Doctopic: Analysis and Interpretation
29_05 Digital medicine
S0140-6736(21)01152-1

Digital medicine

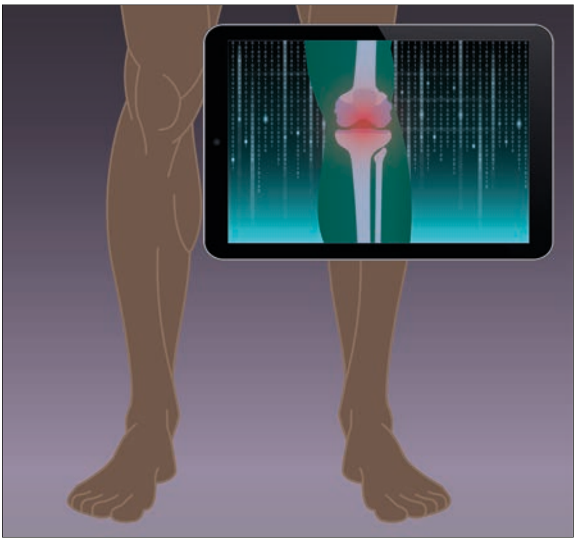
Artificial intelligence, bias, and patients' perspectives

For more on **Digital medicine** see **Comment** *Lancet* 2016; **388:** 740 and **Perspectives** *Lancet* 2021; **397:** 1254

Some of the most exciting applications of machine learning to medicine involve the kinds of data that cannot be analysed with traditional statistical models: medical imaging, waveforms, and videos. Researchers are training algorithms to take in these complex signals, and output a doctor's interpretation—eg, given a particular retinal fundus photograph, would an ophthalmologist identify diabetic retinopathy? Algorithms based on datasets that pair images or waveforms with "labels" assigned by a doctor have the potential to drive improvements in efficiency and diagnostic accuracy. However, the strength of this approach can also be its weakness: by matching the performance of doctors, algorithms will also incorporate their inherent limitations.

Take the example of pain. Decades of research have shown poor correlation between so-called objective findings on imaging and patients' reports of pain. There is much research on how stress and other contextual factors can mediate experiences of pain. This work is often cited in relation to the "pain gap" between Black and White patients, above and beyond radiographic severity, and the part played by factors such as socioeconomic determinants and the way pain is managed in an unjust medical system. But what do we mean by radiographic severity? One commonly used grading system for defining knee osteoarthritis, for example, is largely based on data from coal miners in the UK during the 1950s. Tellingly, the original reports do not even mention the sex or ethnicity of the cohort, presumably because all the participants were White men.

Could algorithms that use imaging data overcome some of the limitations of such a grading system? The standard machine learning approach, however, will falter for such a task. By training an algorithm to predict what a radiologist would say about the image—eg, its Kellgren and Lawrence grade—we are also constraining it. We are preventing the algorithm from seeing past the doctor's limitations and biases. The performance of artificial intelligence (AI) algorithms has typically been compared with doctors' performance, but what about patients' experiences?



Faith Hark/Scripps Research Translational Institute

Research by one of us (ZO), with colleagues, has produced an algorithm trained to predict the knee pain reported by the patient, rather than the x-ray interpretation of the doctor. This approach explained more of all patients' pain compared with standard measures of radiographic severity, and its explanatory power for pain was particularly useful for underserved groups of patients, such as Black patients or patients with low income and low education. Relative to standard measures of radiographic severity, the algorithm's better performance was rooted in the diversity of the patient population from which it learned. The algorithm could also potentially be useful in relation to addressing disparities in access to knee replacement surgeries. In this study, we replicated clinical guidelines for eligibility for knee replacement surgeries but replaced the radiologist's judgment with the algorithm's severity score. Doing so doubled the proportion of Black patients' knees that were eligible for knee replacement.

There are two key lessons here for medical applications of AI. First, algorithms can scale up racial bias—or they can fight against it. Which one depends on the technical choices we make and the values we instantiate when training our algorithms. Do we teach them to listen to the doctor or the patient? The closer we can align algorithms with patient experiences and outcomes, rather than the way patients are treated by the health-care system, the more algorithms will work to redress inequities rather than reinforce them.

Second, by training algorithms to predict labels related to clinical outcomes, rather than doctors' judgments, we can start to push forward a new kind of clinical science. For example, by grounding patient reports of pain in objective radiographic features, we might develop a more comprehensive understanding of what causes pain. By not being doctor-centric and incorporating the patient's perspective, machine learning has added potential for unravelling important mysteries of medicine.

Ziad Obermeyer, Eric J Topol
UC Berkeley School of Public Health, Berkeley, CA 94720, USA (ZO);
Scripps Research Translational Institute, La Jolla, CA, USA (EJT)
zobermeyer@berkeley.edu

Further reading

Anderson KO, Green CR, Payne R. Racial and ethnic disparities in pain: causes and consequences of unequal care. *J Pain* 2009; **10:** 1187–204

Allen KD, Oddone EZ, Coffman CJ, et al. Racial differences in osteoarthritis pain and function: potential explanatory factors. *Osteoarthr Cartil* 2010; **18:** 160–67

Mullainathan S, Obermeyer Z. On the inequity of predicting A while hoping for B. *AEA Papers Proc* 2021; **3:** 37–42

Pierson E, Cutler DM, Leskovec J, Mullainathan S, Obermeyer Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nat Med* 2021; **27:** 136–40

Poleshuck EL, Green CR. Socioeconomic disadvantage and pain. *Pain* 2008; **136:** 235

ZO is supported by grants from the US National Institutes of Health/National Institute on Aging, Schmidt Futures, and the Gordon and Betty Moore Foundation and he reports equity in Dandelion Health and LookDeep Health. EJT is supported by the US National Institutes of Health/National Center for Advancing Translational Sciences grant UL1TR001114.