 1**).

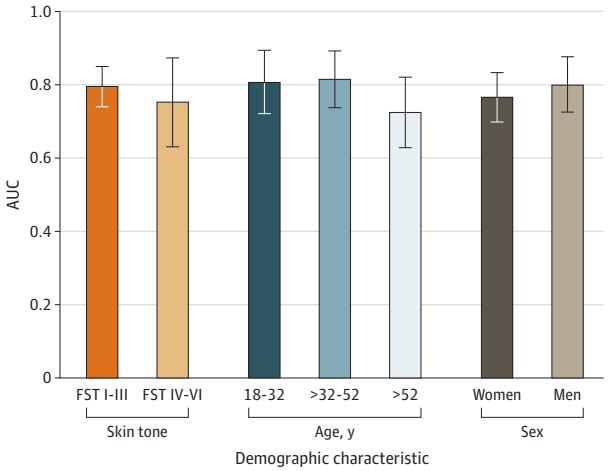
We also analyzed AI algorithm performance across demographic subgroups in the retrospective data set. In **Figure 2**, we compared ROC-AUC across (1) diverse skin tones, (2) age, and (3) sex. We evaluated skin type into 2 groups based on FST: FST I-III (with phototypes I [pale white skin, blue or green eyes, and blonde or red hair], II [fair skin and blue eyes], and III [darker white skin]) and FST IV-VI (with phototypes IV [light brown skin], V [brown skin], and VI [dark brown or black skin]). For skin tone, FST I-III had an ROC-AUC of 0.794 (n = 289) and FST IV-VI had an ROC-AUC of 0.751 (n = 63) ($P = .52$; DeLong test). There was no statistical difference between younger patients (aged 18-32 y: n = 112, AUC of 0.806; aged 32-52 y: n = 120, ROC-AUC of 0.814) and older patients (aged >52 y:

Figure 1. Retrospective Data Evaluation



A, Overall good quality. B, Blurry. C, Lighting issues. Receiver operator curve area under the curve plots showing performance of the quality classifier and quality explanation classifiers. AUC indicates area under the curve.

Figure 2. Retrospective Data Evaluation



Comparison of receiver operator curve area under the curve across demographic splits for overall quality classifier. The smaller the gap within a group, the more robust our model is across that group. Error bars indicate 95% CIs. AUC indicates area under the curve; FST, Fitzpatrick skin type (with phototypes I [pale white skin, blue or green eyes, and blonde or red hair], II [fair skin and blue eyes], III [darker white skin], IV [light brown skin], V [brown skin], and VI [dark brown or black skin]).

n = 120, ROC-AUC of 0.723) (P = .14; DeLong test) and no statistical difference was found between male patients (n = 159, ROC-AUC of 0.800) and female patients (n = 193, AUC of 0.766) (P = .51; DeLong test).

As an additional analysis of data external to Stanford, we performed a retrospective analysis on skin images pulled from the internet. Our model performance was consistent with the observed performance on the clinical pilot study data, suggesting our model may generalize well. More discussion is provided in the eMethods section of Supplement 1.

Clinical Pilot Study Analysis

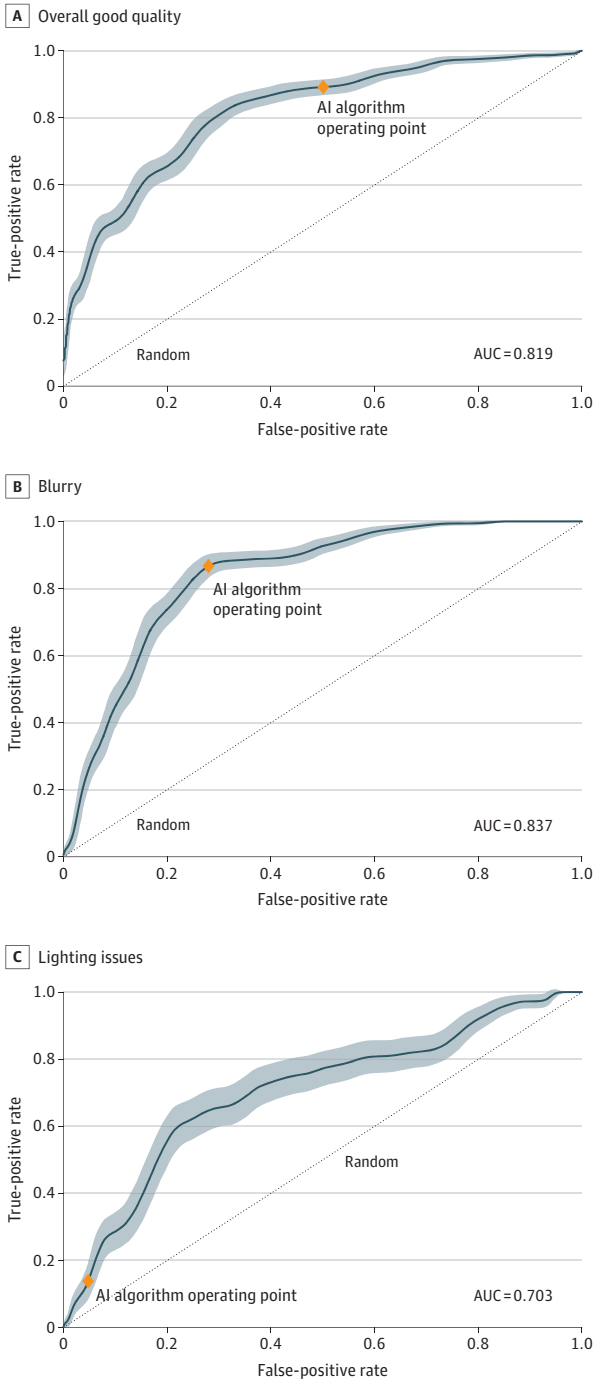
Ninety-eight patients were recruited. The mean (SD) age was 49.8 (17.6) years, and 51% of the patients were male. The skin tone composition of the patients and patient-taken images are detailed in eTable 3 and eTable 4 in Supplement 1. Since patients were limited to 4 attempts, a portion of the patients (n = 13) did not generate an image that was considered good quality by the AI algorithm; for these patients, the best quality image as determined by the AI algorithm was submitted as the final image. Overall, patients took a mean (SD) of 1.7 (0.9) images and spent an additional mean (SD) of 30 (51.9) seconds taking additional photos (eTable 4 in Supplement 1).

Among the initial image submissions, 65.3% were good quality and the mean (SD) quality score was 1.15 (0.98). An AI algorithm reduced the number of patients with a poor-quality image by 68.0%. The predominant reason for poor quality was blur in the area of interest, followed by lighting issues and zoom/cropping issues where the lesion was not adequately shown in the photo. We analyzed the performance of the AI algorithm both at the patient level and individual image level.

Image-Level Analysis

We plotted ROC-AUC performance of the AI algorithm across all images submitted by patients in **Figure 3.** Quality labels were assessed by our labelers. We achieved a ROC-AUC of 0.819 for assessing overall quality of an image, and a ROC-AUC of 0.837

Figure 3. Clinical Pilot Study



A, Overall good quality. B, Blurry. C, Lighting issues. Plots showing performance of the quality classifier and quality explanation classifiers. Receiver operator curve area under the curve computed across all images captured in the clinical pilot study. Operating point used in the clinical study is marked. AI indicates artificial intelligence; AUC, area under the curve.

for identifying blur, and a ROC-AUC of 0.703 for identifying poor lighting. These AUC values are similar to those observed during the retrospective image analysis, suggesting our algorithm generalized well during the clinical pilot study setting.

Patient-Level Analysis

We also analyzed the patient-level benefit of the AI decision support tool. We compared the quality of the initial and final images submitted by patients. Note that when the AI algorithm assesses the first image to have good quality, the initial and final images are identical. For patients for whom the AI algorithm never assessed an image to have good quality, we selected the image with best quality for analysis as assessed by the AI algorithm. We performed analysis for all patients as well as the subgroup for whom the AI algorithm identified a good-quality image (eTable 4 in Supplement 1).

Across all patients, use of the AI decision support tool was associated with a significant improvement in photo quality (**Table 2**). Use of the AI decision support tool was associated with improved quality for patients whose initial image was rated a 2 (P = .003, Wilcoxon signed rank test, with a mean [SD] improvement of 0.71 [0.95]; a 1-point improvement was needed for good quality) and 3 (P = .01, Wilcoxon signed rank test, with a mean [SD] improvement of 1.75 [1.09]; a 2-point improvement was needed for good quality). These findings are important for clinical care as they correspond to a reduction in the number of patients with poor-quality images: 54% (2-quality initial image), 70% (3-quality initial image), and 56% overall. There were no images with a score of 4 (the worst score) in the clinical pilot study.

Discussion

COVID-19 rapidly accelerated the adoption of telehealth with an initial 78-fold in telehealth use compared with prior to the pandemic.[14] This rapid uptake has led to changes in regulatory policies and increased familiarity with telehealth among clinicians and patients. Thus, even as medical practices have begun seeing patients in person again, telehealth usage is still 38 times higher than before the pandemic.[14]

Telehealth visits for skin disease often require patients to send in a photograph since video quality is inadequate for skin disease assessment. However, dermatologists have reported that patient photos that do not meet the quality standard for making a clinical assessment interrupt the flow of clinical care.[3] Moreover, the influx of low quality images can lead to increased physician time and burnout.[5,15] In a retrospective assessment of 1700 images of skin disease from 650 patients who had Stanford Healthcare telemedicine visits from March 2020 to June 2021, we found that 37.6% of images did not meet the quality threshold for making a clinical assessment. This finding is in line with a 2022 study[5] in which dermatologists assessed 1200 telemedicine-submitted images and found that 37.8% were of insufficient quality. In the present study, we found the most common reasons for insufficient quality were blurry images or poor lighting. The current standard for image quality assessment is manual review after images are submitted, which requires a substantial amount of time and may not be sustainable long term. Moreover, images sent right before the appointment cannot be reviewed for quality prior to the virtual encounter, potentially disrupting clinical care.

Table 2. Clinical Pilot Study Data Evaluation

	Quality score[b]				
Initial quality[a]	0	1	2	3	4
No. of patients	30	34	24	10	0
Quality improvement, mean (SD)	−0.03 (0.18)	0.09 (0.38)	0.71 (0.95)[c]	1.75 (1.09)[c]	NA

Abbreviation: NA, not applicable.

[a] Initial quality scores and the mean (SD) quality improvement after using a machine learning algorithm are reported. A positive quality improvement corresponds to an increase in image quality. Statistically significant improvement is seen for the worst quality images: those with an initial quality score of 2 (P = .003) or 3 (P = .01). None of the images in the study received the lowest quality grade, 4.

[b] Photo scoring: 0 = crisp, clear, perfect photo; 1 = generally good quality with minor imperfections, but I can tell what is happening; 2 = I think I can tell what is going on, but the quality is not great; 3 = can barely discern what is happening in the photo; 4 = cannot tell what is going on in the photo.

[c] Statistically significant improvement.

We developed an AI algorithm on retrospective telehealth data that could identify skin images that were of insufficient quality for making a clinical determination. We focused on the outcome important to clinicians: the ability to assess the skin disease. On retrospective data, we found that the AI decision support tool could identify insufficient quality patient-captured images sent for telemedicine use. We found similar retrospective performance on an external data set of skin images pulled from publicly available data on the internet, suggesting generalizability of the AI decision support tool.

However, algorithms that perform well on retrospective data will often have a performance degradation in clinical practice.[16] Additionally, the AI algorithm is an algorithm that interacts with the user; its efficacy is based on the user taking the feedback and producing an improved image. Thus, the pilot study was key for assessing the clinical utility of the AI algorithm.

We found that real-time algorithmic feedback was associated with an improvement in the quality of skin disease images taken by patients in our early clinical evaluation. Our findings suggest that this algorithm could serve as a useful tool for improving the quality of images sent by patients for telemedicine evaluation. In turn, this could reduce the manual labor of clinicians having to review photos beforehand. As the scope of telemedicine has also recently grown to include remote clinical trials, decision support AI tools could likewise be useful for remote trials involving skin disease.

Limitations

This early clinical pilot study has limitations: (1) it was conducted in a clinic, where lighting is more ideal than the at-home setting, (2) patients were provided feedback about why their image was of insufficient quality but not with instructions on how to improve the images, (3) a lack of power to do subgroup analysis, such as by skin tone, sex, or age. We did not compare with the current standard of manual review, as this baseline is not a sustainable model, (4) a focus on English-speaking patients, (5) evaluation at a single institution. Future studies will assess how giving advice through AI decision support tools (eg, "please tap to focus the camera") affects patients' photo taking behavior after taking a poor-quality image. Moreover, future iterations will include translations to other languages.

While we show the AI decision support tool performed similarly across demographic characteristics in the retrospective analysis, this retrospective analysis was not specifically powered for skin tone analysis. A larger trial is needed in the future, although a lack of differential performance is notable. Representing diversity of skin tones in AI trials is important, with most previous AI algorithms being built from data that lack FST IV-VI.[17,18] Both our retrospective and prospective studies drew from the Stanford patient population without targeting specific demographic characteristics. We note a lack of FST VI in the prospective trial as a limitation that needs to be addressed in future larger trials. Additionally, we did not have associated disease labels with the images; however, since we did not target any specific population, our data likely represents the diverse spectrum of disease assessed at Stanford Dermatology.

Conclusions

Results of this quality improvement study suggest that artificial intelligence support tools could assist patients in taking photos for telemedicine use and lead to a higher percentage of sufficient-quality photos submitted.

ARTICLE INFORMATION

Accepted for Publication: January 12, 2023.

Published Online: March 15, 2023. doi:10.1001/jamadermatol.2023.0091

Author Affiliations: Department of Electrical Engineering, Stanford University, Stanford, California (Vodrahalli, Zou); Department of Dermatology, Stanford School of Medicine, Redwood City, California (Ko, Chiou, Novoa, Phung, Yekrang, Petrone, Daneshjou); Department of Pathology, Stanford School of Medicine, Stanford, California (Novoa); Hugging Face, New York, New York (Abid); Department of Biomedical Data Science, Stanford School of Medicine, Stanford, California (Zou, Daneshjou); Department of Computer Science, Stanford University, Stanford, California (Zou); Chan-Zuckerberg Biohub, San Francisco, California (Zou).

Author Contributions: Dr Daneshjou and Mr Vodrahalli had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Drs Zou and Daneshjou contributed equally.
Concept and design: Vodrahalli, Ko, Zou, Daneshjou.
Acquisition, analysis, or interpretation of data: Vodrahalli, Ko, Chiou, Novoa, Abid, Phung, Yekrang, Petrone, Daneshjou.
Drafting of the manuscript: Vodrahalli, Chiou, Petrone, Daneshjou.
Critical revision of the manuscript for important intellectual content: Vodrahalli, Ko, Chiou, Novoa, Abid, Phung, Yekrang, Zou, Daneshjou.

Statistical analysis: Vodrahalli, Daneshjou.
Obtained funding: Ko, Daneshjou.
Administrative, technical, or material support: Ko, Phung, Yekrang, Petrone, Zou, Daneshjou.
Supervision: Ko, Chiou, Novoa, Zou, Daneshjou.

Conflict of Interest Disclosures: Mr Vodrahalli reported receiving grants from the National Science Foundation (NSF) Graduate Research Fellowships Program during the conduct of the study; in addition, Mr Vodrahalli had a patent for Automated Clinical Image Quality Assessment pending for the work in this study. Dr Ko reported receiving grants from Stanford Medicine in support of the work through the Catalyst program during the conduct of the study; personal fees from Enspectra, grants from Google Health Research collaboration, and personal fees from Skin Analytics outside the submitted work; in addition, Dr Ko reported having a patent for US Patent Application 17/937,714 pending; and being a chair of the American Academy of Dermatology Committee on Augmented Intelligence. Dr Chiou reported holding a patent for Systems and Methods for Automated Clinical Image Quality Assessment and a pending provisional patent on TrueImage. Dr Novoa reported receiving grants from Melanoma Research Alliance during the conduct of the study; in addition, Dr Novoa reported holding a patent for US patent 17/937,714 pending. Dr Zou reported holding a patent for US patent 17/937,714 pending. Dr Daneshjou reported receiving personal fees from DWA, personal fees from Pfizer, personal fees from L'Oreal, personal fees from VisualDx, stock options from MDAlgorithms and Revea outside the submitted work; in addition, Dr Daneshjou reported holding a patent for TrueImage pending. No other disclosures were reported.

Funding/Support: Mr Vodrahalli is supported by an NSF graduate research fellowship and a Stanford Graduate Fellowship award. Drs Ko, Chiou, and Novoa are supported by the Melanoma Research Alliance's L'Oréal Dermatological Beauty Brands-MRA Team Science Award. Dr Chiou is supported by a Dermatology Foundation Medical Dermatology Career Development Award. Dr Zou is supported by NSF CAREER 1942926. Dr Daneshjou is supported by 5T32AR007422-38 and the Stanford Catalyst Program.

Role of the Funder/Sponsor: The funding organizations had no role 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.

Data Sharing Statement: See Supplement 2.

REFERENCES

1. Wilmer EN, Gustafson CJ, Ahn CS, Davis SA, Feldman SR, Huang WW. Most common dermatologic conditions encountered by dermatologists and nondermatologists. *Cutis*. 2014;94(6):285-292.

2. Briggs SM, Lipoff JB, Collier SM. Using implementation science to understand teledermatology implementation early in the COVID-19 pandemic: cross-sectional study. *JMIR Dermatol*. 2022;5(2):e33833. doi:10.2196/33833

3. Vodrahalli K, Daneshjou R, Novoa RA, Chiou A, Ko JM, Zou J. *TrueImage: A Machine Learning Algorithm To Improve the Quality of Telehealth Photos*. Biocomputing; 2021:220-231.

4. Irvine E, Sayed L, Johnson N, Dias J. The ability of patients to provide standardized, patient-taken photographs for the remote assessment of Dupuytren disease. *Hand (N Y)*. 2023;18(1):139-144. doi:10.1177/15589447211006834

5. Jiang SW, Flynn MS, Kwock JT, et al. Quality and perceived usefulness of patient-submitted store-and-forward teledermatology images. *JAMA Dermatol*. 2022;158(10):1183-1186. doi:10.1001/jamadermatol.2022.2815

6. Muraco L. Improved medical photography: key tips for creating images of lasting value. *JAMA Dermatol*. 2020;156(2):121-123. doi:10.1001/jamadermatol.2019.3849

7. TensorFlow.js. How DermAssist uses TensorFlow.js for on-device image quality checks. October 11, 2021. Accessed May 9, 2022. https://blog.tensorflow.org/2021/10/how-DermAssist-uses-TensorFlowJS.html

8. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. Accessed February 1, 2023. https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf

9. Guo Z, Zhang L, Zhang D. A completed modeling of local binary pattern operator for texture classification. *IEEE Trans Image Process*. 2010;19(6):1657-1663. doi:10.1109/TIP.2010.2044957

10. Abid A, Abdalla A, Abid A, Khan D, Alfozan A, Zou J. Gradio: hassle-free sharing and testing of ML models in the wild. *ArXiv*. Preprint posted online June 6, 2019. doi:10.48550/arXiv.1906.02569

11. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837-845. doi:10.2307/2531595

12. Sun X, Xu W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Process Lett*. 2014;21(11):1389-1393. doi:10.1109/LSP.2014.2337313

13. Kazeev N. DeLong Test Implementation. GitHub. Accessed February 1, 2023. https://github.com/yandexdataschool/roc_comparison

14. McKinsey & Co. Telehealth: a quarter-trillion-dollar post-COVID-19 reality? July 9, 2021. Accessed July 1, 2022. https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/telehealth-a-quarter-trillion-dollar-post-covid-19-reality

15. Borre ED, Nicholas MW. The disproportionate burden of electronic health record messages with image attachments in dermatology. *J Am Acad Dermatol*. 2022;86(2):492-494. doi:10.1016/j.jaad.2021.09.026

16. Han SS, Kim YJ, Moon IJ, et al. Evaluation of artificial intelligence-assisted diagnosis of skin neoplasms: a single-center, paralleled, unmasked, randomized controlled trial. *J Invest Dermatol*. 2022;142(9):2353-2362.e2. doi:10.1016/j.jid.2022.02.003

17. Wen D, Khan SM, Ji Xu A, et al. Characteristics of publicly available skin cancer image datasets: a systematic review. *Lancet Digit Health*. 2022;4(1):e64-e74. doi:10.1016/S2589-7500(21)00252-1

18. Daneshjou R, Smith MP, Sun MD, Rotemberg V, Zou J. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review. *JAMA Dermatol*. 2021;157(11):1362-1369. doi:10.1001/jamadermatol.2021.3129