HHS Public Access

Author manuscript

N Engl J Med. Author manuscript; available in PMC 2018 May 16.

Published in final edited form as:

N Engl J Med. 2017 June 29; 376(26): 2507–2509. doi:10.1056/NEJMp1702071.

Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations

Jonathan H. Chen, M.D., Ph.D. and **Steven M. Asch, M.D., M.P.H.**

Department of Medicine, Stanford University, Stanford (J.H.C., S.M.A.), and the Center for Innovation to Implementation (Ci2i), Veteran Affairs Palo Alto Health Care System, Palo Alto (S.M.A.) — both in California

Big data, we have all heard, promise to transform health care with the widespread capture of electronic health records and high-volume data streams from sources ranging from insurance claims and registries to personal genomics and biosensors.[1] Artificial-intelligence and machine-learning predictive algorithms, which can already automatically drive cars, recognize spoken language, and detect credit card fraud, are the keys to unlocking the data that can precisely inform real-time decisions. But in the "hype cycle" of emerging technologies, machine learning now rides atop the "peak of inflated expectations."[2]

Prediction is not new to medicine. From risk scores to guide anticoagulation (CHADS2) and the use of cholesterol medications (ASCVD) to risk stratification of patients in the intensive care unit (APACHE), data-driven clinical predictions are routine in medical practice. In combination with modern machine learning, clinical data sources enable us to rapidly generate prediction models for thousands of similar clinical questions. From early-warning systems for sepsis to superhuman imaging diagnostics, the potential applicability of these approaches is substantial.

Yet there are problems with real-world data sources. Whereas conventional approaches are largely based on data from cohorts that are carefully constructed to mitigate bias, emerging data sources are typically less structured, since they were designed to serve a different purpose (e.g., clinical care and billing). Issues ranging from patient self-selection to confounding by indication to inconsistent availability of outcome data can result in inadvertent bias, and even racial profiling, in machine predictions. Awareness of such challenges may keep the hype from outpacing the hope for how data analytics can improve medical decision making.

Machine-learning methods are particularly suited to predictions based on existing data, but precise predictions about the distant future are often fundamentally impossible. Prognosis models for *HER2*-negative breast cancer had to be inverted in the face of targeted therapies, and the predicted efficacy of influenza vaccination varies with disease prevalence and community immunization rates. Given that the practice of medicine is constantly evolving in

Disclosure forms provided by the authors are available at NEJM.org.

response to new technology, epidemiology, and social phenomena, we will always be chasing a moving target.

The rise and fall of Google Flu remind us that forecasting an annual event on the basis of 1 year of data is effectively using only a single data point and thus runs into fundamental time-series problems.[3] Yet if the future will not necessarily resemble the past, simply accumulating mass data over time has diminishing returns. Research into decision-support algorithms that automatically learn inpatient medical practice patterns from electronic health records reveals that accumulating multiple years of historical data is worse than simply using the most recent year of data. When our goal is learning how medicine should be practiced in the future, the relevance of clinical data decays with an effective “half-life” of about 4 months.[4] To assess the usefulness of prediction models, we must evaluate them not on their ability to recapitulate historical trends, but instead on their accuracy in predicting future events.

Although machine-learning algorithms can improve the accuracy of prediction over the use of conventional regression models by capturing complex, nonlinear relationships in the data, no amount of algorithmic finesse or computing power can squeeze out information that is not present. That's why clinical data alone have relatively limited predictive power for hospital readmissions that may have more to do with social determinants of health.

The apparent solution is to pile on greater varieties of data, including anything from sociodemographics to personal genomics to mobile-sensor readouts to a patient's credit history and Web-browsing logs. Incorporating the correct data stream can substantially improve predictions, but even with a deterministic (non-random) process, chaos theory explains why even simple nonlinear systems cannot be precisely predicted into the distant future. The so-called butterfly effect refers to the future's extreme sensitivity to initial conditions. Tiny variations, which seem dismissible as trivial rounding errors in measurements, can accumulate into massively different future events. Identical twins with the same observable demographic characteristics, lifestyle, medical care, and genetics necessarily generate the same predictions — but can still end up with completely different real outcomes.

Though no method can precisely predict the date you will die, for example, that level of precision is generally not necessary for predictions to be useful. By reframing complex phenomena in terms of limited multiple-choice questions (e.g., Will you have a heart attack within 10 years? Are you more or less likely than average to end up back in the hospital within 30 days?), predictive algorithms can operate as diagnostic screening tests to stratify patient populations by risk and inform discrete decision making.

Research continues to improve the accuracy of clinical predictions, but even a perfectly calibrated prediction model may not translate into better clinical care. An accurate prediction of a patient outcome does not tell us what to do if we want to change that outcome — in fact, we cannot even assume that it's possible to change the predicted outcomes.

Machine-learning approaches are powered by identification of strong, but theory-free, associations in the data. Confounding makes it a substantial leap in causal inference to

identify modifiable factors that will actually alter outcomes. It is true, for instance, that palliative care consults and norepinephrine infusions are highly predictive of patient death, but it would be irrational to conclude that stopping either will reduce mortality. Models accurately predict that a patient with heart failure, coronary artery disease, and renal failure is at high risk for postsurgical complications, but they offer no opportunity for reducing that risk (other than forgoing the surgery). Moreover, many such predictions are "highly accurate" mainly for cases whose likely outcome is already obvious to practicing clinicians. The last mile of clinical implementation thus ends up being the far more critical task of predicting events early enough for a relevant intervention to influence care decisions and outcomes.[5]

With machine learning situated at the peak of inflated expectations, we can soften a subsequent crash into a "trough of disillusionment"[2] by fostering a stronger appreciation of the technology's capabilities and limitations. Before we hold computerized systems (or humans) up against an idealized and unrealizable standard of perfection, let our benchmark be the real-world standards of care whereby doctors grossly misestimate the positive predictive value of screening tests for rare diagnoses, routinely overestimate patient life expectancy by a factor of 3, and deliver care of widely varied intensity in the last 6 months of life.

Although predictive algorithms cannot eliminate medical uncertainty, they already improve allocation of scarce health care resources, helping to avert hospitalization for patients with low-risk pulmonary embolisms (PESI) and fairly prioritizing patients for liver transplantation by means of MELD scores. Early-warning systems that once would have taken years to create can now be rapidly developed and optimized from real-world data, just as deep-learning neural networks routinely yield state-of-the-art image-recognition capabilities previously thought to be impossible.

Whether such artificial-intelligence systems are "smarter" than human practitioners makes for a stimulating debate — but is largely irrelevant. Combining machine-learning software with the best human clinician "hardware" will permit delivery of care that outperforms what either can do alone. Let's move past the hype cycle and on to the "slope of en-lightenment,"[2] where we use every information and data resource to consistently improve our collective health.

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. Gartner, Inc. Stamford, CT: Gartner; 2016. identifies three key trends that organizations must track to gain competitive advantage in its 2016 hype cycle for emerging technologies. http://www.gartner.com/newsroom/id/3412017
3. Lazer D, Kennedy R, King G, Vespignani A. Big data — the parable of Google Flu: traps in big data analysis. Science. 2014; 343:1203–5. [PubMed: 24626916]
4. Chen JH, Alagappan M, Goldstein MK, Asch SM, Altman RB. Decaying relevance of clinical data towards future decisions in data-driven inpatient clinical order sets. Int J Med Inform. 2017; 102:71–9. [PubMed: 28495350]

5. Escobar GJ, Turk BJ, Ragins A, et al. Piloting electronic medical record-based early detection of inpatient deterioration in community hospitals. J Hosp Med. 2016; 11(1):S18–S24. [PubMed: 27805795]