JAMA Network Open

Original Investigation | Dermatology

Assessment of Accuracy of an Artificial Intelligence Algorithm to Detect Melanoma in Images of Skin Lesions

Michael Phillips, MMedSci; Helen Marsden, PhD; Wayne Jaffe, MB, ChB, FRCS; Rubeta N. Matin, PhD, MBBS, MRCP; Gorav N. Wali, MA, BMBCh, MRCP; Jack Greenhalgh, PhD; Emily McGrath, BMBS, MRCP; Rob James, BSc, RGN; Evmorfia Ladoyanni, PGCME, DTM&H, FRCP; Anthony Bewley, MB ChB, FRCP; Giuseppe Argenziano, MD, PhD; Ioulios Palamaras, MD, PhD

Abstract

IMPORTANCE A high proportion of suspicious pigmented skin lesions referred for investigation are benign. Techniques to improve the accuracy of melanoma diagnoses throughout the patient pathway are needed to reduce the pressure on secondary care and pathology services.

OBJECTIVE To determine the accuracy of an artificial intelligence algorithm in identifying melanoma in dermoscopic images of lesions taken with smartphone and digital single-lens reflex (DSLR) cameras.

DESIGN, SETTING, AND PARTICIPANTS This prospective, multicenter, single-arm, masked diagnostic trial took place in dermatology and plastic surgery clinics in 7 UK hospitals. Dermoscopic images of suspicious and control skin lesions from 514 patients with at least 1 suspicious pigmented skin lesion scheduled for biopsy were captured on 3 different cameras. Data were collected from January 2017 to July 2018. Clinicians and the Deep Ensemble for Recognition of Malignancy, a deterministic artificial intelligence algorithm trained to identify melanoma in dermoscopic images of pigmented skin lesions using deep learning techniques, assessed the likelihood of melanoma. Initial data analysis was conducted in September 2018; further analysis was conducted from February 2019 to August 2019.

INTERVENTIONS Clinician and algorithmic assessment of melanoma.

MAIN OUTCOMES AND MEASURES Area under the receiver operating characteristic curve (AUROC), sensitivity, and specificity of the algorithmic and specialist assessment, determined using histopathology diagnosis as the criterion standard.

RESULTS The study population of 514 patients included 279 women (55.7%) and 484 white patients (96.8%), with a mean (SD) age of 52.1 (18.6) years. A total of 1550 images of skin lesions were included in the analysis (551 [35.6%] biopsied lesions; 999 [64.4%] control lesions); 286 images (18.6%) were used to train the algorithm, and a further 849 (54.8%) images were missing or unsuitable for analysis. Of the biopsied lesions that were assessed by the algorithm and specialists, 125 (22.7%) were diagnosed as melanoma. Of these, 77 (16.7%) were used for the primary analysis. The algorithm achieved an AUROC of 90.1% (95% CI, 86.3%-94.0%) for biopsied lesions and 95.8% (95% CI, 94.1%-97.6%) for all lesions using iPhone 6s images; an AUROC of 85.8% (95% CI, 81.0%-90.7%) for biopsied lesions and 93.8% (95% CI, 91.4%-96.2%) for all lesions using Galaxy S6 images; and an AUROC of 86.9% (95% CI, 80.8%-93.0%) for biopsied lesions and 91.8% (95% CI, 87.5%-96.1%) for all lesions using DSLR camera images. At 100% sensitivity, the algorithm achieved

(continued)

Key Points

Question How accurate is an artificial intelligence–based melanoma detection algorithm, which analyzes dermoscopic images taken by smartphone and digital single-lens reflex cameras, compared with clinical assessment and histopathological diagnosis?

Findings In this diagnostic study, 1550 images of suspicious and benign skin lesions were analyzed by an artificial intelligence algorithm. When compared with histopathological diagnosis, the algorithm achieved an area under the receiver operator characteristic curve of 91.8%. At 100% sensitivity, the algorithm achieved a specificity of 64.8%, while clinicians achieved a specificity of 69.9%.

Meaning As the burden of skin cancer increases, artificial intelligence technology could play a role in identifying lesions with a high likelihood of melanoma.

+ Supplemental content

Author affiliations and article information are listed at the end of this article.

Abstract (continued)

a specificity of 64.8% with iPhone 6s images. Specialists achieved an AUROC of 77.8% (95% CI, 72.5%-81.9%) and a specificity of 69.9%.

CONCLUSIONS AND RELEVANCE In this study, the algorithm demonstrated an ability to identify melanoma from dermoscopic images of selected lesions with an accuracy similar to that of specialists.

JAMA Network Open. 2019;2(10):e1913436. doi:10.1001/jamanetworkopen.2019.13436

Introduction

When compared with other forms of skin cancer, malignant melanoma is relatively uncommon; however, the incidence of melanoma is increasing faster than any other form of cancer, and it is responsible for the majority of skin cancer deaths.[1] Patients in whom melanoma is diagnosed at stage I have more than a 95% 5-year relative survival rate compared with 8% to 25% when the disease is diagnosed at stage IV.[1,2]

A person with a suspicious pigmented skin lesion will go through several steps before a definitive diagnosis of melanoma: self-evaluation, evaluation by a primary care physician, assessment by a specialist, and excision and assessment by histopathology. Current practice guidelines recommend appropriately trained health care professionals assess all suspicious skin lesions using dermoscopy.[3] Techniques to improve diagnostic accuracy can be used at each step to improve differentiation between harmless and potentially harmful lesions, thus reducing the pressure on later services from lesions that were unnecessarily referred further.[4-6] The Cochrane Skin Cancer Diagnostic Test Accuracy Group recently published a comprehensive series of reviews on the accuracy of different diagnostic techniques, including visual assessment with or without dermoscopy,[7] reflectance confocal microscopy,[8] teledermatology,[9] computer-aided diagnostic techniques,[10] and smartphone applications.[11] Dermoscopy was found to improve diagnostic accuracy over visual inspection alone, and reflectance confocal microscopy was more accurate than dermoscopy alone, but the data supporting the widespread adoption of teledermatology, computer-assisted diagnostic techniques, or smartphone applications are limited and/or of poor quality. Clinical diagnostic accuracy also depends on the experience of the examiners,[12] and the equipment required for reflectance confocal microscopy is expensive.

A large number of smartphone applications for melanoma detection have been released, although there is little evidence of clinical validation.[11,13] Of 39 skin cancer applications assessed, 19 involved smartphone photography and 4 provided an estimate of the probability of malignancy, but none were assessed for diagnostic accuracy.[14] Poorly designed, inaccurate, and/or misleading consumer applications may cause harm to patients.[15-18]

However, with appropriate development and suitable evaluation, modern electronic technology could improve diagnostic accuracy. Indeed, artificial intelligence (AI) algorithms categorizing photographs of lesions have recently been shown to be capable of classifying melanoma with a level of competence comparable with dermatologists.[19,20] Deep Ensemble for Recognition of Malignancy, developed by Skin Analytics Limited, is an AI algorithm that is designed to be used as a decision support tool for health care professionals by determining the likelihood of skin cancer from dermoscopic images of pigmented skin lesions. It was developed using deep learning techniques that identify and assess features of lesions that are associated with melanoma, using more than 7000 archived dermoscopic images, and it has been shown to identify melanoma with accuracy similar to that of specialist physicians.[21]

The aim of this study was to evaluate the ability of the Deep Ensemble for Recognition of Malignancy algorithm to detect melanoma from images of both biopsied and nonbiopsied pigmented

skin lesions, prospectively captured in dermatology and plastic surgery clinics, and to compare this with clinical diagnoses made by specialists.

Methods

We conducted a prospective, multicenter, single-arm, masked diagnostic trial. Ethical approval for the study was granted by the Leicester South National Research Ethics Service committee. This study followed the Standards for Reporting of Diagnostic Accuracy (STARD) reporting guideline.

Patients attending dermatology or plastic surgery clinics (on 2-week wait and general referral pathways) with at least 1 skin lesion referred for histological evaluation were eligible for the study. A total of 514 patients provided written informed consent for the study, which was conducted between January 2017 and July 2018. Recruitment was on a consecutive, competitive recruitment basis in 7 participating hospitals, resulting in 551 suspicious lesions scheduled for biopsy and 999 control lesions. Overall, 13 patients were ineligible or withdrawn, primarily because no skin lesion was biopsied (eFigure 1 in the Supplement). One lesion was not biopsied because of specialist decision. Race was defined by the participant.

Lesions not previously biopsied or excised and 2 control lesions (ie, lesions believed to be benign) were selected if they were less than 15 mm in diameter and not located in an anatomical site unsuitable for photographing or in an area of visible scarring. These were photographed by 3 different cameras: iPhone 6s (Apple, Inc), Galaxy S6 (Samsung), and digital single-lens reflex (DSLR) camera D5500 (Nikon). Operating systems were kept up to date during the study. The dermoscopic lens attachments were DermLite DL1 for the iPhone 6s and Galaxy S6 and DermLite Foto II Pro for the DSLR (DermLite). Clinicians assessed the lesions for likelihood of melanoma on an ordinal scale (range, 1-4; 1 indicates unlikely and 4, highly likely).

Histopathology results were collected on biopsied excised lesions and categorized as melanoma, dysplastic nevi, or other. Melanoma in situ was classified as melanoma. When there was histopathological uncertainty in the diagnosis, the most likely diagnosis was reported. Clinicians and patients were not informed of the outcome of the algorithmic assessment, and patients were managed in accordance with the usual standard of care. Images were stored electronically until completion of recruitment.

Prior to the study, the algorithm had been trained with published dermoscopic images. A subset of 289 images from this study were extracted from the data set and used to further train the algorithm (36 [12.5%] randomly selected confirmed melanoma lesions; 67 [23.2%] randomly selected nonmelanoma lesions; and 186 [64.4%] control lesions). No single patient had lesions that appeared in both the training and test split to ensure that overfitting did not occur. Images from each camera were used to train a version of the algorithm, which was then used to assess the remaining images from that camera type. Images were screened for quality, and images that were blurred or otherwise flawed were removed from the data set. The remaining 1550 images were assessed by the algorithm between September 2018 and February 2019.

The diagnostic performance of the algorithm was reassessed in August 2019 to optimize the algorithm to better generalize for different cameras. To ensure that any difference in performance was owing to algorithmic improvements rather than the quantity of data, the image data set was limited to only those lesions for which images came from all 3 cameras. In total, images of 731 lesions were used, including 260 (35.6%) biopsied lesions, of which 51 (7.0%) were diagnosed as melanoma. In addition, healthy skin images were included in the training set, but low-quality images remained excluded. This data set was combined with a historical data set[21] and shuffled. The algorithm was retrained and validated using 10-fold cross-validation. Lesions from each patient were kept together in the same fold to avoid overfitting to a single patient. The model was trained on images taken by the DSLR and Galaxy S6, then tested using images taken by the iPhone 6s as the naive device as well as images taken by the DSLR and Galaxy S6 cameras.

The algorithm generated a numerical response to an image, from 0 to 1, which reflects its confidence that the lesion is melanoma (0 indicates certainly benign and 1, certainly malignant melanoma). A decision threshold defined the point above which a lesion is classed as melanoma.

Statistical Analysis

Receiver operator curves using a nonparametric method with bootstrapped estimation of uncertainty[22] were used to examine the association of the algorithm's confidence scale with the criterion-standard, histopathology-confirmed diagnosis for each camera and for the clinical assessment. The assessment of the clinical accuracy of likelihood of melanoma was estimated on all biopsied lesions (including the lesions whose images were missing and those used for training the algorithm). In the direct comparison, only those lesions with images from every camera were included in the analysis. Area under the receiver operating characteristic curve (AUROC) and other diagnostic accuracy indices (specificity, predictive values, false-positive rates, and false-negative rates) at the decision thresholds that generated a sensitivity of 100% and 95% were compared with the clinical assessment of the likelihood of melanoma. The 100% sensitivity threshold was used because all lesions referred for biopsy were assessed by histopathology. A χ^2 test of the equality of the AUROC between assessments was conducted.

We examined the extent of influence of covariates on the accuracy of each method of melanoma detection. The following patient-related covariates were examined: patients' level of concern, age, sex, race (ie, white vs nonwhite), Fitzpatrick skin type, hair color (ie, blond or red vs not), patient or family history of melanoma and other skin cancer, and lesion-specific body location.

The potential consequences of missing values (from missing or poor-quality images) were assessed using the Little missing completely at random test.[23] We corrected receiver operating characteristic curve estimates using the selection model, as developed by Cook and Rajbhandari.[24] A *P* value of less than .05 was regarded as statistically significant, and all tests were 2-tailed. Statistical estimates of accuracy are reported with 95% CIs. Statistical analysis was conducted using Stata statistical software version 15 (StataCorp).

Results

The study population of 514 patients included 279 women (55.7%) and 484 white patients (96.8%), with a mean (SD) age of 52.1 (18.6) years (**Table 1**). Most participants had Fitzpatrick skin sensitivity categories I to III (417 [81.1%]) and had no family or personal history of any form of skin cancer (352 [71.3%]).

A total of 1550 images of skin lesions were included in the analysis (551 [35.6%] biopsied lesions; 999 [64.4%] control lesions). Of the 551 lesions biopsied, 125 (22.7%) were identified as melanoma by histopathology, 148 (26.8%) were dysplastic nevi, and 278 (50.5%) received other diagnoses (eTable in the Supplement). To understand the types of lesions categorized other, a poststudy audit of 119 biopsied lesions from 1 hospital found that, of 40 lesions categorized as other, 3 (7.5%) were diagnosed as basal cell carcinoma and the rest were benign. The most frequent melanoma subtypes were superficial spreading (67 [53.6%]) and melanoma in situ or lentigo maligna (39 [31.2%]). A total of 32 melanoma cases (25.6%) had a Breslow thickness greater than 1.0 mm. Of the 50 biopsied lesions clinicians identified as highly likely to be melanoma, 36 (72.0%) were histologically confirmed as melanoma and 6 (12.0%) were diagnosed as dysplastic nevi. Of the 124 biopsied lesions clinicians identified as unlikely to be melanoma, 9 (7.3%) were diagnosed as melanoma and 35 (28.2%) were dysplastic nevi (**Table 2**). Table 2 shows the similarity between the algorithmic and clinical assessment at approximately similar levels of sensitivity and specificity. No adverse events were recorded in the study.

The algorithm training data set included 858 images of 286 lesions from 92 patients. A further 627 images were missing owing to technical issues with the camera equipment, data extraction, or an inability to link images to clinical information, and 222 images failed an image quality check (eFigure 1

in the Supplement). Of the remaining lesions used for the analysis, 77 (16.7%) were identified as melanoma.

The analysis of images of biopsied lesions using the algorithm trained on published images of lesions produced AUROCs for each camera as follows: iPhone 6s, 87.9% (82.8%-89.9%); Galaxy S6, 82.3% (76.9%-87.2%); DSLR, 85.0% (77.3%-90.9%).[21] Using the training clinical images from this study, the retrained algorithm produced AUROCs for each camera as follows: iPhone 6s, 90.1% (95% CI, 86.3%-94.0%); Galaxy S6, 85.8% (95% CI, 81.0%-90.7%); and DSLR, 86.9% (95% CI, 80.8%-93.0%) (**Figure 1**). The AUROC for clinician assessment of melanoma likelihood for biopsied lesions was 77.8% (95% CI, 72.5%-81.9%).

Assuming none of the control lesions were melanoma, all methods show an improvement in accuracy when assessing all lesions compared with biopsied lesions alone. The AUROC for clinical assessment was 90.8% (95% CI, 88.0%-93.6%). The AUROC for the algorithmic assessment of

Table 1. Characteristics of 501 Patients

Characteristic	No. (%)
Age, mean (SD), y	52.1 (18.6)
Sex	
Male	222 (44.3)
Female	279 (55.7)
White race	
No	16 (3.20)
Yes	484 (96.8)
Missing	1 (0.1)
No. of nevi	
≤10	173 (35.4)
11-50	232 (47.4)
>50	84 (17.2)
Missing	8 (1.6)
Fitzpatrick skin type	
Type I, highly sensitive	61 (12.4)
Type II, very sun sensitive	172 (34.9)
Type III, sun-sensitive skin	184 (37.3)
Type IV, minimally sun sensitive	62 (12.6)
Type V, sun-insensitive skin	10 (2.0)
Type VI, sun insensitive, never burns	4 (0.8)
Missing	8 (1.6)
Hair color	
Blond	110 (22.8)
Red	39 (8.1)
Brown	298 (61.8)
Black	35 (7.3)
Missing	19 (3.8)
Freckles	
No	214 (43.4)
Yes	279 (56.6)
Missing	8 (1.6)
History of skin cancer	
No history	352 (71.3)
Family nonmelanoma	30 (6.1)
Family melanoma	38 (7.7)
Patient nonmelanoma	34 (6.9)
Patient melanoma	40 (8.1)
Missing	7 (1.4)

images taken with the iPhone 6s was 95.8% (95% CI, 94.1%-97.6%), of images taken with the Galaxy S6, 93.8% (95% CI, 91.4%-96.2%), and of images taken with the DSLR, 91.8% (95% CI, 87.5%-96.1%) (**Figure 2**). A comparison of the 715 lesions that were assessed by all devices and clinicians using a χ^2 test found a statistically significant difference between the AUROCs (χ^2 = 22.6; P < .001), with the iPhone images being most accurate (Figure 1 and Figure 2).

After a third round of training to control for overfitting, the algorithm produced AUROCs for each camera as follows: iPhone 6s, 94.0% (95% CI, 91.3%-96.7%); Galaxy S6, 92.6% (95% CI, 89.6%-95.5%); and DSLR, 92.2% (95% CI, 88.6%-95.7%). These figures show a minor attenuation from the trained AUROC, suggesting that overfitting was small (eFigure 2 in the Supplement). However, the AUROC for all devices showed an improvement of 2%, increasing from 90.1% (95% CI, 85.8%-93.2%) to 92.3% (95% CI, 90.4%-94.1%).

At 100% sensitivity, the algorithm achieved a specificity of 64.8% with iPhone 6s images. When using a decision-threshold that created a sensitivity of 95%, the specificity of biopsied and all lesions for each camera were as follows: iPhone 6s, 50.6% and 78.1%, respectively; Galaxy S6, 46.9% and 75.6%, respectively; and DSLR, 27.6% and 45.5%, respectively. These compare with a specificity of 69.9% on all lesions by clinicians. Similarly, the number of biopsies needed to identify 1 case of melanoma (number needed to biopsy) at a 95% sensitivity was 3.04 for biopsied lesions and 4.00 for all lesions for images taken with the iPhone 6s; 3.22 and 4.39, respectively, for images taken with the Galaxy S6; and 4.23 and 9.02, respectively, for images taken with the DSLR. This compares with a number needed to biopsy by clinicians of 4.92. When the negative predictive value was fixed at 100%, the positive predictive value was 20.3% for clinician assessment; 17.9% for images taken with the iPhone 6s; 13.4% for images taken with the Galaxy S6; and 9.5% for images taken with the DSLR (**Table 3**).

Age was a significant covariate for all methods, and Fitzpatrick skin type was also significant for clinicians. Clinical assessment was more accurate in younger patients (accuracy increased by 0.73%

Table 2. Assessment by Histopathology

			No. (%)			
Assessment Method	Probability of Melanoma	Sensitivity, %/Specificity, %[a]	Other Lesion Type	Dysplasia	Melanoma	Total
Clinician	Unlikely	93/27	80 (64.5)	35 (28.2)	9 (7.3)	124 (100)
	Equivocal	68/83	155 (57.6)	82 (30.5)	32 (11.9)	269 (100)
	Likely	29/97	35 (32.4)	25 (23.2)	48 (44.4)	108 (100)
	Highly likely	0/100	8 (16.0)	6 (12.0)	36 (72.0)	50 (100)
	Total	NA	278 (50.5)	148 (26.8)	125 (22.7)	551 (100)
Algorithm with iPhone 6s image	Unlikely	100/0	19 (76.0)	6 (24.0)	0	25 (100)
	Equivocal	79/87	157 (60.2)	87 (33.3)	17 (6.5)	261 (100)
	Likely	37/98	24 (35.8)	10 (14.9)	33 (49.3)	67 (100)
	Highly likely	0/100	5 (13.9)	2 (5.5)	29 (80.6)	36 (100)
	Total	NA	205 (52.7)	105 (30.0)	79 (20.3)	389 (100)
Algorithm with Galaxy S6 image	Unlikely	92/50	102 (65.4)	48 (30.8)	6 (3.8)	156 (100)
	Equivocal	71/84	72 (60.5)	25 (37.9)	3 (8.6)	202 (100)
	Likely	40/98	25 (37.9)	17 (25.7)	24 (36.4)	66 (100)
	Highly likely	0/100	3 (8.6)	2 (5.7)	30 (85.7)	35 (100)
	Total	NA	202 (53.7)	98 (26.1)	76 (20.2)	376 (100)
Algorithm with DSLR image	Unlikely	86/58	81 (60.5)	46 (34.3)	7 (5.2)	134 (100)
	Equivocal	75/89	44 (59.5)	24 (32.4)	6 (8.1)	74 (100)
	Likely	41/99	12 (30.8)	10 (25.6)	17 (43.6)	39 (100)
	Highly likely	0/100	3 (12.5)	0	21 (87.5)	24 (100)
	Total	NA	140 (51.7)	80 (29.5)	51 (18.8)	271 (100)

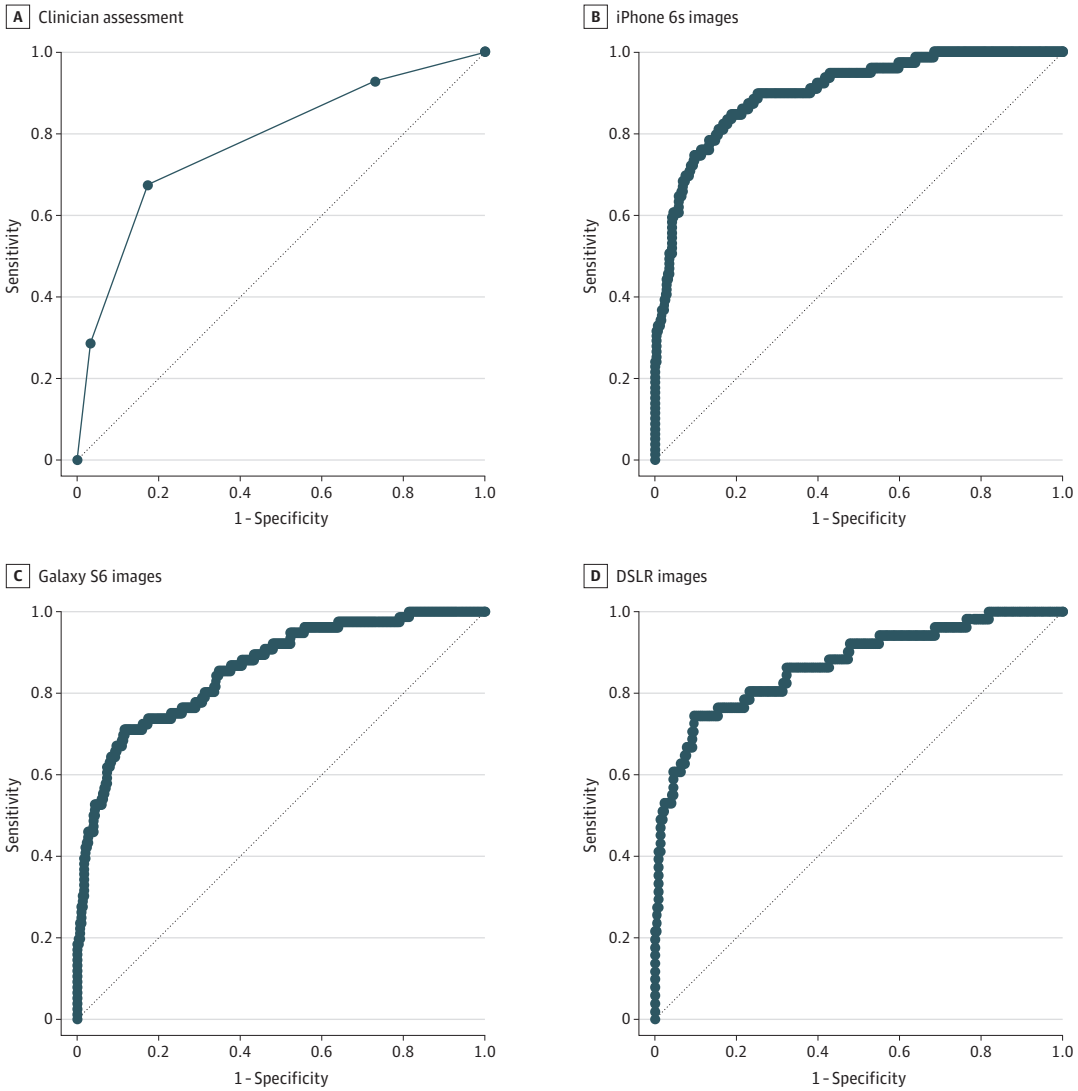
Abbreviations: DSLR, digital single-lens reflex; NA, not applicable.

[a] Sensitivity/specificity shows the accuracy of the level of likelihood for each threshold for detecting melanoma.

for each decreasing year), whereas the algorithm's assessment was more accurate in older patients (for each increasing year, iPhone 6s AUROC increased by 0.76%, Galaxy S6 AUROC increased by 1.9%, and DSLR AUROC increased by 4.3%). Because clinician accuracy is also influenced by Fitzpatrick skin type (ie, AUROC increased by 10% for each level of skin sensitivity above Fitzpatrick type IV), it is possible that clinicians overweighted these risk factors in their determination of the likelihood of melanoma.

Missing images were independent of patient-related or lesion-related characteristics (missing completely at random test results: clinical likelihood: $P = .26$; n = 1581; iPhone 6s image algorithmic assessment: $P = .56$; n = 1110; Galaxy S6 image algorithmic assessment: $P = .17$; n = 1078; and D5500 image algorithmic assessment: $P = .26$; n = 809). However, a degree of bias did appear to be introduced by the missing image–based assessments, as the AUROC based on the Heckman analysis was generally lower for every method of assessment of all lesions than the empirical AUROC. Nevertheless, there is very little difference between the empirical and the Heckman AUROC for all assessment methods.

Figure 1. Receiver Operating Characteristic Curves for Clinical and Trained Algorithm Assessment of Biopsied Lesions



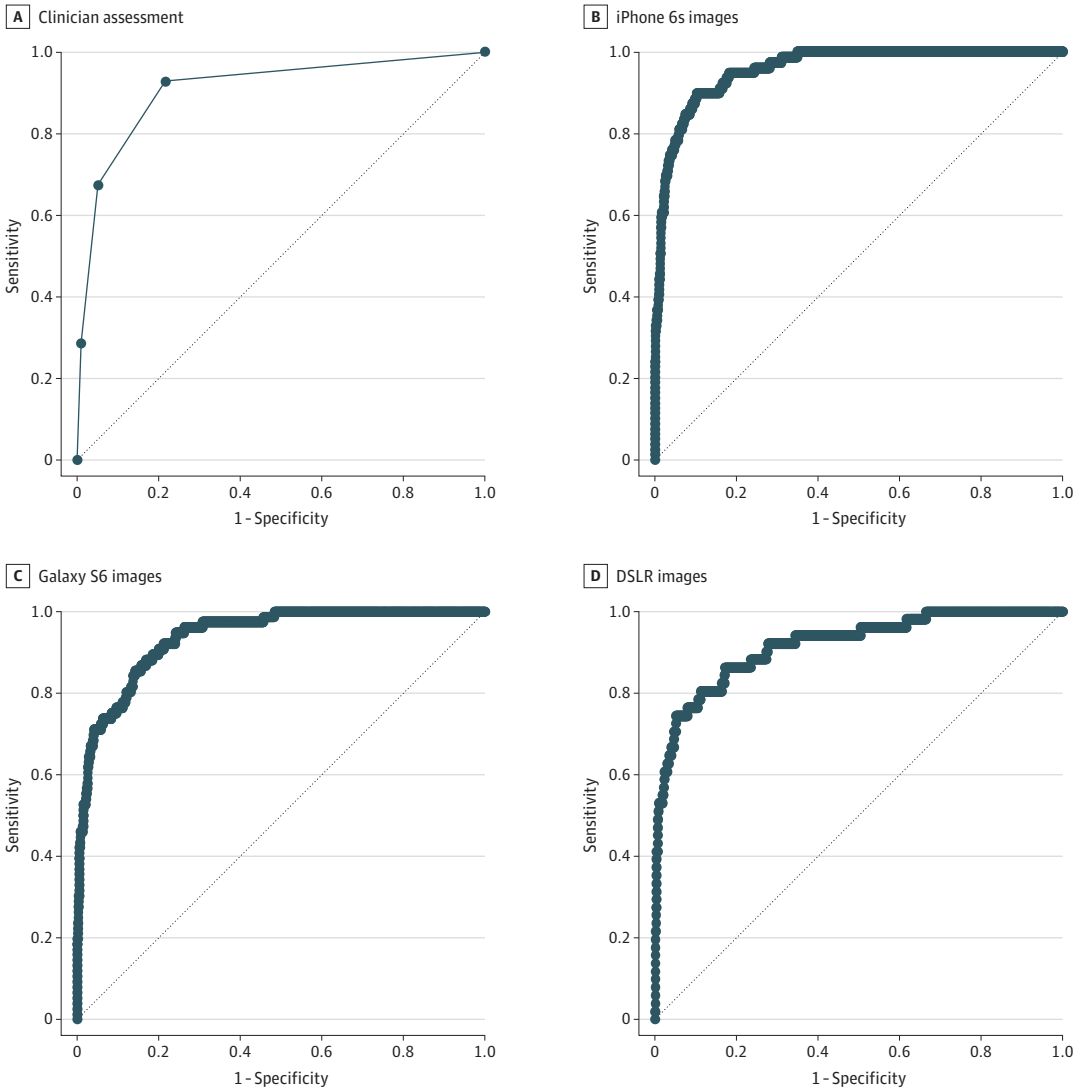
A, Area under receiver operator characteristic curve, 0.779. B, Area under receiver operator characteristic curve, 0.902. C, Area under receiver operator characteristic curve, 0.858. D, Area under receiver operator characteristic curve, 0.869. DSLR indicates digital single-lens reflex.

Discussion

The results of the study showed that the algorithm and specialists identified melanoma in selected suspicious pigmented skin lesions at a similar level of accuracy. More than half of the melanoma diagnoses were either in situ or less than 1 mm deep, indicating that the algorithm could play a role in detecting thin or early-stage lesions. The inclusion of control (ie, believed to be benign) lesions in the analysis enabled us to show that the algorithm maintains a high specificity when control lesions are assessed.

Primary prevention of melanoma based on so-called sun-safe behaviors has not proved to be successful,[25] and a US government review of secondary prevention concluded that there is insufficient evidence to recommend skin cancer screening.[2] While this opinion has been challenged[26] and guidance in other countries recommends screening in high-risk populations,[27-29] the cost of screening in populations at lower risk of developing melanoma means it is not cost-effective to deploy widespread screening programs.[30] To address this issue, Skin Analytics developed an AI

Figure 2. Receiver Operating Characteristic Curves for Clinical and Trained Algorithm Assessment of All Lesions



A, Area under receiver operator characteristic curve, 0.909. B, Area under receiver operator characteristic curve, 0.959. C, Area under receiver operator characteristic curve, 0.938. D, Area under receiver operator characteristic curve, 0.918. DSLR indicates digital single-lens reflex.

algorithm with the intention of it being used as a decision support tool throughout the patient pathway, most notably in the primary care setting.

In terms of reduced numbers of onward referrals and biopsies, the clinical benefit of integrating the algorithm into the patient pathway was not tested in this study. However, the results indicate that the algorithm is sufficiently accurate to detect suspicious lesions in a high-risk, referred population.

Further research to investigate the ability of the algorithm to identity nonmelanoma skin cancer and benign conditions has started. Assessment of the impact of the algorithm on decision-making in primary care, quality of life, and health economics is planned.

Limitations

This study has limitations. First, the analysis of the control lesions with no biopsy was subject to validation bias,[22] and we observed the expected improvement in sensitivity and reduction in specificity (the AUROC was used as the primary assessment because it is less vulnerable to validation bias).[31] Second, caution is needed in extrapolating these results into other settings in which the patient population, frequency of different skin lesion types, risk of skin cancer, and the experience of the physician making the assessment are different. Third, the research was conducted in a secondary-care setting, where there was a high yield of melanoma. The positive and negative predicative values will have been influenced and may not generalize to other clinical settings, such as primary care or screening.

Fourth, AI systems frequently suffer from overfitting,[32] resulting in a lack of generalization. This algorithm was developed and trained using historical dermoscopic images of skin lesions, and it is likely that biases exist in these data sets (eg, patient demographics, melanoma subtypes, or imaging methods). However, it is very difficult to determine whether such biases exist, driving the need for clinical validation. The accuracy of the algorithm observed in this study is similar to, or better than, previously reported accuracy on historical images,[21] providing confidence that it minimizes overfitting and generalizes to out-of-distribution data sets. The accuracy of the algorithm was initially highest for images taken with the iPhone 6s and lowest for images taken with the DSLR, which could reflect poor generalization across different cameras. However, the retraining of the algorithm using images from the Galaxy S6 and DSLR reduced this and demonstrated that overfitting was minimal.

Table 3. Comparison of Algorithm With Clinicians

Method	No. of Images	% Sensitivity	Specificity	FPR	FNR	PPV	NPV	NNB
All Lesions								
Clinician	1582	100	69.90	30.13	0.00	20.30	100.00	4.92
Algorithm with iPhone 6s image	1110	100	64.80	35.18	0.00	17.90	100.00	5.58
		94.90	78.10	21.87	5.06	25.00	99.50	4.00
Algorithm with Galaxy S6 image	1078	100	51.20	48.83	0.00	13.40	100.00	7.44
		94.90	75.60	24.42	5.13	22.80	99.50	4.39
Algorithm with DSLR image	809	100	30.00	70.03	0.00	9.48	100.00	10.55
		92.80	45.50	54.52	7.25	11.10	98.80	9.02
Biopsied Lesions								
Clinician	553	100	NA	NA	NA	NA	NA	NA
Algorithm with iPhone 6s image	388	100	31.30	68.71	0.00	27.10	100	NA
		95.00	50.60	49.35	5.06	32.90	98.00	NA
Algorithm with Galaxy S6 image	376	100	18.40	81.55	0.00	23.60	100	NA
		95.00	46.90	53.07	5.13	31.10	97.00	NA
Algorithm with DSLR image	271	100	3.15	96.85	0.00	19.90	100	NA
		93.00	27.60	72.38	7.25	23.60	94.00	NA

Abbreviations: DSLR, digital single-lens reflex; FNR, false-negative rate; FPR, false-positive rate; NA, not applicable; NNB, number needed to biopsy (to identify 1 case of melanoma); NPV, negative predictive value; PPV, positive predictive value.

Fifth, there were technical issues with the equipment used for the study, leading to a high proportion of lesions that could not be photographed by at least 1 of the 3 cameras or images that could not be analyzed by the algorithm. These issues show the influence that the usability of equipment can have on the effectiveness of a technology, which needs to be reliable for health care professionals to use it and have confidence in it. Reasons for the missing images included technology failures, such as camera malfunction, and poor optimization of the equipment for use in a clinical setting (particularly the DSLR, which was awkward for the study staff to use). Despite this, statistical analysis of the influence of missing data suggests that there was little selection bias generated by the missing data. This provides evidence for the robustness of the algorithm in accurately detecting melanoma.

Sixth, there were only 3 types of cameras used in this trial. It remains to be tested whether the results apply to other smartphones or other cameras.

Conclusions

The findings of this diagnostic trial demonstrated that an AI algorithm, using different camera types, can detect melanoma with a similar level of accuracy as specialists. The development of low-cost screening methods, such as AI-based services, could transform patient diagnosis pathways, enabling greater efficiencies throughout the health care service.

ARTICLE INFORMATION

Accepted for Publication: August 27, 2019.

Published: October 16, 2019. doi:10.1001/jamanetworkopen.2019.13436

Corresponding Author: Michael Phillips, MMedSci, Centre for Medical Research, University of Western Australia, Perth, WA 6008, Australia (michael.phillips@perkins.uwa.edu.au); Helen Marsden, PhD, Skin Analytics Limited, One Phipp Street, The Frames, London EC2A 4PS, United Kingdom (helen@skinanalytics.co.uk).

Author Affiliations: Harry Perkins Institute of Medical Research, Perth, Western Australia, Australia (Phillips); Centre for Medical Research, University of Western Australia, Perth, Western Australia, Australia (Phillips); Skin Analytics Limited, London, United Kingdom (Marsden, Greenhalgh); Royal Stoke University Hospital, University Hospital North Midlands, Stoke, United Kingdom (Jaffe); Oxford University Hospitals NHS Foundation Trust, Oxford, United Kingdom (Matin, Wali); Royal Devon and Exeter NHS Foundation Trust, Exeter, United Kingdom (McGrath, James); Dudley Group NHS Foundation Trust, Corbett Hospital, Stourbridge, United Kingdom (Ladoyanni); Barts Health, London, United Kingdom (Bewley); Queen Mary School of Medicine, University of London, London, United Kingdom (Bewley); Dermatology Unit, University of Campania, Naples, Italy (Argenziano); Barnet and Chase Farm Hospitals, Royal Free NHS Foundation Trust, London, United Kingdom (Palamaras).

Author Contributions: Mr Phillips had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: Phillips, Marsden, Greenhalgh, Argenziano, Palamaras.

Acquisition, analysis, or interpretation of data: Phillips, Marsden, Jaffe, Matin, Wali, Greenhalgh, McGrath, James, Ladoyanni, Bewley, Palamaras.

Drafting of the manuscript: Phillips, Marsden, Greenhalgh, McGrath, Palamaras.

Critical revision of the manuscript for important intellectual content: Jaffe, Matin, Wali, Greenhalgh, McGrath, James, Ladoyanni, Bewley, Argenziano, Palamaras.

Statistical analysis: Phillips, Greenhalgh.

Administrative, technical, or material support: Marsden, Jaffe, McGrath, James, Bewley, Palamaras.

Supervision: Phillips, Jaffe, Wali, Ladoyanni, Palamaras.

Conflict of Interest Disclosures: Mr Phillips reported having a familial relationship with Skin Analytics Limited. Dr Marsden reported working for Skin Analytics Limited and receiving share options during the conduct of the study.

Dr Matin reported receiving grants from Barco outside the submitted work and being coauthor of a suite of Cochrane diagnostic test accuracy systematic reviews, including the diagnosis of melanoma. Dr Greenhalgh reported working for Skin Analytics Limited during the conduct of the study and holding patent US 20150254851 A1. Dr Bewley reported working as an ad hoc consultant for Almirall, AbbVie, Galderma, LEO Pharma, Eli Lilly and Co, Sanofi, Novartis, and Janssen Pharmaceuticals. No other disclosures were reported.

Funding/Support: This study was funded by Skin Analytics Limited, which developed and owns Deep Ensemble for Recognition of Malignancy. The Royal Perth Hospital Medical Research Fund supported the analysis and interpretation of the data and the preparation of the manuscript.

Role of the Funder/Sponsor: Employees of Skin Analytics Limited designed and conducted the study; collected, managed, and analyzed the images; contributed to the interpretation of the data; and prepared, reviewed, and approved the decision to submit the manuscript for publication. Royal Perth Hospital Medical Research Fund had no involvement with the data analysis and interpretation of the data, nor the preparation of the manuscript.

Additional Contributions: We thank all the patients who kindly consented to participate in the study.

REFERENCES

1. Cancer Research UK. Melanoma skin cancer survival statistics. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/melanoma-skin-cancer/survival#heading-Three. Accessed March 21, 2019.

2. Wernli KJ, Henrikson NB, Morrison CC, Nguyen M, Pocobelli G, Whitlock EP; US Preventive Services Task Force Evidence Syntheses formerly Systematic Evidence Reviews. Screening for skin cancer in adults: an updated systematic evidence review for the US Preventive Services Task Force. *JAMA*. 2016;316(4):436-447.

3. National Institute for Health and Care Excellence. Melanoma: assessment and management. https://www.nice.org.uk/guidance/ng14. Accessed September 5, 2019.

4. Welch HG, Woloshin S, Schwartz LM. Skin biopsy rates and incidence of melanoma: population based ecological study. *BMJ*. 2005;331(7515):481. doi:10.1136/bmj.38516.649537.E0

5. Chen SC, Pennie ML, Kolm P, et al. Diagnosing and managing cutaneous pigmented lesions: primary care physicians versus dermatologists. *J Gen Intern Med*. 2006;21(7):678-682. doi:10.1111/j.1525-1497.2006.00462.x

6. Bainbridge S, Cake R, Meredith M, Furness P, Gordon B. Testing times to come: an evaluation of pathology capacity across the UK. https://www.cancerresearchuk.org/sites/default/files/testing_times_to_come_nov_16_cruk.pdf. Accessed September 5, 2019.

7. Dinnes J, Deeks JJ, Chuchu N, et al; Cochrane Skin Cancer Diagnostic Test Accuracy Group. Dermoscopy, with and without visual inspection, for diagnosing melanoma in adults. *Cochrane Database Syst Rev*. 2018;12(12):CD011902. doi:10.1002/14651858.cd011902.pub2

8. Dinnes J, Deeks JJ, Saleh D, et al; Cochrane Skin Cancer Diagnostic Test Accuracy Group. Reflectance confocal microscopy for diagnosing cutaneous melanoma in adults. *Cochrane Database Syst Rev*. 2018;12(12):CD013190. doi:10.1002/14651858.cd013190

9. Chuchu N, Dinnes J, Takwoingi Y, et al; Cochrane Skin Cancer Diagnostic Test Accuracy Group. Teledermatology for diagnosing skin cancer in adults. *Cochrane Database Syst Rev*. 2018;12(12):CD013193. doi:10.1002/14651858.cd013193

10. Ferrante di Ruffano L, Takwoingi Y, Dinnes J, et al; Cochrane Skin Cancer Diagnostic Test Accuracy Group. Computer-assisted diagnosis techniques (dermoscopy and spectroscopy-based) for diagnosing skin cancer in adults. *Cochrane Database Syst Rev*. 2018;12(12):CD013186. doi:10.1002/14651858.cd013186

11. Chuchu N, Takwoingi Y, Dinnes J, et al; Cochrane Skin Cancer Diagnostic Test Accuracy Group. Smartphone applications for triaging adults with skin lesions that are suspicious for melanoma. *Cochrane Database Syst Rev*. 2018;12(12):CD013192. doi:10.1002/14651858.cd013192

12. Kittler H, Pehamberger H, Wolff K, Binder M. Diagnostic accuracy of dermoscopy. *Lancet Oncol*. 2002;3(3):159-165. doi:10.1016/S1470-2045(02)00679-4

13. Brewer AC, Endly DC, Henley J, et al. Mobile applications in dermatology. *JAMA Dermatol*. 2013;149(11):1300-1304. doi:10.1001/jamadermatol.2013.5517

14. Kassianos AP, Emery JD, Murchie P, Walter FM. Smartphone applications for melanoma detection by community, patient and generalist clinician users: a review. *Br J Dermatol*. 2015;172(6):1507-1518. doi:10.1111/bjd.13665

15. Ferrero NA, Morrell DS, Burkhart CN. Skin scan: a demonstration of the need for FDA regulation of medical apps on iPhone. *J Am Acad Dermatol*. 2013;68(3):515-516. doi:10.1016/j.jaad.2012.10.045

16. Wolf JA, Moreau JF, Akilov O, et al. Diagnostic inaccuracy of smartphone applications for melanoma detection. *JAMA Dermatol*. 2013;149(4):422-426. doi:10.1001/jamadermatol.2013.2382

17. Stoecker WV, Rader RK, Halpern A. Diagnostic inaccuracy of smartphone applications for melanoma detection: representative lesion sets and the role for adjunctive technologies. *JAMA Dermatol*. 2013;149(7):884. doi:10.1001/jamadermatol.2013.4334

18. Wolf JA, Ferris LK. Diagnostic inaccuracy of smartphone applications for melanoma detection: reply. *JAMA Dermatol*. 2013;149(7):885. doi:10.1001/jamadermatol.2013.4337

19. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118. doi:10.1038/nature21056

20. Haenssle HA, Fink C, Schneiderbauer R, et al; Reader study level-I and level-II Groups. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol*. 2018;29(8):1836-1842. doi:10.1093/annonc/mdy166

21. Phillips M, Greenhalgh J, Marsden H, Palamaras I. Detection of malignant melanoma using artificial intelligence: an observational study. *Dermatol Pract Concept*. In press.

22. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford, UK: Oxford University Press; 2003.

23. Little RJA. A test of missing completely at random for multivariate data with missing values. *J Am Stat Assoc*. 1988;83(404):1198-1202. doi:10.1080/01621459.1988.10478722

24. Cook JA, Rajbhandari A. Heckroccurve: ROC curves for selected samples. *Stata J*. 2018;18(1):174-183. doi:10.1177/1536867X1801800110

25. Greinert R, Breitbart E, Mohr P, Volkmer B. Health initiatives for the prevention of skin cancer. In: Reichrath J, ed. *Sunlight, Vitamin D and Skin Cancer*. New York, NY: Springer; 2008:485-499.

26. Johnson MM, Leachman SA, Aspinwall LG, et al. Skin cancer screening: recommendations for data-driven screening guidelines and a review of the US Preventive Services Task Force controversy. *Melanoma Manag*. 2017;4(1):13-37. doi:10.2217/mmt-2016-0022

27. Cancer Council Australia. Position statement: screening and early detection of skin cancer. https://dermcoll.edu.au/wp-content/uploads/2014/05/PosStatEarlyDetectSkinCa.pdf. Accessed September 5, 2019.

28. Royal Australian College of General Practitioners. Guidelines for preventive activities in general practice, 9th edition. https://www.racgp.org.au/download/Documents/Guidelines/Redbook9/17048-Red-Book-9th-Edition.pdf. Accessed September 5, 2019.

29. Marsden JR, Newton-Bishop JA, Burrows L, et al; British Association of Dermatologists (BAD) Clinical Standards Unit. Revised UK guidelines for the management of cutaneous melanoma 2010. *J Plast Reconstr Aesthet Surg*. 2010;63(9):1401-1419. doi:10.1016/j.bjps.2010.07.006

30. Robinson JK, Halpern AC. Cost-effective melanoma screening. *JAMA Dermatol*. 2016;152(1):19-21. doi:10.1001/jamadermatol.2015.2681

31. Diamond GA. The wizard of odds: Bayes theorem and diagnostic testing. *Mayo Clin Proc*. 1999;74(11):1179-1182. doi:10.4065/74.11.1179

32. Obermeyer Z, Emanuel EJ. Predicting the future: big data, machine learning, and clinical medicine. *N Engl J Med*. 2016;375(13):1216-1219. doi:10.1056/NEJMp1606181

SUPPLEMENT.

eTable. Biopsied Lesion Characteristics

eFigure 1. STARD Flow Chart

eFigure 2. ROC Curves of Algorithm Performance Following Retraining