nature machine intelligence

PERSPECTIVE

https://doi.org/10.1038/s42256-021-00331-0

Direct-to-consumer medical machine learning and artificial intelligence applications

Boris Babic[1,2], Sara Gerke[3], Theodoros Evgeniou[1,2] and I. Glenn Cohen[4]✉

Direct-to-consumer medical artificial intelligence/machine learning applications are increasingly used for a variety of diagnostic assessments, and the emphasis on telemedicine and home healthcare during the COVID-19 pandemic may further stimulate their adoption. In this Perspective, we argue that the artificial intelligence/machine learning regulatory landscape should operate differently when a system is designed for clinicians/doctors as opposed to when it is designed for personal use. Direct-to-consumer applications raise unique concerns due to the nature of consumer users, who tend to be limited in their statistical and medical literacy and risk averse about their health outcomes. This creates an environment where false alarms can proliferate and burden public healthcare systems and medical insurers. While similar situations exist elsewhere in medicine, the ease and frequency with which artificial intelligence/machine learning apps can be used, and their increasing prevalence in the consumer market, calls for careful reflection on how to effectively regulate them. We suggest regulators should strive to better understand how consumers interact with direct-to-consumer medical artificial intelligence/machine learning apps, particularly diagnostic ones, and this requires more than a focus on the system's technical specifications. We further argue that the best regulatory review would also consider such technologies' social costs under widespread use.

The global market for medical artificial intelligence or machine learning applications (medical AI/ML apps) is currently valued at above US$10 billion[1]. Some of these apps are marketed directly to healthcare professionals, but there is a burgeoning industry of medical AI/ML apps that are marketed directly to consumers for personal use. We call these direct-to-consumer (DTC) medical AI/ML apps. Most such apps are predicated around a predictive or diagnostic function—they offer cheap and purportedly accurate diagnoses of various conditions. A prominent example is the Apple Watch irregular rhythm notification feature, an app that is marketed directly to consumers for personal screening suggestive of certain heart disorders, particularly a condition known as atrial fibrillation (AFib)[2].

Whether or not such DTC AI/ML apps are subject to regulatory review by regulators such as the US Food and Drug Administration (FDA) depends on whether the apps are considered to be medical devices and their risks to patients. For example, while the Apple Watch irregular rhythm notification feature has received marketing authorization as a Class II (moderate risk) device[2], it is rather an exception, and many DTC AI/ML apps are not reviewed by the FDA at all. However, even if the FDA undertakes a review, we argue that for reasons beyond just individual user risks, the level of scrutiny such apps have so far received is not as high as it should be. In particular, DTC medical AI/ML apps may seem innocuous because they provide additional information, often in the form of a diagnostic assessment, which consumers are not obliged to rely on; and when in doubt, consumers can seek professional healthcare advice. But herein lies the problem. DTC medical AI/ML apps are marketed to imperfectly rational and risk-averse decision-makers and are designed for cheap, instantaneous and repeated use. As a result, errors—particularly false positive judgements—can proliferate quickly, thereby generating substantial negative externalities and placing an undue burden on the healthcare system as a whole through unnecessary doctor visits, over-testing of people who are otherwise not at risk for disease and other strains on scarce medical resources. Our key argument is that even if the risk borne by any given individual from failures of such devices may be low, the aggregate cost on public healthcare systems and private insurers can be quite large. To be sure, the cost of proliferating false positive judgements is not realized exclusively at the social level. The personal costs of overdiagnosis and overtreatment can be substantial, but in what follows, we focus especially on the social costs.

This discrepancy between individual utility and social externalities is not dissimilar from the situation we face in other contexts, such as the oversubscription of antibiotics. From the individual perspective of a person's safety, the marginal round of antibiotics is generally safe, though in many instances may add little or no value to the patient's recovery. But from a social perspective, the aggregate cost of widespread overreliance on antibiotics creates negative externalities in the form of antibiotic-resistant bacteria, which then require further expenditures[3]. Both are instances of a healthcare tragedy of the commons where individual incentives can be contrary to social welfare considerations.

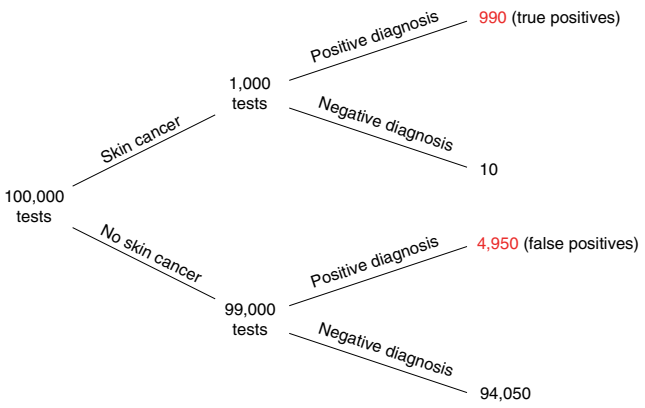
For this reason, if regulators employ a narrow individual user-based approach for determining whether to review and permit marketing of DTC medical AI/ML apps, they miss an important part of the picture. The best form of regulatory review—which may require additional statutory authority—would also consider the aggregate social costs generated through such devices' mass usage and would pay particular attention to how consumers behave when they interact with these apps rather than simply on the apps' accuracy levels. While we recognize that such a review may be difficult for many regulators under their current institutional design and statutory authority (the US FDA, for example, draws a firm line at not regulating the practice of medicine), we make several recommendations for how regulators can move in the right direction.

While our focus is on DTC AI/ML, some of our recommendations may also be relevant to DTC apps in general. We note,

[1]INSEAD, Fontainebleau, France. [2]INSEAD, Singapore, Singapore. [3]The Petrie-Flom Center for Health Law Policy, Biotechnology, and Bioethics at Harvard Law School, The Project on Precision Medicine, Artificial Intelligence, and the Law (PMAIL), Cambridge, MA, USA. [4]Harvard Law School, Cambridge, MA, USA. ✉e-mail: igcohen@law.harvard.edu

Box 1 | The (dis)value of information

Suppose that a skin cancer screening app is marketed directly to consumers, which takes an image (a pigmented skin lesion) as its input and provides a binary 'yes'/'no' disease prediction as its output. Once downloaded, a consumer is able to perform tests on any lesion of her body. Furthermore, she is able to retest any particular lesion (for example, from a different angle, at an alternative distance or under different lighting). Suppose that the underlying disease is present in 1% of skin lesions population-wide, and that this app is highly accurate, with 99% sensitivity and 95% specificity. If a patient receives a positive diagnosis, then the rational Bayesian posterior that she has the disease should be 16.7%. To make things more realistic, suppose that the patient performs five tests on a single lesion, obtaining two negative diagnoses, followed by three positive diagnoses. If we assume for the sake of illustration that the tests are independent, conditional on disease state, then following all five tests, the probability of disease will be only 0.86%. These are the rational Bayesian verdicts. The reason these numbers look strikingly low is because for medically rare diseases such as skin cancer, while most negatives are true negatives, most positives are false positives. The figure illustrates this situation after 100,000 tests. Out of 5,940 positive diagnoses, 4,950 of them (84%) are false positives. Meanwhile, out of 94,060 negative diagnoses, only 10 (0.01%) are false negatives.



Comparison of true positives to false positives with a highly accurate hypothetical AI/ML skin cancer screening app after 100,000 diagnoses.

However, the situation in practice is probably much different. Most users of AI/ML apps are not domain experts and are thereby not aware of the prevalence of disease. Moreover, they are extra sensitive to adverse health outcomes. As a result, and as previous research in judgement and decision-making indicates, they are likely to assume that the prior odds are closer to 1:1 (ref. [10]), thereby putting more weight on the machine prediction than a rational Bayesian should. In this case of assuming prior odds of 1:1, the posterior probability that a person has the disease, after observing one positive diagnosis, will be approximately 95% (a sixfold increase from the Bayesian answer, which is 16.7%). Meanwhile, the posterior probability that one has the disease, after observing two negative diagnoses followed by three positive ones, will be 46.2% (a 53-fold increase from the rational Bayesian answer, which is 0.86%). While this is only a theoretical illustration, it points to the perils of overestimating posterior probability of disease from introducing AI/ML assessments into an imperfectly rational and risk-averse population.

however, that AI/ML products have certain relevant characteristics that other products do not. Most importantly, their outputs tend to be probabilistic, unlike other conventional devices that report exact measurements. This requires a user to incorporate more nuanced information into their own body of belief. This is exacerbated by the fact that the probabilistic assessment is typically produced using an algorithm that is not interpretable, even by very informed users, thereby making it even more difficult to contextualize the output.

The FDA's current regulatory approach

The FDA can only regulate systems that meet the definition of a medical device under Section 201(h) of the Federal Food, Drug, and Cosmetic Act (FDCA). Some DTC AI/ML apps do not meet the definition of a device as they are intended "for maintaining or encouraging a healthy lifestyle" and are "unrelated to the diagnosis, cure, mitigation, prevention, or treatment of a disease or condition" (Section 520(o)(1)(B) of the FDCA). For example, a DTC AI/ML app that records and monitors food consumption to manage weight is not considered a medical device. However, some DTC AI/ML apps meet the definition of a device, such as a general wellness product that notifies users to keep skin out of direct sunlight when the ultraviolet index is too high to mitigate the risk of skin cancer[4].

If a DTC AI/ML app meets the definition of a device (that is, it is a DTC medical AI/ML app), the FDA may still exercise enforcement discretion. The overarching principle of the FDA's review is to consider the degree of risk the app poses to individual user safety if it were not to function as intended[5]. At present, the agency does not intend to enforce compliance with the regulatory requirements for DTC medical AI/ML apps that pose only a low risk to the public[5]. Examples include DTC AI/ML apps that may meet the definition of a medical device and give motivational guidance to smokers trying to quit[5].

If a DTC medical AI/ML app poses a moderate to high risk to individual user safety, then the FDA intends to enforce compliance with the regulatory requirements[5]. Examples of such devices include Apple's electrocardiogram (ECG) and AFib features that have received marketing authorization by the FDA as Class II medical devices[2,6].

Social cost of widespread use of DTC medical AI/ML apps

While regulators such as the FDA focus primarily on the risks to a user's safety, marketing directly to consumers presents unique concerns that call for particular care.

To begin, let us distinguish two important concepts. First, we have the technical performance of an AI/ML system as given by, say, its specificity/sensitivity (or equivalently, given a prevalence level, its false positive and false negative error rates). Such metrics are often used to describe a system's accuracy. Second, we have the beliefs/judgements of individual consumers that are formed in part by taking into account the information provided by the AI/ML system. These too can be described as false positive (believing one has a disease that is absent) or false negative (failing to believe one has a disease that is present). We can articulate the same idea probabilistically: high confidence in having a disease that is absent is a mistake in the false positive error direction, whereas low confidence in a disease that is present is a mistake in the false negative error direction[7]. But the veracity of a belief is not the same as the accuracy of an AI/ML system.

These two can substantially come apart when an AI/ML system is introduced into actual practice settings. In Box 1, we provide an example of how the divergence of personal belief and diagnostic

accuracy can create major problems for DTC medical AI/ML apps. As we illustrate, even with extremely high sensitivity/specificity, the frequency of false positive judgements among consumers may still be very high. This is in part because consumers are not medical experts and are probably unable to place DTC medical AI/ML app diagnoses in context. Without reliable prior information about the likelihood of a disease's occurrence, they are liable to identify their own posterior belief—that is, the probability that they have the disease—with the app's diagnosis. This mistake—base-rate neglect—is perhaps the most common fallacy in medical decision-making. Doctors themselves often fall into the trap, but non-expert consumers will be particularly prone to it[8,9]. This discrepancy between personal belief and diagnostic accuracy can be further exacerbated by the way the system's conclusion is expressed and the gravity of the disease. For example, if patients are given a conclusion with very precise language (for example, '83.7% chance of disease'), they may overestimate its reliability. Likewise, if the disease is very serious, even a low-probability risk can be quite daunting.

The problem of exaggerated false positive judgements may be further compounded by two factors. First, DTC medical AI/ML apps are targeted to a large, generally young, target demographic. Consider, for example, the Apple Watch irregular rhythm notification feature's user base. Within such a heterogeneous and overall healthy younger population, diseases such as AFib will be very rare. This has the effect of deflating the base rate of disease, which increases the probability of a false positive judgement. Second, DTC medical AI/ML apps are marketed for quick and inexpensive use; and using them many times is effortless and instantaneous. For example, apps designed to detect skin cancer from mobile phone images of lesions on a person's body: consumers can retest the same spot at an alternative distance, or under different lighting, for example (see Box 1). This further increases the probability of false positive judgements by increasing the overall number of tests performed. Thus, an app's ability to detect the presence of disease can be excellent, while the probability that someone actually has a disease on receiving a positive diagnosis may still be very low. Failing to recognize this can lead to unreasonably high assessments by users of the likelihood that they have a disease after receiving a positive diagnosis.

More generally, this is because for relatively rare diseases, whereas most negatives are indeed true negatives, most positives are false positives (see figure in Box 1). This is a well-known point in Bayesian decision theory, but it tends to be overlooked in the DTC medical AI/ML app discussion, and indeed, it is not something ordinary consumers can be expected to understand. For example, in a well-known study, patients in US clinics were asked to interpret diagnostic tests for several well-known diseases, including HIV and strep throat. They were asked to compare the probability of a positive test when the disease is present (sensitivity) against the probability that a person has the disease if she tests positive (posterior belief/positive predictive value). Most patients estimated these probabilities to be nearly the same[10]. This implies that even though a regulator such as the FDA can determine the optimal accuracy threshold (and false positive versus false negative ratio) required for approval, the observed proportion of mistakes under a system's actual use can be quite different, which can result in unexpected costs for the system as a whole.

Finally, the typical consumer is risk and/or loss averse, so the cost of a positive diagnosis will loom larger in their mind than the benefit of a negative one, a well-known effect in behavioural economics and decision-making[11]. Indeed, in the context of DTC genetic counselling, we have seen that when people learn that they are susceptible to a certain disease, they routinely overestimate their risk of contracting that disease[12]. In other words, people unduly amplify low-probability adverse outcomes in their personal decision-making. Note that while the result in this example is similar to base-rate neglect, the point here is slightly different: risk and loss aversion can further increase base-rate neglect when the outcome being estimated is harmful (compare, instead, estimating a stranger's health outcomes). This further exacerbates the discrepancy between a positive diagnosis and the veracity of a person's belief that they have a certain disease. This issue has been recognized for other medical diagnostics, such as HIV self-testing, where researchers have called for confirmatory lab testing, which would reduce the effect of base-rate neglect, and for pre- and post-test counselling, to help patients interpret their results[13]. The discrepancy between the assessment and the personal belief may be made worse by the fact that AI/ML DTC apps, as opposed to ordinary DTC apps, typically produce verdicts using opaque algorithms that even informed users are unable to interpret and thus appropriately weigh against their prior body of evidence.

Note that from the standard regulatory perspective of risk to an individual's safety, the typical worry when it comes to DTC medical AI/ML apps would probably be false negative, not false positive, judgements. That is, if a diagnostic device fails to identify a present disease, it may lull the patient into a false sense of security, thereby leaving the condition untreated and letting it deteriorate over time. This is indeed a risk, capable of generating both individual and social costs—for example, if people delay treatment it may be more costly to treat them later.

However, as we begin to see, a particularly important risk we face when placing DTC medical AI/ML apps into the hands of consumers is from false positive judgements: when someone falsely believes they have a disease, they will probably schedule a doctor's visit to confirm the diagnosis. They may also see one or more specialists or request unnecessary prophylactic medication (for example, blood thinners, angiotensin converting enzyme inhibitors and so on). While each of these steps is trivial at an individual level, collectively, they can generate a substantial misuse of scarce medical resources.

Accordingly, while regulators typically focus on a device's technical specifications (for example, specificity/sensitivity), what also matters for DTC medical AI/ML apps is the veracity of users' beliefs and their subsequent actions. Even though a positive diagnosis from a very accurate system would not lead a rational Bayesian to a high probability that they actually have a disease, the typical user is unlikely to be a perfectly rational Bayesian and may thereby consume disproportionate healthcare resources. When we reframe the problem this way, it becomes clear that more attention should be paid to how consumers interact with medical AI/ML apps and the negative externalities on healthcare infrastructure generated by their collective behaviour. We discuss some possible approaches below.

Guidelines for a more effective DTC AI/ML regulatory regime

The ideal form of regulatory review would take as its goal improving the veracity of DTC medical AI/ML apps in the hands of consumers rather than focusing on their performance in a formal algorithmic testing environment (that is, the system's accuracy in classification tasks with known labels). Moreover, for the mistaken judgements that inevitably remain, it would be ideal if the regulatory system made device makers internalize the social costs of these technologies.

We say 'ideal' because we recognize that many regulators may not be well positioned to achieve this. First, they may lack the expertise to measure the potential social costs and to determine what changes in design will reduce them. Second, the social costs may be hard to estimate in a pre-marketing review and regulators may find it difficult to impose post-market surveillance or to force recalls once the horse has left the barn. Third, regulators may face legal limits on their authority in these areas—the US FDA, for example, takes it as one of its prime directives that it does not regulate the practice of medicine. Thus, while many regulators will not be able to

embody the ideal, the perfect should not be the enemy of the good. Accordingly, we believe there are several steps regulators can take to move in the right direction.

First, regulators should encourage device makers to adopt a 'system view'[14] and focus on observing through clinical trials or field research-based human factors testing how consumers actually interact with DTC medical AI/ML apps rather than unduly relying on reported sensitivity/specificity rates. Instead of focusing on a technology's performance alone, regulators should pay attention to the combined human/AI environment. As a helpful analogy, the FDA sometimes requires actual use and label comprehension studies for over-the-counter drugs[15]. This would help us understand more directly what we care about: how people incorporate diagnostic information into their belief system, and what they do with it in practice.

Second, regulators should consider requiring manufacturers to pair some DTC medical AI/ML apps with virtual doctor appointments. This would mitigate the risk of erroneous diagnoses early. Such pairing would, for example, be particularly useful for skin health or heart function screening apps while it may be less ideal for more generalized and ubiquitous conditions such as obesity. For regulators with this power, or who receive appropriate statutory authority, that could involve requiring the device maker to bear some of the costs of involving physicians in the product's life cycle as a condition of marketing authorization. Doing so would serve to internalize the costs that might otherwise be borne by the public through insurance pools or medical tax expenditures. While this might appear radical at first, there are precedents for it in the medical community. For example, in Singapore, the Ministry of Health recently launched a regulatory sandbox initiative for innovating medical services through early partnerships with industry[16]. As part of this initiative, the government is expanding physician-driven telemedicine services that provide direct clinical care, including diagnoses. Such initiatives can be easily combined with the use of DTC medical AI/ML apps. This is in keeping with the spirit of confirmatory testing and post-test counselling that scholars have urged regulators to adopt in the context of HIV self-testing[13].

Third, and similarly, regulators should consider stratifying the general mobile health market so that some DTC medical AI/ML apps are not available to the general public. One can envision a system where a function such as Apple's ECG monitor could be activated only after a doctor's prescription. For example, under Germany's recent Digital Healthcare Act, insurance coverage for certain medical AI/ML apps is explicitly conditioned on their use with a physician's prescription or an insurer's approval[17]. Our suggestion could be executed through an activation code that only one's physician can provide, much in the way that proprietary software can only be activated with a unique code provided by the manufacturer. Indeed, the system could be further streamlined by generating a QR code so that the patient would simply scan the doctor's prescription with her phone and activate the ECG functionality.

While this may arguably limit consumer autonomy to some extent, we believe that in light of the risks described above, a doctor's prescription for some DTC medical AI/ML apps would be justified by the expertise medical professionals can bring to shape the population affected by such apps. In other words, if doctors prescribe a DTC medical AI/ML app only to consumers at higher risk of the relevant disease, this can curtail the magnitude of mistaken judgements, as illustrated in Box 1. For example, if the underlying prevalence of disease in the relevant subpopulation is 10% (instead of 1%, as in Box 1), with all other details remaining the same, then the probability that a patient has the disease, given a positive test, increases from 16.7% to 68.75%. For this reason, we recommend to involve doctors, rather than both doctors and insurers, as is the case under Germany's Digital Healthcare Act. In either case, stratifying the market would strike a compromise between limiting apps to physicians and targeting consumers at large—the app would remain in the hands of consumers provided it is prescribed (Box 2).

Box 2 | Summary of recommendations

To better manage the risk of DTC medical AI/ML apps, regulators should require (or encourage) device makers to:

- Adopt a 'system view' and conduct clinical studies or field research to better understand how consumers actually behave with the devices in their hands.
- Pair DTC medical AI/ML apps with virtual doctor screenings, perhaps through regulatory sandbox initiatives with partial cost being internalized by the device maker.
- Stratify the market so that DTC medical AI/ML apps can be activated only after a doctor has prescribed their use.

Precision medicine in the form of DTC medical AI/ML apps can bring tremendous benefits. But its costs are often understated because they accumulate from seemingly harmless individual behaviour. In this Perspective, we have tried to highlight these hidden social costs, illustrate how surprisingly prevalent they may be due to the underlying behavioural factors giving rise to them and offer effective guidance for regulators to mitigate them.

Received: 19 October 2020; Accepted: 9 March 2021;
Published online: 20 April 2021

References

1. *Market Research Report* (Fortune Business Insights, 2020); https://www.fortunebusinessinsights.com/mhealth-apps-market-102020
2. *Marketing Authorization for Irregular Rhythm Notification Feature* DEN180042 (FDA, 2018); https://www.accessdata.fda.gov/cdrh_docs/pdf18/DEN180042.pdf
3. Outterson, K. et al. Repairing the broken market for antibiotic innovation. *Health Aff.* **34**, 277–285 (2015).
4. *General Wellness: Policy for Low Risk Devices* (FDA, 2019).
5. *Policy for Device Software Functions and Mobile Medical Applications* (FDA, 2019).
6. *Marketing Authorization for ECG App* DEN180044 (FDA, 2018); https://www.accessdata.fda.gov/cdrh_docs/pdf18/DEN180044.pdf
7. Babic, B. A theory of epistemic risk. *Phil. Sci.* **86**, 522–550 (2019).
8. Gigerenzer, G. et al. Helping doctors and patients make sense of health statistics. *Psychol. Sci. Public Interest* **8**, 53–96 (2007).
9. Casscells, W., Schoenberger, A. & Graboys, T. B. Interpretations by physicians of clinical laboratory results. *N. Engl. J. Med.* **299**, 99–1001 (1978).
10. Hamm, R. M. & Smith, S. L. The accuracy of patients' judgments of disease probability and test sensitivity and specificity. *J. Fam. Pract.* **47**, 44–52 (1998).
11. Rosen, A. B. et al. Variations in risk attitude across race, gender, and education. *Med. Decis. Making* **23**, 511–517 (2003).
12. Ransohoff, D. F. & Khoury, M. J. Personal genomics: information can be harmful. *Eur. J. Clin. Invest.* **40**, 64–68 (2010).
13. Stevens, D. R. et al. A global review of HIV self-testing: themes and implications. *AIDS Behav.* **22**, 497–512 (2018).
14. Gerke, S., Babic, B., Evgeniou, T. & Cohen, I. G. The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. *npj Digit. Med.* **3**, 53 (2020).
15. *Guidance for Industry: Label Comprehension Studies for Nonprescription Drug Products* (FDA, 2010).
16. *Licencing Experimentation and Adaptation Programme (LEAP) - A MOH Regulatory Sandbox* (Singapore MOH, 2019); https://www.moh.gov.sg/home/our-healthcare-system/licensing-experimentation-and-adaptation-programme-(leap)---a-moh-regulatory-sandbox
17. Gerke, S., Stern, A. D. & Minssen, T. Germany's digital health reforms in the COVID-19 era: lessons and opportunities for other countries. *npj Digi. Med.* **3**, 94 (2020).

Acknowledgements

S.G. and I.G.C. were supported by a grant from the Collaborative Research Program for Biomedical Innovation Law, a scientifically independent collaborative research program supported by a Novo Nordisk Foundation grant (NNF17SA0027784).

Author contributions

All authors contributed equally to the analysis and drafting of the paper.

Competing interests

I.G.C. served as a bioethics consultant for Otsuka on their Abilify MyCite product. I.G.C. is a member of the Illumina Ethics Advisory Board. The other authors declare no competing interests.

Additional information

Correspondence should be addressed to I.G.C.

Peer review information *Nature Machine Intelligence* thanks Jan Brauner, Geoff Tison and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

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