DOI: 10.1111/jdv.15873 JEADV

ORIGINAL ARTICLE

Poor agreement between the automated risk assessment of a smartphone application for skin cancer detection and the rating by dermatologists

Y. Chung,[1,†] A.A.J. van der Sande,[1,†] K.P. de Roos,[1,2] M.W. Bekkenk,[3] E.R.M. de Haas,[4] N.W.J. Kelleners-Smeets,[5] N.A. Kukutsch[6,*]

[1]Dutch Society of Dermatology and Venereology, Utrecht, The Netherlands
[2]Dermapark, Uden, The Netherlands
[3]Department of Dermatology, Academic Medical Center and Vrije University Medical Center, Amsterdam, The Netherlands
[4]Department of Dermatology, Erasmus Medical Center, Rotterdam, The Netherlands
[5]Department of Dermatology, Maastricht University Medical Center, Maastricht, The Netherlands
[6]Department of Dermatology, Leiden University Medical Center, Leiden, The Netherlands
*Correspondence: N. Kukutsch. E-mail: n.a.kukutsch@lumc.nl

Abstract

Background Several smartphone applications (app) with an automated risk assessment claim to be able to detect skin cancer at an early stage. Various studies that have evaluated these apps showed mainly poor performance. However, all studies were done in patients and lesions were mainly selected by a specialist.

Objectives To investigate the performance of the automated risk assessment of an app by comparing its assessment to that of a dermatologist in lesions selected by the participants.

Methods Participants of a National Skin Cancer Day were enrolled in a multicentre study. Skin lesions indicated by the participants were analysed by the automated risk assessment of the app prior to blinded rating by the dermatologist. The ratings of the automated risk assessment were compared to the assessment and diagnosis of the dermatologist. Due to the setting of the Skin Cancer Day, lesions were not verified by histopathology.

Results We included 125 participants (199 lesions). The app was not able to analyse 90 cases (45%) of which nine BCC, four atypical naevi and one lentigo maligna. Thirty lesions (67%) with a high and 21 with a medium risk (70%) rating by the app were diagnosed as benign naevi or seborrhoeic keratoses. The interobserver agreement between the ratings of the automated risk assessment and the dermatologist was poor (weighted kappa = 0.02; 95% CI −0.08-0.12; $P = 0.74$).

Conclusions The rating of the automated risk assessment was poor. Further investigations about the diagnostic accuracy in real-life situations are needed to provide consumers with reliable information about this healthcare application.

Received: 27 February 2019; Accepted: 19 July 2019

Conflicts of Interest

None declared.

Funding sources

None.

Introduction

The smartphone has become an integral part of our daily lives. There are many smartphone applications (apps) on the consumer market for the evaluation of moles, facilitating follow-up of lesions, performing automated risk assessments (ARAs) or providing an e-consultation with a dermatologist.[1] The SkinVision app (SVA) (SkinVision, Amsterdam, The Netherlands) has been found to be the most frequently downloaded app of this kind on the Android store.[1] Consumers can get lesions automatically analysed and receive an instant rating as low, medium or high risk. Lesions can be reviewed by a dermatologist. Various studies[2–5] that have evaluated apps using automated analysis including the SVA or similar apps showed mainly poor performance. However, all studies[2–6] so far were done in patients instead of in

[†]These authors share senior authorship

the general population and only in one pilot study[2] patients themselves were asked to indicate the lesions they were worried about. In all other studies,[3–6] the lesions were selected by a physician.

The aim of this study was to evaluate the performance of the ARA of the SVA by investigating the interobserver agreement between the dermatologist and the ARA in lesions indicated by participants of the study.

Methods

Study design and participants

The National Skin Cancer Day in the Netherlands was an annually recurring campaign to raise awareness of skin cancer and provide information on sun safety measures. It was mainly promoted through a poster campaign and local press releases. Posters were distributed to general practitioners, dermatologists and pharmacies. At the National Skin Cancer Day, participants could consult a dermatologist for a free skin check-up and receive medical advice at participating clinics. Participants were advised to consult their general practitioner if further treatment or a diagnostic procedure was necessary. Due to the legal aspects of the campaign day, it was not possible to excise suspicious lesion for histopathological investigation. At four academic hospitals in the Netherlands (Leiden University Medical Center, Amsterdam, Medical Center/VU University Medical Center, Maastricht University Medical Center and Erasmus University Medical Center, Rotterdam), participants of the National Skin Cancer Day, from the age of 18 years, were asked to take part in this study.

After written informed consent, participants were asked to point out a maximum of two lesions they were concerned about and in which case they would have potentially used the app themselves. A researcher entered the risk profile of the lesion and imaged the lesions with an iPhone using the in-app automatic camera which was facilitated by the SVA. The at that time point available version of the app was used in all centres. Between centres, different iPhone models were used, all of them supporting the app and the instant rating (green, yellow, red) was noted. Participants were blinded to the outcome of this procedure. A maximum of five attempts was made to capture the lesion; if the attempts were unsuccessful, the lesion was tagged as 'unable to perform analysis'. Afterwards, the same lesions were rated and diagnosed by an expert dermatologist in skin cancer (dermatologists working in skin cancer centres of the four aforementioned academic referral hospitals with more than 5 years of experience in skin cancer diagnosis, treatment and research in skin cancer) who was blinded for the results previously acquired by the SVA. Dermoscopy was used as a complementary diagnostic tool. Benign lesions were rated as green; lesions like atypical naevi, which should be followed up by the visitor, were rated as yellow; and (pre)malignant lesions that needed physician's consultation were rated as red. Due to the setting of the Skin Cancer Day, lesions were not verified by histopathology. At the Leiden University Medical Center and the Maastricht University Medical Center, two dermatologists rated in consensus. The local medical ethical committees approved the study.

Data were collected in Excel, and descriptive statistics were performed in SPSS (IBM SPSS Statistics for Windows, version 24.0. Armonk, NY, U.S.A: IBM Corp). Weighted Kappa (Kw) and 95% confidence intervals (CIs) were calculated to determine SVA-dermatologist agreement. Kappa values were interpreted based on the Landis and Koch guidelines.[7] The chi-square test was used to compare the characteristics of persons and lesions that could or could not be analysed by the SVA. Median and interquartile range (IQR) were measured for age. Statistical significance was stated as $P < 0.05$.

Results

On 20 May 2017, about 3500 participants were seen in 83 dermatological practices. A total of 297 participants were registered at the four academic centres on the Dutch National Skin Cancer Day. One hundred twenty-five participants (199 lesions) agreed to take part in this study. Characteristics of participants and the localization of lesions are detailed in Table 1. In 45% of the 199 lesions, the app was not able to take a picture for analysis. ARA was green (low risk) in 31% of the lesions, yellow (moderate risk) in 28% and red (high risk) in 41%, while the dermatologist rating was green in 84%, yellow in 8% and red in 8% of the lesions (Table 2a). Of the ARA red rating, 84% of the lesions were rated as green by the dermatologist, 2% was rated as yellow and 13% were rated as red. Of all green ARA, 79% were also rated green by the dermatologist. In 21% of cases, the dermatologist reported a yellow or red rating. Of the lesions where the app was unable to perform analysis, 81% were rated as green, 6% as yellow and 13% as red by the dermatologist (Table 2b). The interobserver agreement between the ARA and the

Table 1 Characteristics of participants

Number of participants	125
Sex, *n* (%)	
Male	31 (31)
Female	69 (69)
Total male and female	100 (100)
Missing	25
Age in years (median, IQR)	50 (40–62)
Number of participants with x lesions, *n* (%)	
Participants with one lesion	51 (41)
Participants with two lesions	74 (59)
Total number of lesions, *n*	199
Location of lesion, *n* (%)	
Head/neck	39 (20)
Trunk	101 (51)
Upper extremity	32 (16)
Lower extremity	27 (14)

dermatologist was poor (weighted kappa = 0.02; 95% CI -0.08-0.12; $P = 0.74$). Sixty-seven per cent of the lesions with a red rating and 70% with a yellow rating by the ARA were diagnosed as a benign naevus or a seborrhoeic keratosis by the dermatologist. Two lesions, which were diagnosed as basal cell carcinoma (BCC) by the dermatologist, got either a green or yellow rating by the ARA (Table 3a). In nine cases of BCC, four atypical naevi and one lentigo maligna, the app was unable to perform the analysis (Table 3b). Comparison between the characteristics of patients and location per lesion that could or could not be analysed by the app only showed that in the head/neck area in a significantly higher proportion of lesions, the app was unable to perform the analysis (Table S1).

Table 2 (a) ARA vs rating of the dermatologist; (b) ARA unable to perform analysis vs rating of the dermatologist

(a)		Rating dermatologist, *n* (%)†			Total of rating ARA, *n*
		Green	Yellow	Red	
ARA	**Green**	27 (79*)	5 (15**)	2 (6***)	34
	Yellow	26 (87*)	3 (10**)	1 (3***)	30
	Red	38 (84*)	1 (2**)	6 (13***)	45
Total rating dermatologist, *n*		91	9	9	109

(b)	Rating dermatologist, *n* (%)			Total, *n* (%)
	Green	Yellow	Red	
ARA unable to perform analysis, *n*	73 (81)	5 (6)	12 (13)	90 (100)

†Percentage of the ARA that was given, respectively, a green*/yellow**/red*** rating by the dermatologist.
ARA, automated risk assessment.
Benign lesions were rated as green; lesions like atypical naevi, which should be followed up by the participant, were rated as yellow; and (pre)malignant lesions that needed physician's consultation were rated as red by the dermatologist.

Discussion

This study showed a poor agreement of the rating of skin lesions selected by participants of a skin cancer campaign day between the ARA and the rating and the diagnosis of the dermatologist. This is in line with the findings of previous studies[2,3] testing the same app on patients visiting a pigmented lesion clinic. The current study differs from the previous ones in that in this study, a large cohort of participants of a campaign day and not patients were included in a multicentre trial. In the previous studies, the group was preselected by including patients who already intended to see a dermatologist. Nabil *et al.*[2] suggested that their outcome would even be stronger in the general population. They argued[2] that laypersons are not able to distinguish between cancerous lesions and benign lookalikes such as a benign naevus and seborrhoeic keratosis and that the app is not able to perform well in unselected lesions, either. The current study group represents the general population better since the participants were visitors of a campaign day and not patients. In accordance with this fact, most lesions indicated as suspicious by the SVA (66% of the red rating) were benign naevi and seborrhoeic keratoses. In another study,[8] only lesions suspicious for melanoma were included after selection by a dermatologist. A sensitivity of 73% and a specificity of 83% were found for melanoma. In this setting, most benign lesions like banal naevi and seborrheic keratoses had already been excluded. Thissen *et al.*[9] validated a modification on the SVA algorithm and found a sensitivity of 80% and a specificity of 78% for (pre)malignant lesions. Also in this study,[9] lesions were selected by a dermatologist and it is not clear from the study design if specific lesions were selected for the purpose of validation. Other studies[3–5] which also included malignant lesions testing the SVA or a similar anonymous app showed a low accuracy of the app. Ngoo *et al.*[3] who tested the SVA and several other apps on both benign and malignant lesions concluded that the agreement between the apps and the

Table 3 (a) Rating of the ARA vs diagnosis of the dermatologist; (b) ARA unable to perform analysis vs diagnosis of the dermatologist

(a)		Diagnosis dermatologist, *n* (%)										Total, *n*
		Benign naevus	Atypical naevus	Seb. keratosis	Actinic keratosis	BCC	M. Bowen	Solar lentigo	Dermatofibroma	Angioma	Other benign	
Rating ARA, *n*	**Green**	20 (59)	4 (12)	4 (12)	0 (0)	1 (3)	0 (0)	3 (9)	2 (6)	0 (0)	0 (0)	34
	Yellow	19 (63)	2 (7)	2 (7)	0 (0)	1 (3)	0 (0)	3 (10)	2 (7)	1 (3)	0 (0)	30
	Red	15 (33)	1 (2)	15 (33)	1 (2)	4 (9)	1 (2)	2 (4)	3 (7)	0 (0)	3 (7)	45
Total *n*		54	7	21	1	6	1	8	7	1	3	109

(b)	Diagnosis dermatologist, *n (%)*†										Total, *n* (%)
	Benign naevus	Atypical naevus	Seb. keratosis	Actinic keratosis	BCC	Solar lentigo	Lentigo Maligna	Dermatofibroma	Angioma	Other benign	
ARA unable to perform analysis, *n*	35 (39)	4 (4)	14 (16)	8 (9)	9 (10)	7 (8)	1 (1)	3 (3)	3 (3)	6 (7)	90 (100)

†Percentage of the ARA rating that was given a specific diagnosis by the dermatologist.
Abbreviations: ARA, automated risk assessment; BCC, basal cell carcinoma.

dermatologist was limited. Doraiay *et al.*[5] also demonstrated a very low diagnostic accuracy for a similar app and showed that eight of the nine high-risk lesions were missed. Also, Wolf *et al.*[4] who tested three apps with an automated rating found that 30% or more of the melanomas were classified as unconcerning by the apps. A recent review[10] concluded that 'existing automated apps are unreliable' and that certificates do not implicate good performance.

What was furthermore remarkable in this study was the fact that 45% of the lesions could not be analysed by the app. Investigators in two earlier studies[3,8] also had difficulties analysing suspicious skin lesions. They reported that they failed to analyse up to 26% of the lesions and Ngoo *et al.*[3] wondered how consumers would respond. They[3] also demonstrated that using the iOS platform, which we used in our study, resulted in less failures than using the android system. Among others, the lesions that could not be analysed by the app in this study were nine basal cell carcinomas, one lentigo maligna and four atypical naevi. We do not know if consumers would have initiated a medical consultation in these cases. Furthermore, it might have been more problematic to acquire adequate pictures in the head/neck area. We found a significant higher proportion of lesions that could not be analysed by the app in that anatomic area (Table S1).

The fact that most lesions were benign stresses our opinion that the current study setting comes close to a real-life situation where most lesions that will be tested by the consumer are benign.

Our study has certain limitations. Histopathological assessment was not carried out since it was a campaign day. Furthermore, the app was unable to make an assessment in a large number of cases, even though all steps were taken to ensure optimal imaging. Thus, this study could have underestimated the capabilities of the app. Conversely, the aim of this study was to test the app in a daily practice setting in which app failures are inherently and unavoidably present.

Based on the results of this study, it is not possible to make statements about SVA's ability to detect melanoma. Only one lentigo maligna was diagnosed by the dermatologists in our study. The app was unable to perform an assessment on that lesion. The strength of our study is the broad spectrum of participants due to the multicentre setting at an open day. Moreover, participants selected the lesions themselves and not the dermatologist. The option to get a review of the picture by an in-house dermatologist which was just introduced at that point was not investigated. At the moment of writing, pictures of lesions with a red rating can be reviewed by a dermatologist for free. However, a recent review article[11] about teledermatology found a variable diagnostic agreement of 51–85% with the reference standard (histopathology for excised lesions and clinical diagnosis for others) for the diagnosis of skin cancer. The conclusion of a Cochrane review[12] about teledermatology for the diagnosis of skin cancer was that '. . .the evidence base to support its ability to accurately diagnose lesions and to triage lesions from primary to secondary care is lacking. . .'.

We did not investigate the question if the use of different iPhone models might have influenced our results. However, the rating provided by the app should be the same with different phone models, since all of them supported the app.

Despite the fact that participants of the study were not patients but visitors of the National Skin Cancer day, we still cannot exclude selection bias. Most participants were female (69%) with a median age of 50 from four geographical regions in the Netherlands covering most of the population. We expect that in real life, users of the app might be younger but also predominately female since women are more likely to perform skin self-examination.[13,14]

Other applications of the app like the option to follow-up a lesion and the possibility to send a picture to a dermatologist while a patient is under follow-up were not investigated in this study but might be of potential benefit.

With technology further evolving, new and updated versions of apps, emerging deep learning algorithms and artificial intelligence networks[15–17] will inevitably come to play in skin cancer care in the future. But before integrating them into daily practice, safety and efficacy need to be proven.

Our results highlight that caution is warranted before recommending and using these applications in a real-life setting. A European CE certification only means that a product has technically met EU health, safety and environmental requirements and is based on self-certification (ce.europe.eu). For consumers, it is important to know at a glance if there is reliable scientific support for the claims of an app.[6] Therefore, regulations integrating scientific research and validation in real-life situations are urgently needed to provide consumers with reliable information.

Acknowledgements

The authors would like to thank MH Vermeer for his assistance in data acquisition at the Dutch National Skin Cancer Day, L Teligui and M Hofhuis for support with the analysis and J Goeman for his valuable advice about the concept of the paper and the statistical analysis.

References

1 Ngoo A, Finnanae A, Mc Meniman E *et al.* Fighting melanoma with Smartphones: a snapshot of where we are a decade after app stores opened their doors. *Int J Med Inform* 2018; **118**: 99–112.
2 Nabil R, Bergman W, Kukutsch NA. Poor agreement between a mobile phone application for the analysis of skin lesions and the clinical diagnosis of the dermatologist, a pilot study. *Br J Dermatol* 2017; **177**: 583–584.
3 Ngoo A, Finnane A, McMeniman E *et al.* Efficacy of smartphone applications in high-risk pigmented lesions. *Australas J Dermatol* 2018; **59**: e175–e182.
4 Wolf JA, Ferris LK. Diagnostic inaccuracy of smartphone applications for melanoma detection–reply. *JAMA Dermatol* 2013; **149**: 422–426.
5 Dorairaj JJ, Healy GM, McInerney A *et al.* Validation of a melanoma risk assessment smartphone application. *Dermatol Surg* 2017; **43**: 299–302.

6 Buechi R, Faes L, Bachmann LM *et al.* Evidence assessing the diagnostic performance of medical smartphone apps: a systematic review and exploratory meta-analysis. *BMJ Open* 2017; **7**: e018280. https://doi.org/10.1136/bmjopen-2017-018280.
7 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**: 159–174.
8 Maier T, Kulichova D, Schotten K *et al.* Accuracy of a smartphone application using fractal image analysis of pigmented moles compared to clinical diagnosis and histological result. *J Eur Acad Dermatol Venereol* 2015; **29**: 663–667.
9 Thissen M, Udrea A, Hacking M *et al.* mHealth app for risk assessment of pigmented and nonpigmented skin lesions-a study on sensitivity and specificity in detecting malignancy. *Telemed J E Health* 2017; **23**: 948–954.
10 Rat C, Hild S, Rault Sérandour J *et al.* Use of smartphones for early detection of melanoma: systematic review. *J Med Internet Res* 2018; **20**: e135.
11 Finnane A, Dallest K, Janda M, Soyer HP. Teledermatology for the diagnosis and management of skin cancer: a systematic review. *JAMA Dermatol* 2017; **153**: 319–327.
12 Chuchu N, Dinnes J, Takwoingi Y *et al.* Teledermatology for diagnosing skin cancer in adults. *Cochrane Database Syst Rev* 2018; **12**: CD013193.
13 Weinstock MA, Martin RA, Riscia PM *et al.* Thorough skin examination for the early detection of melanoma. *Am J Prev Med* 1999; **17**: 169–175.
14 Robinson JK, Fisher SG, Turrisi RJ. Predictors of skin self-examination performance. *Cancer* 2002; **95**: 135–146.
15 Esteva A, Kuprel B, Novoa RA *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; **542**: 115–118.
16 Haenssle HA, Fink C, Schneiderbauer R *et al.* Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncol* 2018; 1–7. https://doi.org/10.1093/annonc/mdy166.
17 Tschandl P, Rosendahl C, Akay BN *et al.* Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks. *JAMA Dermatol* 2019; **155**: 58–65.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Table S1. Comparison between the characteristics of participants and location per lesion that could or could not be analyzed by the ARA.