Published in final edited form as:
JAMA Dermatol. 2013 April ; 149(4): 422–426. doi:10.1001/jamadermatol.2013.2382.

Diagnostic Inaccuracy of Smart Phone Applications for Melanoma Detection

Joel Wolf, BA, **Jacqui Moreau, BA**, **Oleg Akilov, MD**, **Timothy Patton, DO**, **Joseph C English III, MD**, **Jon Ho, MD**, and **Laura Korb Ferris, MD, PhD**[*]
Department of Dermatology, University of Pittsburgh Medical Center, Pittsburgh, PA, USA

Abstract

Objective—To measure the performance of smart phone applications which evaluate photographs of skin lesions and provide the user feedback as to their likelihood of malignancy.

Design—Case-control diagnostic accuracy study

Setting—Academic dermatology department

Participants—Digital clinical images of pigmented cutaneous lesions (60 melanoma cases and 128 benign lesion controls), all with histologic diagnosis rendered by a board-certified dermatopathologist, obtained prior to biopsy in patients undergoing lesion removal as part of routine care.

Main Outcome Measures—Sensitivity, specificity, and positive and negative predictive values of four smart phone applications designed to aid non-clinician users in determining if their skin lesion is benign or malignant.

Results—Sensitivity of the four tested applications ranged from 6.8% to 98.1%. Specificity ranged from 30.4% to 93.7%. Positive predictive value ranged from 33.3% to 42.1%, and negative predictive value ranged from 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 and the lowest sensitivity was observed for applications that use automated algorithms to analyze images.

Conclusions—The performance of smart phone applications in assessing melanoma risk is highly variable, and 3 out of 4 smart phone 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, has the potential to delay the diagnosis of melanoma and to harm users.

[*]Corresponding author: Laura Korb Ferris, MD, PhD, 3601 Fifth Avenue, 5th Floor, Pittsburgh, PA 15213, Phone: 412-647-4200, Fax: 412-647-4832, ferrislk@upmc.edu.

Disclosures: Dr. Ferris has served as an investigator and consultant for MELA Sciences. The remaining authors have no conflicts to declare.

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: Ferris, Wolf, Moreau. Acquisition of data: Ferris, Akilov, Patton, English, Ho, Wolf. Analysis and interpretation of data: Ferris, Moreau, English, Patton, Ho. Drafting of the manuscript: Ferris, Wolf, Moreau. Critical revision of the manuscript for important intellectual content: English, Patton, Ho, Akilov. Study supervision: Ferris.

INTRODUCTION

As smart phones usage increases, these devices are being used for more than communication and entertainment, and are often becoming tools that are intimately involved 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 side effects of medications, to track their caloric intake and expenditure to manage weight loss, and to log their menstrual cycles to monitor fertility. While such applications have the potential to improve patient awareness and physician-patient communication, there is also the potential that applications that provide any type of medical advice could actually result in harm to the patient if that advice is incorrect or misleading.

A review of the applications available for the two most popular smart phone platforms reveals several that are marketed to non-clinician users to assist them in deciding if a skin lesion is potentially a melanoma, or otherwise concerning, and in need of medical attention or is likely benign based upon analysis of a digital clinical image. Such applications are available either for free or for a relatively low cost compared to seeking medical consultation in person. 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 mistakenly believe that the evaluation given by such an application is, in fact, a substitute for medical advice. This risk is particularly concerning for economically disadvantaged and uninsured patients. Because a substantial percentage of melanomas are initially patient-detected[1–4], the potential impact 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. malignant nature of a series of images of pigmented skin lesions using histology 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 were already in existence at the start of the study. The images of skin lesions were selected from our database of images that are routinely captured prior to skin lesion removal to allow clinical-pathologic 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 either excluded or cropped to remove the identifiable features or information. As histologic diagnosis was used as the reference standard for subsequent analysis, we only used images for which there was a clear histologic diagnosis rendered by a board-certified dermatopathologist. 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 one 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 assessments from one of the applications were made 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 the iPhoto program and did not compromise the integrity of the images. Two of the investigators (JW and LKF) then reviewed all images for image quality and omitted those that were of poor quality or resolution.

Smart phone applications

We searched the application stores of the two most popular smart phone operating systems for applications that claim or suggest an ability to assist users in determining if 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 if they use a photograph of a skin lesion to make assessments and if they suggest any type of diagnosis or estimation of malignancy risk. These applications were then evaluated to determine whether they could be used with an existing image (i.e. 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 then selected for further evaluation. Our search yielded a total of four applications that met our criteria. Since the purpose of our study was to determine the accuracy of such applications in general and not to directly make a statement about a particular application, we have chosen not to identify the applications by their commercial name but rather to refer them as applications 1, 2, 3, and 4.

Application 1 uses an automated algorithm to detect the border of the lesion, although it also allows manual input to confirm or change the detected border. It is the only application we tested that 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, “ok,” which we considered to be a negative test, 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 either “melanoma,” which we considered to be a positive test, or “looks good” which we considered to be a negative test. If the image could not be analyzed a message of “skin condition not found” was given and we considered the image 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, or “medium risk” or “low risk,” both of which we considered to be a negative test. The presence of a medium risk category in this application presented some difficulty in analysis as it was the only application tested that gave an intermediate output. Thus, we did perform sensitivity and specificity analysis with "medium risk" lesions counting as a positive test as well since it is not clear 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 either a smart phone or from a website. This program differs from the others in that it does not use an automated analysis algorithm to evaluate images but 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 it is not known if all the images were read by the same dermatologist or by several different dermatologists. The output given is either "atypical," which we considered to be a positive test, or "typical" which we considered to be a negative test. 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 four applications was presented with each eligible pigmented skin lesion image, and evaluation was attempted. We recorded output as a test result of positive, negative, or unevaluable as described above. 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 on 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 to each other, the relative sensitivities of each application were compared using McNamar's 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. Statistical analysis was performed using Stata 12.1 software (StataCorp, College Station, TX).

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 quality, containing identifiable patient information or features, or lacking sufficient clinical or histologic information. A total of 188 lesions were evaluated using the four applications. Of these lesions 60 were melanomas (44 invasive and 16 in situ). The remaining 128 lesions were benign. The categorization of all lesions is outlined 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 either "positive," "negative," or "unevaluable" as outlined in the methods section. The primary endpoint of our study was sensitivity to melanoma categorization because most of the lesions removed in our practice are removed due to concern for malignancy, and thus we expected the specificity to be low.

As reported in Table 2, the applications considered 84.6% – 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% – 98.1%. Application 3 had the lowest sensitivity when a readout of "medium risk" was considered a negative test result. When analysis was performed considering the "medium risk" readout to be a positive test result, this resulted in a calculated sensitivity of 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 compared to each other, application 4 had higher sensitivity than the other 3 applications (p <.001 vs. applications 1 and 3, p= .02 vs. application 2).

We also calculated the PPV and NPV and confidence intervals for each application. The results are shown in Table 3. The PPVs ranged between 33.3% and 42.1%. The NPV ranged from 65.4% to 97.0%..

DISCUSSION

Over 13,000 healthcare 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 smart phone 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 Alzheimer's patients,[7] to assist in the delivery of cardiac life support,[8] and in diabetes management.[9] However, this type of evaluation is not common for applications marketed directly to consumers.

In the field of dermatology, there are several applications 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 actually give an assessment of risk or probability that a lesion is benign or malignant. This is particularly concerning as patients may substitute these readouts for standard medical consultation. Three of the four applications we evaluated do not involve a physician at any point in their evaluation. Even the best performing among these three classified 30% of the melanomas in our study as "ok."

The explosion of smart phone applications geared at health-related decision making has not gone unnoticed by the Food and Drug Administration (FDA). In July of 2011, the FDA announced plans to regulate smart phone applications that pair with medical devices that are already regulated by the FDA, such as cardiac monitors and radiologic imaging devices.[10] In June of 2012 congress approved the FDA Safety and Innovation act,[11] which allows the FDA to regulate some medical applications on smart phones. However, it is not clear how this will occur and which applications will be subject to this regulation and which will be exempt. Clarification of these questions, however, is not expected for another 18 months. In 2011, the Federal Trade Commission (FTC) fined the developers of two applications that made unsubstantiated claims to treat acne using colored light that could be shined upon the skin from a smart phone application. Both 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" (i.e. benign). While 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 ranges from free to $4.99 for evaluation of an unlimited number of lesions. Also, while applications 1, 2 and 3 provided immediate feedback on lesions (in under one minute on average), the evaluation given by application 4 was received in about 24 hours.

Our study has some intrinsic limitations. To adequately power this pilot study while restricting our inclusion criteria to lesions for which histopathology as reference standard for diagnosis was available, we were limited to the use of existing photographs of lesions that had been removed prior to the start of the study. This has several implications. First, our images were primarily of lesions that were considered to be atypical in clinical appearance by at least one dermatologist. For this reason, and because of the potentially devastating consequences of missing a melanoma (as compared to classifying a benign lesion as concerning), we made sensitivity our primary endpoint. 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, since 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 only included evaluable lesions in our analyses.

Technologies that improve the rate of melanoma self-detection have potential to improve melanoma mortality and would be welcome additions to our efforts to decrease melanoma mortality through early detection. However, there must be extreme care taken to avoid harming patients in the process. Despite disclaimers presented by each of these applications of being designed for education 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 potentially be misused. This is particularly concerning in times of economic hardship when uninsured, and even insured patients deterred by the cost of medical visit copayments, may turn to these applications as alternatives to physician evaluation. It is important for physicians to be aware of these applications, as the use of medical applications seems to be increasing over time and it is unclear if and when such applications may be subject to regulatory oversight, nor if this is appropriate. However, given the recent media and legislative interest in such applications, it is helpful for the dermatologist to be aware of those relevant to our field to aid us in protecting and educating our patients.

Acknowledgments

The project described was supported by the National Institutes of Health through Grant Numbers UL1 RR024153 and UL1TR000005.

References

1. Brady MS, Oliveria SA, Christos PJ, et al. Patterns of detection in patients with cutaneous melanoma. Cancer. Jul 15; 2000 89(2):342–347. [PubMed: 10918164]
2. Epstein DS, Lange JR, Gruber SB, Mofid M, Koch SE. Is physician detection associated with thinner melanomas? JAMA. Feb 17; 1999 281(7):640–643. [PubMed: 10029126]
3. Kantor J, Kantor DE. Routine dermatologist-performed full-body skin examination and early melanoma detection. Arch Dermatol. Aug; 2009 145(8):873–876. [PubMed: 19687416]
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. Jun; 147(6):673–678. [PubMed: 21690529]
5. Morris DP. Health-care apps for smartphones pit FDA against tech industry. The Washington Post. Jun 22.2012
6. Senior K. Smart phones: new clinical tools in oncology? Lancet Oncol. May; 2011 12(5):429–430. [PubMed: 21536224]
7. De Leo G, Brivio E, Sautter SW. Supporting autobiographical memory in patients with Alzheimer's disease using smart phones. Appl Neuropsychol. Jan; 2011 18(1):69–76. [PubMed: 21390903]
8. Low D, Clark N, Soar J, et al. A randomised control trial to determine if use of the iResus(c) application on a smart phone improves the performance of an advanced life support provider in a simulated medical emergency. Anaesthesia. Apr; 66(4):255–262. [PubMed: 21401537]
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. Jul; 2010 4(4):958–960. [PubMed: 20663462]
10. FDA. FDA outlines oversight of mobile medical applications. 2011.
11. FDA. FDA Safety and Innovation Act. 2012.
12. Commission FT. Mobile App Marketers Will Drop Baseless Claims Under FTC Settlements. 2011. Acne Cure.

Table 1

Histologic diagnosis of lesions evaluated

	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)
Dematofibroma	4 (2.1)

Table 2

Sensitivity and specificity of applications using evaluable images

Application	n (%) Evaluable	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)

CI – confidence interval

Table 3

PPV and NPV of applications using evaluable images

Application	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)

CI – confidence interval; PPV – positive predictive value; NPV –negative predictive value