COMMENT

AI-based smartphone apps for risk assessment of skin cancer need more evaluation and better regulation

Rubeta N. Matin[1] and Jacqueline Dinnes [ID][2,3]

Smartphone applications ("apps") with artificial intelligence (AI) algorithms are increasingly used in healthcare. Widespread adoption of these apps must be supported by a robust evidence-base and app manufacturers' claims appropriately regulated. Current CE marking assessment processes inadequately protect the public against the risks created by using smartphone diagnostic apps.

MAIN

With an ever increasing skin cancer burden on healthcare, technologically enhanced diagnostic tools such as smartphones have potential to improve triage and provide earlier and more accurate diagnosis of all skin cancers, in the hope of improving both morbidity and mortality.[1] In the UK, 8 in 10 adults now own a smartphone[2] and a wealth of smartphone applications ("apps") with a dermatological focus are available. Between 2014 and 2017, 235 new dermatology smartphone apps became available to download; teledermatology apps having the largest market share.[3] With the apparent explosion of artificial intelligence (AI) applications in medicine and given the routine acquisition of images of suspicious skin lesions in dermatology, the skin cancer field is set for exploitation of new and evolving machine learning techniques.[1] Colour- and symmetry-based analyses of images of suspicious skin lesions, combined with a simple graphical user interface theoretically allow an immediate risk assessment and subsequent 'next steps' recommendation to app users. If sufficiently accurate, AI-based apps have the potential not only to encourage those with high risk lesions to quickly seek appropriate specialist advice, but also to reassure the 'worried well' that their risk of skin cancer is low. On the other hand, apps with poor diagnostic performance risk false reassurance and inappropriate delays in obtaining medical assessment, and have potential to further overwhelm health care services if benign lesions are wrongly flagged as high risk.

What is the evidence base?

A recent systematic review of algorithm-based smartphone apps identified evidence of diagnostic accuracy for only six apps for skin lesion risk stratification.[4] Only two of the six apps are currently available to download (SkinVision and TeleSkin's skinScan app); the remaining four could not be found online or had been withdrawn from the market.[4] No published evidence was identified for skinScan, however three studies of SkinVision showed small improvements in diagnostic accuracy over time. A subsequently published study of SkinVision reported considerably higher sensitivity of 95% and specificity 78% for identification of malignant or pre-malignant lesions.[5]

A sensitivity of 95% is an impressive headline result, however not only do serious flaws in the evidence base call into question the validity of this result,[6] but there is considerable potential for harm. Used in a low prevalence, real-world setting, e.g. using UK age-standardised incidence for non-melanoma skin cancer of 257 per 100,000,[7] an app with hypothetical sensitivity of 95% and specificity 80%, would have a positive predictive value of only 1.2%. With around 20,000 false positive results for every 100,000 app users, the potential consequences to healthcare services are huge.

Flaws in the evidence

Available clinical evaluations have a number of serious flaws. Firstly, and most importantly, AI algorithms have been developed and evaluated using images of highly suspicious skin lesions that have been selected for examination by a skin specialist and have subsequently undergone biopsy, or have incorporated app users' data in a way that will have biased results.[4,6] Such studies include a narrow spectrum of lesion types, inflating estimates of both sensitivity and specificity.

Secondly, the images used have been taken by study investigators using study phones often under optimal conditions rather than by smartphone users with their own devices. Variable image quality will affect how well an algorithm performs and lead to unevaluable images, which are often excluded from smartphone app evaluations. Up to 10 attempts at image acquisition have been reported,[4] seriously affecting the usability of apps in practice.

Finally, most studies have relied on expert diagnosis to confirm the presence of a benign skin lesion with no clinical follow-up to identify any false-negative results, resulting in overstated claims regarding app sensitivity.

Regulatory approval

Under the EU Medical Device Directive, smartphone apps that make a medical claim (for example to 'detect' or 'aim to catch' melanoma

[1]Department of Dermatology, Oxford University Hospitals NHS Foundation Trust, Oxford, UK; [2]Test Evaluation Research Group, Institute of Applied Health Research, University of Birmingham, Edgbaston, Birmingham, UK and [3]NIHR Birmingham Biomedical Research Centre, University Hospitals Birmingham NHS Foundation Trust and University of Birmingham, Birmingham, UK
Correspondence: Jacqueline Dinnes (j.dinnes@bham.ac.uk)

or skin cancer at an earlier stage[4]) are class 1 medical devices that do not require independent regulatory inspection. App developers effectively 'self-certify' and apply CE (Conformit Europenne) marking, usually advertising disclaimers that app results cannot replace healthcare advice. There is no requirement for app manufacturers to provide evidence of how well the app performs in the population in which it will be used in practice. Even when the EU Medical Device Regulations and the new UK Conformity assessment "UKCA" come into force in May 2021 and January 2021, respectively, the minimal performance requirement is for the app to perform as it claims to. This is an incredibly low bar for a technology that could inform an individual's decision-making regarding whether or not to seek healthcare advice for a potentially fatal skin lesion.

The US Food and Drug Administration (FDA) has a stricter assessment process for smartphone apps, considering "a risk to a patient's safety if the mobile app were to not function as intended".[8] It is perhaps telling that no skin cancer risk stratification smartphone app has received FDA approval to date. With the changing regulatory system for medical devices in the UK on leaving the EU, there is an opportunity for regulators to enforce more stringent requirements, to ensure app 'performance' is established in a clinically relevant cohort and in a clinically meaningful way. Moreover, it is essential that both healthcare professionals and regulators are alerted to the potential harm that poorly performing diagnostic or risk stratification apps create.

Informed choice?
With the exponential growth in healthcare apps, it is challenging for smartphone users and clinicians to make informed choices. App user 'ratings' have been shown to be poor indicators of the clinical utility or usability of health-related apps, with few apps addressing the needs of the patients who could benefit the most.[9] Skin cancer specialists must encourage more informed app choice by patients while acknowledging the limitations. In the absence of more robust regulation, a generic framework such as the five-level pyramid for app evaluation and selection[10] provides one possible route to ensuring that app choice is informed in regard to safety, data privacy, clinical evidence, usability, and data integration.

To date, there is no high-quality evidence for the accuracy of algorithm-based smartphone apps for risk stratification of skin cancer when used by the general population of smartphone app users. Regulators should take action not only to require independent appraisal of the clinical evidence supporting smartphone apps, but to require clinically relevant evidence in order to protect the public from potential harms. It is the role of healthcare professionals to be aware of the limitations of these apps to reliably identify serious skin cancers such as melanoma and to educate their patients about this.

ADDITIONAL INFORMATION
Ethics approval and consent to participate Not applicable.

Consent to publish Not applicable.

Data availability Not applicable.

Competing interests The authors declare no competing interests.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES
1. Ferrante di Ruffano, L., Takwoingi, Y., Dinnes, J., Chuchu, N., Bayliss, S.E., Davenport, C. et al. Computer-assisted diagnosis techniques (dermoscopy and spectroscopy-based) for diagnosing skin cancer in adults. *Cochrane Database Syst. Rev.* **12**, CD013186 (2019).
2. Ipsos-MORI. Technology tracker Q3. https://www.ipsos.com/sites/default/files/ct/publication/documents/2018-10/techtracker_q3_2018_final2.pdf (2018).
3. Flaten, H. K., St Claire, C., Schlager, E., Dunnick, C. A. & Dellavalle R. P. Growth of mobile applications in dermatology—2017 update. *Dermatol Online J.* **24**, 13030/qt3hs7n9z6 (2018)
4. Freeman, K., Dinnes, J., Chuchu, N., Takwoingi, Y., Bayliss, S.E., Matin, R.N. et al. Algorithm based smartphone apps to assess risk of skin cancer in adults: systematic review of diagnostic accuracy studies. *BMJ* **368**, m127 (2020).
5. Udrea, A., Mitra, G.D., Costea, D., Noels, E.C., Wakkee, M., Siegel, D.M. et al. Accuracy of a smartphone application for triage of skin lesions based on machine learning algorithms. *J. Eur. Acad. Dermatol. Venereol.* **34**, 648–655 (2020).
6. Deeks, J. J., Dinnes, J. & Williams, H. C. Sensitivity and specificity of SkinVision are likely to have been overestimated. *J. Eur. Acad. Dermatol. Venereol.* **34**, e582–e583 (2020).
7. Cancer Research UK. Non-melanoma skin cancer incidence statistics (2016).
8. FDA. Mobile medical applications—guidance for industry and food and drug administration staff (Food and Drug Administration, Rockville MD, 2015).
9. Singh, K., Drouin, K., Newmark, L.P., Lee, J., Faxvaag, A., Rozenblum, R. et al. Many mobile health apps target high-need, high-cost populations, but gaps remain. *Health Aff. (Millwood)* **35**, 2310–2318 (2016).
10. Torous, J. B., Chan, S.R., Gipson, S.Y.M.T., Kim, J.W., Nguyen, T.Q., Luo, J. et al. A hierarchical framework for evaluation and informed decision making regarding smartphone apps for clinical care. *Psychiatr. Serv.* **69**, 498–500 (2018).