Published in final edited form as:

N Engl J Med. 2018 March 15; 378(11): 981–983. doi:10.1056/NEJMp1714229.

Implementing Machine Learning in Health Care — Addressing Ethical Challenges

Danton S. Char, M.D., Nigam H. Shah, M.B., B.S., Ph.D., and David Magnus, Ph.D.

Department of Anesthesiology, Division of Pediatric Cardiac Anesthesia (D.S.C.), the Center for Biomedical Ethics (D.S.C., D.M.), and the Center for Biomedical Informatics Research (N.S.), Stanford University School of Medicine, Stanford, CA

The incorporation of machine learning into clinical medicine holds promise for substantially improving health care delivery. Private companies are rushing to build machine learning into medical decision making, pursuing both tools that support physicians and algorithms designed to function independently of them. Physician-researchers are predicting that familiarity with machine-learning tools for analyzing big data will be a fundamental requirement for the next generation of physicians and that algorithms might soon rival or replace physicians in fields that involve close scrutiny of images, such as radiology and anatomical pathology.¹

However, consideration of the ethical challenges inherent in implementing machine learning in health care is warranted if the benefits are to be realized. Some ethical challenges are straightforward and need to be guarded against, such as concerns that algorithms may mirror human biases in decision making. Others, such as the possibility for algorithms to become the repository of the collective medical mind, have less obvious risks but raise broader ethical concerns.

Algorithms introduced in non-medical fields have already been shown to make problematic decisions that reflect biases inherent in the data used to train them. For example, programs designed to aid judges in sentencing by predicting an offender's risk of recidivism have shown an unnerving propensity for racial discrimination.²

It's possible that similar racial biases could inadvertently be built into health care algorithms. Health care delivery already varies by race. An algorithm designed to predict outcomes from genetic findings will be biased if there have been few (or no) genetic studies in certain populations. For example, attempts to use data from the Framingham Heart Study to predict the risk of cardiovascular events in nonwhite populations have led to biased results, with both overestimations and underestimations of risk.³

Subtle discrimination inherent in health care delivery may be harder to anticipate; as a result, it may be more difficult to prevent an algorithm from learning and incorporating this type of bias. Clinicians already consider neurodevelopmental delays and certain genetic findings when rationing scarce resources, such as organs for transplantation. Such considerations

may lead to self-fulfilling prophecies: if clinicians always withdraw care in patients with certain findings (extreme prematurity or a brain injury, for example), machine-learning systems may conclude that such findings are always fatal. On the other hand, it's also possible that machine learning, when properly deployed, could help resolve disparities in health care delivery if algorithms could be built to compensate for known biases or identify areas of needed research.

The intent behind the design of machine-learning systems also needs to be considered. Algorithms can be designed to perform in unethical ways. A recent high-profile example is Uber's software tool Greyball, which was designed to predict which ride hikers might be undercover law-enforcement officers, thereby allowing the company to identify and circumvent local regulations. More complex deception might involve algorithms designed to cheat, such as Volkswagen's algorithm that allowed vehicles to pass emissions tests by reducing their nitrogen oxide emissions when they were being tested.

Private-sector designers who create machine-learning systems for clinical use could be subject to similar temptations. Given the growing importance of quality indicators for public evaluations and determining reimbursement rates, there may be a temptation to teach machine-learning systems to guide users toward clinical actions that would improve quality metrics but not necessarily reflect better care. Such systems might also be able to skew the data provided for public evaluation or identify when they're being reviewed by potential hospital regulators. Clinical decision-support systems could also be programmed in ways that would generate increased profits for their designers or purchasers (such as by recommending drugs, tests, or devices in which they hold a stake or by altering referral patterns) without clinical users being aware of it.

Potential differences between the intent behind the design of machine-learning systems and the goals of users (the care team and patients) may create ethical strain. In the U.S. health care system, there is perpetual tension between the goals of improving health and generating profit. This tension needs to be acknowledged and addressed in the implementation of machine learning, since the builders and purchasers of machine-learning systems are unlikely to be the same people delivering bedside care.

The use of machine learning in complicated care practices will require ongoing consideration, since the correct diagnosis in a particular case and what constitutes best practice can be controversial. Prematurely incorporating a particular diagnosis or practice approach into an algorithm may imply a legitimacy that is unsubstantiated by data.

As clinical medicine moves progressively toward a shift-based model, the number of clinicians who have followed diseases from their presentation through their ultimate outcome is decreasing. This trend underscores the opportunity for machine learning and approaches based on artificial intelligence in health care — but it could also give such tools unintended power and authority. The collective medical mind is becoming the combination of published literature and the data captured in health care systems, as opposed to individual clinical experience. Although this shift presents exciting opportunities to learn from aggregate data,⁴ the electronic collective memory may take on an authority that was perhaps

never intended. Clinicians may turn to machine learning for diagnosis and advice about treatments — not simply as a support tool. If that happens, machine-learning tools will become important actors in the therapeutic relationship and will need to be bound by the core ethical principles, such as beneficence and respect for patients, that have guided clinicians.

Ethical guidelines can be created to catch up with the age of machine learning and artificial intelligence that is already upon us. Physicians who use machine-learning systems can become more educated about their construction, the data sets they are built on, and their limitations. Remaining ignorant about the construction of machine-learning systems or allowing them to be constructed as black boxes could lead to ethically problematic outcomes.

More broadly, the introduction of algorithms in the provision of medical care raises questions about the nature of the relationship between physicians and patients. At its core, clinical medicine has been a compact — the promise of a fiduciary relationship between a patient and a physician. As the central relationship in clinical medicine becomes that between a patient and a health care system, the meaning of fiduciary obligation has become strained and notions of personal responsibility have been lost.

Medical ethics will need to adapt. With the addition of machine-learning systems to this changing landscape, it becomes increasingly unclear which parties are involved in a fiduciary compact, even if physicians are still the ones providing care. The idea of confidentiality, once a cornerstone of Hippocratic ethics, was long ago described as “decrepit.”⁵ In the era of electronic medical records, the traditional understanding of confidentiality requires that a physician withhold information from the medical record in order to truly keep it confidential. Once machine-learning–based decision support is integrated into clinical care, withholding information from electronic records will become increasingly difficult, since patients whose data aren’t recorded can’t benefit from machine-learning analyses. The implementation of machine-learning systems will therefore require a reimagining of confidentiality and other core tenets of professional ethics. What’s more, a learning health care system will have agency, which will also need to be factored into ethical considerations surrounding patient care.

We believe that challenges such as the potential for bias and questions about the fiduciary relationship between patients and machine-learning systems will have to be addressed as soon as possible. Machine-learning systems could be built to reflect the ethical standards that have guided other actors in health care — and could be held to those standards. A key step will be determining how to ensure that they are — whether by means of policy enactment, programming approaches, task-force work, or a combination of these strategies.

References

1. Obermeyer Z, Emanuel EJ. Predicting the future — big data, machine learning, and clinical medicine. *N Engl J Med*. 2016; 375:1216–9. [PubMed: 27682033]
2. Angwin, J., Larson, J., Mattu, S., Kirchner, L. Machine bias. ProPublica. May 23, 2016. (<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>)

3. Gijsberts CM, Groenewegen KA, Hoefer IE, et al. Race/ethnic differences in the associations of the Framingham risk factors with carotid IMT and cardiovascular events. *PLoS One*. 2015; 10(7):e0132321. [PubMed: 26134404]
4. Longhurst CA, Harrington RA, Shah NH. A 'green button' for using aggregate patient data at the point of care. *Health Aff (Millwood)*. 2014; 33:1229–35. [PubMed: 25006150]
5. Siegler M. Confidentiality in medicine — a decrepit concept. *N Engl J Med*. 1982; 307:1518–21. [PubMed: 7144818]