Confidential: Do not distribute. Pre embargo material.

STUDY

ONLINE FIRST

Diagnostic Inaccuracy of Smartphone Applications for Melanoma Detection

Joel A. Wolf, BA; Jacqueline Moreau, BA; Oleg Akilov, MD; Timothy Patton, DO; Joseph C. English III, MD; Jonhan Ho, MD; Laura K. Ferris, MD, PhD

Objective: To measure the performance of smartphone applications that evaluate photographs of skin lesions and provide the user with feedback about the likelihood of malignancy.

Design: Case-control diagnostic accuracy study.

Setting: Academic dermatology department.

Participants and Materials: Digital clinical images of pigmented cutaneous lesions (60 melanoma and 128 benign control lesions) with a histologic diagnosis rendered by a board-certified dermatopathologist, obtained before biopsy from patients undergoing lesion removal as a part of routine care.

Main Outcome Measures: Sensitivity, specificity, and positive and negative predictive values of 4 smartphone applications designed to aid nonclinician users in determining whether their skin lesion is benign or malignant.

Results: Sensitivity of the 4 tested applications ranged from 6.8% to 98.1%; specificity, 30.4% to 93.7%; positive predictive value, 33.3% to 42.1%; and negative predictive value, 65.4% to 97.0%. The highest sensitivity for melanoma diagnosis was observed for an application that sends the image directly to a board-certified dermatologist for analysis; the lowest, for applications that use automated algorithms to analyze images.

Conclusions: The performance of smartphone applications in assessing melanoma risk is highly variable, and 3 of 4 smartphone applications incorrectly classified 30% or more of melanomas as unconcerning. Reliance on these applications, which are not subject to regulatory oversight, in lieu of medical consultation can delay the diagnosis of melanoma and harm users.

JAMA Dermatol.
Published online January 16, 2013.
doi:10.1001/jamadermatol.2013.2382

As smartphones use increases, these devices are applied to functions beyond communication and entertainment and often become tools that are involved intimately in many aspects of daily life through the use of specialized applications. Several applications in the field of health care, marketed directly to the public, are readily available. Some examples include applications that are intended to aid users in learning about adverse effects of medications, to track their caloric intake and expenditure to manage weight loss, and to log their menstrual cycles to monitor fertility. Although such applications have the potential to improve patient awareness and physician-patient communication, applications that provide any type of medical advice might result in harm to the patient if that advice is incorrect or misleading.

A review of the applications available for the 2 most popular smartphone platforms reveals several that are marketed to nonclinician users to assist them in deciding whether a skin lesion is potentially a melanoma or otherwise of concern and requires medical attention or whether it is likely benign based on analysis of a digital clinical image. Such applications are available for free or for a relatively low cost compared with an in-person medical consultation. These applications are not subject to any sort of validation or regulatory oversight. Despite disclaimers that these applications are intended for educational purposes, they have the potential to harm users who may believe mistakenly that the evaluation given by such an application is a substitute for medical advice. This risk is of particular concern for economically disadvantaged and uninsured patients. Because a substantial percentage of melanomas are detected initially by patients,[1-4] the potential effect of such applications on melanoma detection patterns is particularly relevant. We therefore sought to determine the accuracy of these applications in determining the benign vs malig-

Author Affiliations: Department of Dermatology, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania (Drs Akilov, Patton, English, Ho, and Ferris). Mr Wolf and Ms Moreau are medical students at the University of Pittsburgh, Pittsburgh, Pennsylvania.

©2013 American Medical Association. All rights reserved.

Confidential: Do not distribute. Pre embargo material.

nant nature of a series of images of pigmented skin lesions using the histologic finding as the reference standard.

METHODS

SKIN LESION IMAGES

The University of Pittsburgh institutional review board reviewed this study and determined that it was exempt from full board review provided that all images used did not contain identifiable patient features or data and images were already in existence at the start of the study. The images of skin lesions were selected from our database of images that are captured routinely before skin lesion removal to allow clinicopathologic correlation in making medical management decisions. We only used close-up images of lesions. Images that contained any identifiable features, such as facial features, tattoos, or labels with patient information, were excluded or cropped to remove the identifiable features or information. Because histologic diagnosis was used as the reference standard for subsequent analysis, we only used images for which a clear histologic diagnosis was rendered by a board-certified dermatopathologist (J.H.). Lesions with equivocal diagnoses, such as "melanoma cannot be ruled out" or "atypical melanocytic proliferation," were excluded, as were Spitz nevi, pigmented spindle cell nevus of Reed, and other uncommon or equivocal lesions. We also excluded lesions with moderate or high-grade atypia given the controversy over their management. The remaining images were stratified into 1 of the following categories: invasive melanoma, melanoma in situ, lentigo, benign nevus (including compound, junctional, and low-grade dysplastic nevi), dermatofibroma, seborrheic keratosis, and hemangioma. Because 1 application used assessments by a remote dermatologist, we cropped images to remove rulers or stickers that might reveal that our images were from a dermatologist and not a patient. This process was performed using a computer program (iPhoto; Apple Inc) and did not compromise the integrity of the images. Two investigators (J.W. and L.K.F.) then reviewed all images for image quality and omitted those that were of poor quality or resolution.

SMARTPHONE APPLICATIONS

We searched the application stores of the 2 most popular smartphone operating systems for applications that claim or suggest an ability to assist users in determining whether a skin lesion may be malignant. Our search terms included *skin, skin cancer, melanoma,* and *mole.* We reviewed the descriptions of all applications returned by these searches to determine whether they use a photograph of a skin lesion to make assessments and whether they suggest any type of diagnosis or estimate the risk of malignancy. These applications then were evaluated to determine whether they could be used with an existing image (ie, if an image could be uploaded into the application rather than requiring that the image be captured in real time within the application). Three applications were excluded because they could not use existing photographs. Applications that allowed the use of existing images were selected for further evaluation. Our search yielded a total of 4 applications that met our criteria. Because the purpose of our study was to determine the accuracy of such applications in general and not to make a direct statement about a particular application, we have chosen not to identify the applications by their commercial name but rather to number them.

Application 1 uses an automated algorithm to detect the border of the lesion, although it also allows manual input to confirm or to change the detected border. Of the applications we tested, only application 1 has this feature of user input for border detection. The application then analyzes the image and gives an assessment of "problematic," which we considered to be a positive test result; "okay," which we considered to be a negative test result; or "error" if the image could not be assessed by the application. We categorized the latter group as unevaluable.

Application 2 uses an automated algorithm to evaluate an image that has been uploaded by the user. The output given is "melanoma," which we considered to be a positive test result, or "looks good," which we considered to be a negative test result. If the image could not be analyzed, a message of "skin condition not found" was given and we considered the image to be unevaluable.

Application 3 asks the user to upload an image to the application and then to position it within a box to ensure that the correct lesion is analyzed. The output given by the application is "high risk," which we considered to be a positive test result, or "medium risk" or "low risk," both of which we considered to be a negative test result. The presence of a medium-risk category in application 3 presented some difficulty in analysis because only this application among those tested gave an intermediate output. Thus, we performed sensitivity and specificity analyses with medium-risk lesions counting as a positive test result because we do not know how a user would interpret such a result. Some lesions generated a message of "error," and these were considered unevaluable.

Application 4 can be run on a smartphone or from a website. This program differs from the others because it does not use an automated analysis algorithm to evaluate images; rather, each image is sent to a board-certified dermatologist for evaluation, and that assessment is returned to the user within 24 hours. The identity of the dermatologist is not given, and we do not know whether all the images were read by the same dermatologist or by several different dermatologists. The output given is "atypical," which we considered to be a positive test result, or "typical," which we considered to be a negative test result. For some images we submitted, we were given a response of "send another photograph" or "unable to categorize," and we considered these images to be unevaluable in our analysis.

DETERMINATION OF APPLICATION ACCURACY AND STATISTICAL ANALYSES

Each of the 4 applications was presented with each eligible pigmented skin lesion image, and we attempted evaluation. We recorded output as a test result of positive, negative, or unevaluable as described in the preceding section. We calculated the percentage of images presented to each application that were considered to be evaluable. Subsequent analysis of the overall sensitivity, specificity, positive predicative value (PPV), and negative predictive value (NPV) for each application was performed with 95% confidence intervals. These calculations were performed only for evaluable lesions because we did not have the option of submitting another image, and we did not want this limitation to bias our results. To compare application performances with each other, the relative sensitivities of each application were compared using the McNemar test with Holm-Bonferroni adjustment for multiple comparisons. To perform this calculation, only lesions that were considered evaluable by both applications being compared were included. We performed statistical analysis using commercially available software (Stata, version 12.1; StataCorp).

RESULTS

IMAGES SELECTED FOR EVALUATION

We reviewed a total of 390 images for possible inclusion in this study. We excluded 202 as being of poor image qual-

©2013 American Medical Association. All rights reserved.

Confidential: Do not distribute. Pre embargo material.

Table 1. Histologic Diagnosis of 188 Evaluated Lesions

Lesion	No. (%)
Melanoma	60 (31.9)
Invasive	44 (23.4)
In situ	16 (8.5)
Benign lesions	128 (68.1)
Lentigo	8 (4.3)
Benign nevus	94 (50.0)
Seborrheic keratosis	20 (10.6)
Hemangioma	2 (1.1)
Dermatofibroma	4 (2.1)

Table 2. Sensitivity and Specificity of Applications Using Evaluable Images

Application No.	Evaluable Image, No. (%)	Sensitivity, % (95% CI)	Specificity, % (95% CI)
1	182 (96.8)	70.0 (56.6-80.8)	39.3 (30.7-48.6)
2	185 (98.4)	69.0 (55.3-80.1)	37.0 (28.7-46.1)
3	170 (90.4)	6.8 (2.2-17.3)	93.7 (87.0-97.2)
4	159 (84.6)	98.1 (88.8-99.9)	30.4 (22.1-40.3)

ity, containing identifiable patient information or features, or lacking sufficient clinical or histologic information. A total of 188 lesions were evaluated using the 4 applications. Of these lesions, 60 were melanoma (44 invasive and 16 in situ). The remaining 128 lesions were benign. The categorization of all lesions is given in **Table 1**.

APPLICATION SENSITIVITY, SPECIFICITY, PPV, AND NPV

Each application was presented with each of the 188 lesions in the study, and the test result was recorded as positive, negative, or unevaluable as outlined in the "Smartphone Applications" subsection of the "Methods" section. The primary end point of our study was the sensitivity to melanoma categorization because most of the lesions removed in our practice are removed owing to concern about malignancy, and thus we expected the specificity to be low.

As reported in **Table 2**, the applications considered 84.6% to 98.4% of the images evaluable. Using only those images considered evaluable for each application, we calculated the overall sensitivity and specificity with 95% confidence intervals for each application (Table 2). Sensitivities ranged from 6.8% to 98.1%. Application 3 had the lowest sensitivity when a readout of medium risk was considered to be a negative test result. When analysis was performed considering the medium-risk readout to be a positive test result, the calculated sensitivity was 54.2% (95% CI, 40.8%-67.1%). Application specificities ranged from 30.4% to 93.7%. When the medium-risk result was considered to be a positive test result, the specificity of application 3 dropped to 61.3% (95% CI, 51.5%-70.2%). When we compared the 4 applications with each other, application 4 had higher sensitivity than the other 3 ($P < .001$ vs applications 1 and 3; $P = .02$ vs application 2).

We also calculated the PPV, NPV, and 95% confidence interval for each application. The results are shown in **Table 3**. The PPVs ranged from 33.3% to 42.1%; the NPVs, from 65.4% to 97.0%.

Table 3. Predictive Values of Applications Using Evaluable Images

Application No.	PPV, % (95% CI)	NPV, % (95% CI)
1	36.2 (27.6-45.7)	72.7 (60.2-82.6)
2	33.3 (25.2-42.6)	72.3 (59.6-82.3)
3	36.4 (12.4-68.4)	65.4 (57.4-72.7)
4	42.1 (33.4-51.2)	97.0 (82.5-99.8)

Abbreviations: NPV, negative predictive value; PPV, positive predictive value.

COMMENT

More than 13 000 health care applications marketed to consumers are available in the largest online application store alone, and the mobile health application industry generated an estimated $718 million worldwide in 2011 according to a recent report.[5] Two-thirds of physicians use smartphone applications in their practice.[6] Some of these applications have been evaluated in the peer-reviewed literature, including instruments used to aid autobiographical memory in patients with Alzheimer disease,[7] to assist in the delivery of cardiac life support,[8] and to manage diabetes mellitus.[9] However, this type of evaluation is not common for applications marketed directly to consumers.

In dermatology, several applications are available that offer educational information about melanoma and skin self-examination and that aid the user in tracking the evolution of individual skin lesions. However, the applications we evaluated in our study go beyond aiding patients in cataloging and tracking lesions and additionally give an assessment of risk or probability that a lesion is benign or malignant. This finding is of particular concern because patients may substitute these readouts for standard medical consultation. Three of the 4 applications we evaluated do not involve a physician at any point in the evaluation. Even the best-performing among these 3 applications classified 18 of 60 melanomas (30%) in our study as benign.

The explosion of smartphone applications geared at health-related decision making has not gone unnoticed by the US Food and Drug Administration (FDA). In July of 2011, the FDA announced plans to regulate smartphone applications that pair with medical devices already regulated by the FDA, such as cardiac monitors and radiologic imaging devices.[10] In June 2012, Congress approved the FDA Safety and Innovation Act,[11] which allows the FDA to regulate some medical applications on smartphones. However, how this process will occur, which applications will be subject to this regulation, and which applications will be exempt remain unclear. Although clarification of these guidelines were projected before the end of 2012, at the time of publication their impact remains uncertain. In 2011, the Federal Trade Commission fined the developers of 2 applications that made unsubstantiated claims to treat acne using colored light that could

©2013 American Medical Association. All rights reserved.

Confidential: Do not distribute. Pfizer Clinical Trial Material.

be shone on the skin from a smartphone application. Both applications were withdrawn from the market.[12]

In our study, the application with the highest sensitivity essentially functions as a tool for store-and-forward teledermatology. Using this application, only 1 of the 53 melanomas evaluated was rated as typical (ie, benign). Although our results show that the physician-based method is superior in sensitivity to the applications that use an automated algorithm for analysis, this application was also the most expensive in terms of cost per use at $5 for each lesion evaluated. By contrast, the costs of the other applications range from free to $4.99 for evaluation of an unlimited number of lesions. In addition, although applications 1, 2, and 3 provided immediate feedback on lesions (mean duration, <1 minute), the evaluation given by application 4 was received in about 24 hours.

Our study has some intrinsic limitations. To power this pilot study adequately while restricting our inclusion criteria to lesions for which histopathologic evaluation as the reference standard for diagnosis was available, we were limited to the use of existing photographs of lesions that had been removed before the start of the study. This limitation has several implications. First, our images consisted primarily of lesions that were considered to be atypical in clinical appearance by at least 1 dermatologist. For this reason, and because of the potentially devastating consequences of missing a melanoma (compared with classifying a benign lesion as of concern), we made sensitivity our primary end point. In addition, we could not evaluate the performance of applications that require images to be captured in real time within the application because we limited our study to existing images. However, because we are not comparing applications for the purpose of recommending one over the other, our results still provide valuable information about the general threat that such applications may pose. Finally, because the lesions in our images were no longer present on the patient, we could not retake a photograph if a lesion was considered unevaluable. To compensate for this limitation, we included only evaluable lesions in our analyses.

Technologies that improve the rate of melanoma self-detection have potential to improve mortality due to melanoma and would be welcome additions to our efforts to decrease mortality through early detection. However, extreme care must be taken to avoid harming patients in the process. Despite disclaimers presented by each of these applications that they were designed for educational purposes rather than actual diagnosis and that they should not substitute for standard medical care, releasing a tool to the public requires some thought as to how it could be misused. This potential is of particular concern in times of economic hardship, when uninsured and even insured patients, deterred by the cost of copayments for medical visits, may turn to these applications as alternatives to physician evaluation. Physicians must be aware of these applications because the use of medical applications seems to be increasing over time; whether such applications may be subject to regulatory oversight, whether oversight is appropriate, and when oversight might be applied remain unclear. However, given the recent media and legislative interest in such applications, the dermatologist should be aware of those relevant to our field to aid us in protecting and educating our patients.

Accepted for Publication: October 29, 2012.
Published Online: January 16, 2013. doi:10.1001/jamadermatol.2013.2382
Correspondence: Laura K. Ferris, MD, PhD, Department of Dermatology, University of Pittsburgh Medical Center, 3601 Fifth Ave, Fifth Floor, Pittsburgh, PA 15213 (ferrislk@upmc.edu).
Author Contributions: Dr Ferris 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. *Study concept and design:* Wolf and Ferris. *Acquisition of data:* Wolf, Akilov, Patton, English, Ho, and Ferris. *Analysis and interpretation of data:* Wolf, Moreau, and Ferris. *Drafting of the manuscript:* Wolf, Moreau, and Ferris. *Critical revision of the manuscript for important intellectual content:* Wolf, Moreau, Akilov, Patton, English, Ho, and Ferris. *Statistical analysis:* Moreau. *Obtained funding:* Akilov and Ferris. *Administrative, technical, and material support:* Wolf, Patton, Ho, and Ferris. *Study supervision:* English and Ferris.
Conflict of Interest Disclosures: Dr Ferris reported having served as an investigator and consultant for MELA Sciences, Inc.
Funding/Support: This study was supported by grants UL1RR024153 and UL1TR000005 from the National Institutes of Health.

REFERENCES

1. Brady MS, Oliveria SA, Christos PJ, et al. Patterns of detection in patients with cutaneous melanoma. *Cancer*. 2000;89(2):342-347.
2. Epstein DS, Lange JR, Gruber SB, Mofid M, Koch SE. Is physician detection associated with thinner melanomas? *JAMA*. 1999;281(7):640-643.
3. Kantor J, Kantor DE. Routine dermatologist-performed full-body skin examination and early melanoma detection. *Arch Dermatol*. 2009;145(8):873-876.
4. McGuire ST, Secrest AM, Andrulonis R, Ferris LK. Surveillance of patients for early detection of melanoma: patterns in dermatologist vs patient discovery. *Arch Dermatol*. 2011;147(6):673-678.
5. Morris DP. Health-care apps for smartphones pit FDA against tech industry. *Washington Post*. June 22, 2012. http://www.washingtonpost.com/business/economy/health-care-apps-for-smartphones-pit-fda-against-tech-industry/2012/06/22/gJQAHCcBvV_story_2.html. Accessed August 25, 2012.
6. Senior K. Smart phones: new clinical tools in oncology? *Lancet Oncol*. 2011;12(5):429-430.
7. De Leo G, Brivio E, Sautter SW. Supporting autobiographical memory in patients with Alzheimer's disease using smart phones. *Appl Neuropsychol*. 2011;18(1):69-76.
8. Low D, Clark N, Soar J, et al. A randomised control trial to determine if use of the iResus© application on a smart phone improves the performance of an advanced life support provider in a simulated medical emergency. *Anaesthesia*. 2011;66(4):255-262. doi:10.1111/j.1365-2044.2011.06649.x.
9. Ciemins E, Coon P, Sorli C. An analysis of data management tools for diabetes self-management: can smart phone technology keep up? *J Diabetes Sci Technol*. 2010;4(4):958-960.
10. US Food and Drug Administration. FDA outlines oversight of mobile medical applications. FDA news release. July 19, 2011. http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm263340.htm. Accessed August 25, 2012.
11. US Food and Drug Administration. FDA Safety and Innovation Act. Updated July 31, 2012. http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/Overview/ucm310927.htm. Accessed August 25, 2012.
12. Federal Trade Commission. "Acne cure" mobile app marketers will drop baseless claims under FTC settlements. Released September 8, 2011. http://www.ftc.gov/opa/2011/09/acnecure.shtm. Accessed August 25, 2012.

©2013 American Medical Association. All rights reserved.