HHS Public Access

Author manuscript

Br J Dermatol. Author manuscript; available in PMC 2023 April 01.

Published in final edited form as:

Br J Dermatol. 2022 April ; 186(4): 744–746. doi:10.1111/bjd.20903.

Accuracy of commercially available smartphone applications for the detection of melanoma

M.D. Sun[1,2], **J. Kentley**[1,3], **P. Mehta**[1], **S. Dusza**[1], **A.C. Halpern**[1], **V. Rotemberg**[1]

[1]Dermatology Service, Memorial Sloan Kettering Cancer Center, 1275 York Avenue, New York, NY, USA

[2]Icahn School of Medicine at Mount Sinai, New York, NY, USA

[3]Department of Dermatology, Chelsea and Westminster Hospital, London, UK

Dear Editor,

Artificial intelligence (AI) has shown promise in the analysis of images for detection of melanoma.[1] The number of available dermatology smartphone applications ('apps') is rapidly growing and there is increasing interest in apps that provide diagnosis or triage of skin lesions.[2,3] A 2020 systematic review found that nine studies evaluating six apps had poor study design and high risk of bias.[3] To date, no studies have evaluated the accuracy of apps using an independent test set of clinical images comparable with those submitted through smartphones.

Based on a futility analysis (H1: mean sensitivity and specificity ≥0·5) with a conditional power threshold of 20%, we included clinical images of 15 consecutive histologically proven invasive melanoma cases (pT1a–pT2b) and 15 histologically proven benign naevi, all in patients with lighter skin phototypes. We identified five images of benign naevi in patients with skin of colour (SoC), diagnosed clinically by two dermatologists. Images were obtained as part of usual care at Memorial Sloan Kettering Cancer Center and retrieved from our image database. Median age was 56 years (range 23–87), and 21 patients (60%) were female. Images were cropped to the lesion and are available at the International Skin Imaging Collaboration Archive (https://doi.org/10.34970/401946). Local institutional review board approval was obtained.

Publicly available smartphone and web-based dermatology apps offering AI diagnostics were identified by querying the Apple App store, Google Play store and Google Search. Images were downloaded to an Apple iPhone 11, a Samsung Galaxy S10 and a Microsoft Windows 10 PC, and uploaded to each app. Metadata regarding lesion location, morphology and history were entered when required. If apps did not allow image upload from camera roll, images were displayed on a 4K ultra-high-definition monitor and captured via smartphone.

halperna@mskcc.org .

M.D.S. and J.K. contributed equally to the work and should both be considered first authors.

For apps that output one or more diagnoses, accuracy was calculated based on the presence or absence of melanoma in the top-1 or top-3 diagnosis. For apps that provided risk-category outputs, all possible category combinations were binarized as a positive or negative melanoma output. For apps that returned a continuous risk score, accuracy was calculated based on a cutoff of ≥50% constituting melanoma.[4] Indeterminate outputs were classified as benign. Apps that provided multiple output types were evaluated separately for each.

Of 43 apps identified, 25 claimed to identify melanoma and were functional. Of these, 10 did not allow upload from camera roll and eight required metadata. Four apps rejected ≥1 image. Fifteen of 25 apps returned diagnoses, 12 of 25 risk categories and two of 25 risk scores (Figure 1). Three apps gave >1 output type.

Across top-1 measures for all apps, mean sensitivity was 0·28 [95% confidence interval (CI) 0·17–0·39)], mean specificity was 0·81 (95% CI 0·71–0·91) and mean accuracy was 0·59 (95% CI 0·55–0·62) (Figure 1). For diagnosis-based apps, risk-category-based apps and score-based apps, mean accuracy was 0·56, 0·60 and 0·64, respectively. In the 10 apps that returned at least three ranked diagnoses, mean top-1 accuracy (0·56) was higher than mean top-3 accuracy (0·41). When rejected images were classified as a benign output, mean accuracy dropped from 0·63 to 0·61 across four apps.

This study is limited by the heterogeneous output of apps preventing direct comparison. Furthermore, the retrospective nature of the clinical images, lack of melanoma in the SoC cohort and inability to directly upload all images prevents a true evaluation of real-world accuracy. The case–control study design may result in spectrum bias and an overestimation of test accuracy, as well as poor applicability to the intended population; poor applicability may also result from intended use where patients choose and take images (rather than the clinician, as in this study).

Of the included apps, only SkinScan, Derma AI and SkinVision (which had the highest accuracy, 0·81, in this study) hold a CE mark. Published studies on their performance have been evaluated as ‘low quality’.[3,5] None have US Food and Drug Administration approval.[3] Regardless of disclaimers identifying apps as educational or research tools, this poses potential risks to the public. Accuracy and sensitivity were highly variable and overall low, in keeping with previous literature, which recommends against their use.[3,6] Eight apps failed to identify a single melanoma in their top-1 ranking, and four did not include melanoma in their top-3 (AI – Detect Skin Disease, Derma AI, HealthAI and My Skin App). This risks false reassurance, and reluctance to seek medical care. Three apps demonstrated specificity <0·5, which may contribute to emotional distress in patients and increased healthcare utilization. Clinicians should be aware of app limitations and their widespread accessibility to lay users. Improved regulation and higher quality studies are necessary to bring prospectively validated algorithms to market.[7,8]

Acknowledgments:

we would like to thank Rachel Manci and Nicholas Kurtansky for their assistance with curating the images used in this study.

Conflicts of interest:

J.K. has provided services for Skin Analytics, Ltd. V.R. has provided services for Inhabit Brands, Ltd. A.C.H. has provided services for Canfield Scientific, the Harry J. Lloyd Charitable Trust and SciBase, has ownership/equity interests in HCW Health, LLC, and has a fiduciary role, intellectual property rights and ownership/equity interests in SKIP Derm, LLC.

Data availability:

the data that support the findings of this study are openly available in the International Skin Imaging Collaboration Archive at https://doi.org/10.34970/401946.

References

1. Young AT, Xiong M, Pfau J et al. Artificial intelligence in dermatology: a primer. J Invest Dermatol 2020; 140:1504–12. [PubMed: 32229141]
2. Flaten HK, St Claire C, Schlager E et al. Growth of mobile applications in dermatology – 2017 update. Dermatol Online J 2018; 24:13020/qt3hs7n9z6.
3. Freeman K, Dinnes J, Chuchu N, Takwoingi Y. Algorithm based smartphone apps to assess risk of skin cancer in adults: systematic review of diagnostic accuracy studies. BMJ 2020; 368:m127. [PubMed: 32041693]
4. Brinker TJ, Hekler A, Utikal JS et al. Skin cancer classification using convolutional neural networks: systematic review. J Med Internet Res 2018; 20:e11936. [PubMed: 30333097]
5. Udrea A, Mitra GD, Costea D et al. Accuracy of a smartphone application for triage of skin lesions based on machine learning algorithms. J Eur Acad Dermatol Venereol 2020; 34:648–55. [PubMed: 31494983]
6. Chuchu N, Takwoingi Y, Dinnes J et al. Smartphone applications for triaging adults with skin lesions that are suspicious for melanoma. Cochrane Database Syst Rev 2018; 12:CD013192.
7. Charalambides M, Flohr C, Bahadoran P, Matin RN. New international reporting guidelines for clinical trials evaluating effectiveness of artificial intelligence interventions in dermatology: strengthening the SPIRIT of robust trial reporting. Br J Dermatol 2021; 184:381–3. [PubMed: 33666954]
8. Matin RN, Dinnes J. AI-based smartphone apps for risk assessment of skin cancer need more evaluation and better regulation. Br J Cancer 2021; 124:1749–50. [PubMed: 33742148]

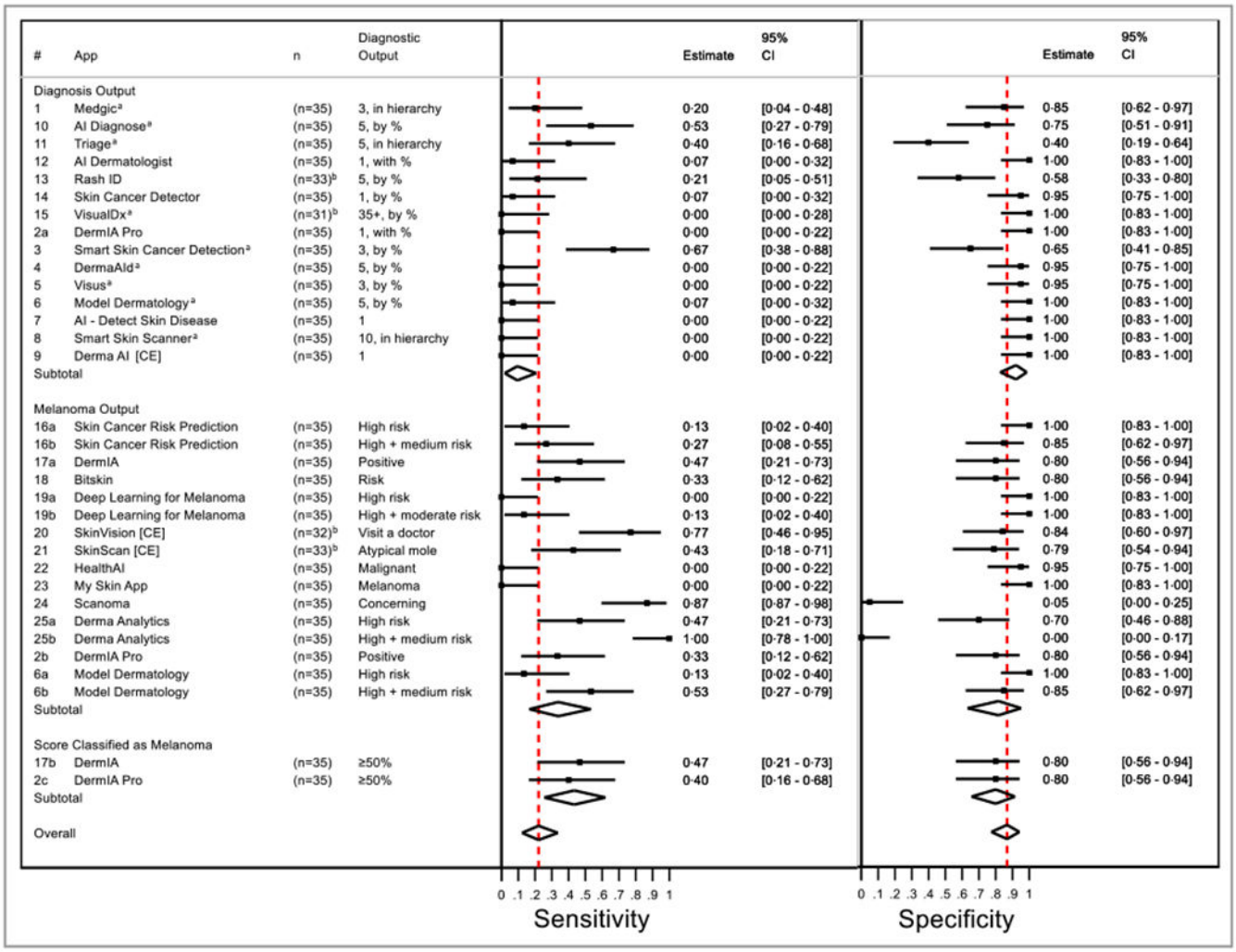

Figure 1.

Forest plots of sensitivity and specificity of commercially available AI smartphone applications for the detection of melanoma, by output category; dotted red line represents overall mean. Paired forest plots with 95% confidence intervals (CI) were estimated using multi-level mixed-effects logit models and the ‘midas’ command in Stata v16.1 (Stata Statistical Software, StataCorp, College Station, TX, USA). For diagnosis-based apps, top-1 sensitivity and specificity are presented. Accuracy = TP + TN/TP + FP + TN + FN; sensitivity = TP/TP + FN; and specificity = TN/TN + FP; where TP = true positive, TN = true negative, FP = false positive, and FN = false negative. The number of diagnoses presented by the app is shown in the ‘Diagnosis Output’ output column, ordered in a hierarchy or by percentage probability (%) for a given diagnosis. For risk-category-based apps the categories interpreted as ‘Melanoma Output’ are presented. For continuous score-based apps the score classified as melanoma is presented. Apps 1, 2a, 2b, 4, 11, 15, 20 and 21 required metadata; 2, 6 and 17 had >1 output type; 20 was based on initial outputs, images later rejected. [a]Rejected at least one image. [b]Shows top-1 accuracy. Top-3 accuracy is presented in the manuscript.