# The poor performance of apps assessing skin cancer risk

These apps are the product of inadequate evaluation and regulation

Jessica Morley *DataLab policy lead*[1], Luciano Floridi *professor of philosophy and ethics of information*[2], Ben Goldacre *DataLab director*[1]

[1]Nuffield Department of Primary Care, University of Oxford, Oxford OX2 6GG, UK; [2]Oxford Internet Institute, University of Oxford, Oxford, UK

Over the past year, technology companies have made headlines claiming that their artificially intelligent (AI) products can outperform clinicians at diagnosing breast cancer,[1] brain tumours,[2] and diabetic retinopathy.[3] Claims such as these have influenced policy makers, and AI now forms a key component of the national health strategies in England, the United States, and China.

It is positive to see healthcare systems embracing data analytics and machine learning. However, there are reasonable concerns about the efficacy, ethics, and safety of some commercial, AI health solutions.[4 5] Trust in AI applications (or apps) heavily relies on the myth of the objective and omniscient algorithm, and our systems for generating and implementing evidence have not yet met the new specific challenges of AI. They may even have failed on the basics. In a linked article, Freeman and colleagues[6] (doi:10.1136/bmj.m127) throw these general concerns into stark relief with a close examination of the evidence on diagnostic apps for skin cancer.

## Exposing inaccuracies

The authors report results from a systematic review of studies evaluating the accuracy of smartphone apps that were offered directly to the public for risk stratification of skin lesions. Nine studies were included, evaluating a total of six apps. Even though methodological decisions made by the instigators of the studies probably led to overestimation of the apps' real world performance, Freeman and colleagues still found evidence for accuracy to be lacking.

Some apps gave conflicting management advice for the same lesions, and their recommendations were commonly inconsistent with clinical histopathological results. In short, little evidence indicates that current AI apps can beat clinicians when assessing skin lesion risk, at least not in a verifiable or reproducible form.

## Misleading regulation

Currently, two apps from the study are available in the UK. Freeman and colleagues found no peer reviewed, published studies evaluating the Teleskin skinScan app. The second, SkinVision, when validated against expert recommendations was found to be poor. Yet both are approved and regulated as "class I medical devices"; and both have a CE mark.

This official approval will give consumers the impression that the apps have been assessed as effective and safe. But "class I" is the European classification for low risk devices, such as plasters and reading glasses. The implicit assumption is that apps are similarly low risk technology. But shortcomings in diagnostic apps can have serious implications: for patients and the public, risks include psychological harm from health anxiety or "cyberchondria," and physical harm from misdiagnosis or overdiagnosis; for clinicians there is a risk of increased workload, and changes to ethical or legal responsibilities around triage, referral, diagnosis, and treatment; for the system, there is a risk of inappropriate resource use, and even loss of credibility for digital technology in general.

## Doing better

The current regimen is clearly unsatisfactory. Collectively as a society we must decide what amounts to good evidence when evaluating health apps; who is responsible for generating, validating, and appraising this evidence; and how post-market monitoring of regularly updated software should be organised. These are complex questions.

Regulators clearly have a role. We must decide which activities they will regulate: risk stratification apps clearly perform a medical function; "wellness" apps for meditation and mindfulness are a grey area, but could nonetheless cause psychological harm. Regulators most accustomed to managing medicines will need new skills to evaluate digital technology. But wherever the perimeter is drawn, they must avoid false reassurance: when regulators are notevaluating technology, they should clearly flag this to patients and policy makers.

Softer governance measures (eg, policies and standards) from governing bodies such as NHS England, can facilitate the

Correspondence to: J Morley jessica.morley@phc.ox.ac.uk

creation of rational and transparent markets. Clinicians, patients, and commissioners are all potential customers for health apps. Guidance, such as that recently produced by Public Health England on evaluating digital health products,[7] can help ensure that they each know enough to require, find, understand, critically evaluate, and apply good evidence, within reasonable limits. This is likely to help drive better innovation, by rewarding only products that deliver tangible benefits. Clinicians should also be trained to evaluate the tools they recommend to patients, avoid the pitfalls of automation bias, and identify the clinical tasks that can be automated safely.

Lastly, we need a cultural shift. It must become the norm, or social expectation, that all those developing health apps and AI solutions support third party access to data, in a trustworthy manner; and code, within the parameters of technical feasibility, while respecting ownership of intellectual property. This would facilitate competition, reproducibility, audit, and error correction,[8] driving up the overall quality of solutions available on the market. It would also enable more independent real world evaluations of market solutions to be conducted, provided that funders are willing to support this type of research.

Collectively, these actions will improve evidence and transparency across the whole algorithm lifecycle.[9 10] Reliable evaluations must find the truth, purchasers must require and use those truths, and regulators and other governing bodies must support and enhance these processes. Without better information patients, clinicians, and other stakeholders cannot be assured of an app's efficacy, and safety.

Competing interests: We have read and understood BMJ policy on declaration of interests. JM is a recent employee of NHSX, the governing body for digital, data, and technology policy in the NHS, and has received a research grant from the Digital Catapult in the past 12 months. Neither organisation has been involved in the writing of this editorial.

1 McKinney SM, Sieniek M, Godbole V, etal . International evaluation of an AI system for breast cancer screening. *Nature* 2020;577:89-94. 10.1038/s41586-019-1799-6 31894144
2 Hollon TC, Pandian B, Adapa AR, etal . Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. *Nat Med* 2020;26:52-8. 10.1038/s41591-019-0715-9 31907460
3 De Fauw J, Ledsam JR, Romera-Paredes B, etal . Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med* 2018;24:1342-50. 10.1038/s41591-018-0107-6 30104768
4 Shaw J, Rudzicz F, Jamieson T, Goldfarb A. Artificial intelligence and the implementation challenge. *J Med Internet Res* 2019;21:e13659. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85069551590&doi=10.2196%2f13659&partnerID=40&md5=93d1414c9b5bc483b20ef58d9df41fb410.2196/13659 31293245
5 Van Calster B, Wynants L, Timmerman D, Steyerberg EW, Collins GS. Predictive analytics in health care: how can we know it works?*J Am Med Inform Assoc* 2019;26:1651-4. 10.1093/jamia/ocz130 31373357
6 Freeman K, Dinnes J, Chuchu N, etal . Algorithm based smartphone apps to assess risk of skin cancer in adults: systematic review of diagnostic accuracy studies. *BMJ* 2020;368:m127.
7 Public Health England. Evaluating digital health products. 2020. https://www.gov.uk/government/collections/evaluating-digital-health-products
8 Goldacre B, Morton CE, DeVito NJ. Why researchers should share their analytic code. *BMJ* 2019;367:l6365. 10.1136/bmj.l6365 31753846
9 Crawford K, Calo R. There is a blind spot in AI research. *Nature* 2016;538:311-3. 10.1038/538311a 27762391
10 Morley J, Floridi L, Kinsey L, Elhalal A. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Sci Eng Ethics* 2019;10.1007/s11948-019-00165-5. 31828533