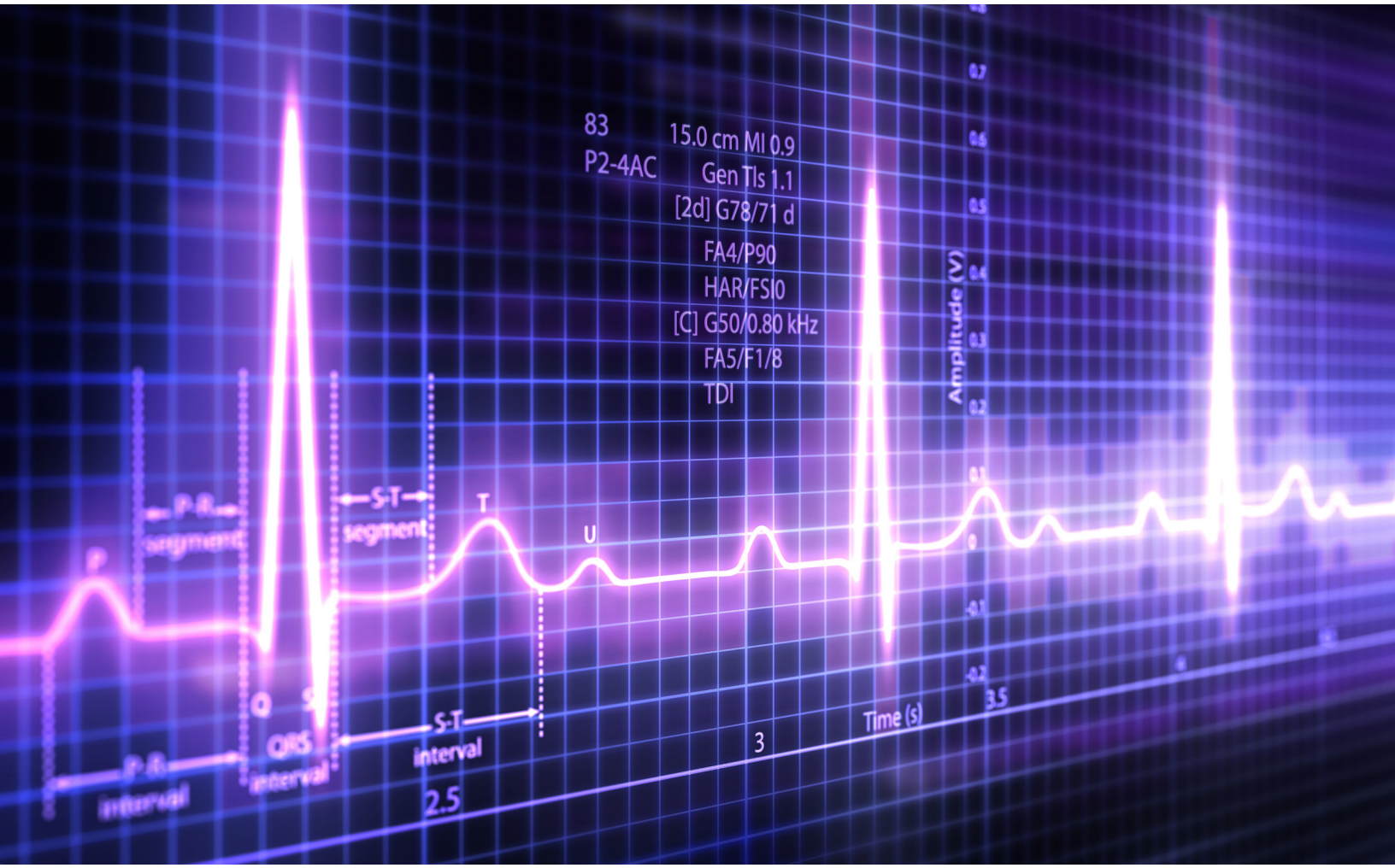
Machine learning and therapeutics 2.0: Avoiding hype, realizing potential

Six levers can help healthcare and pharma players achieve better outcomes when using machine learning.

David Champagne, Sastry Chilukuri, Martha Imprialou, Saif Rathore, and Jordan VanLare



The US healthcare system generates approximately one trillion gigabytes of data annually.[1] These prodigious quantities of data have been accompanied by an increase in cheap, large-scale computing power. Together, they raise the possibility that artificial intelligence—and machine learning, in particular—can generate insights both to improve the discovery of new therapeutics and to make the delivery of current ones more effective. Although we have seen early advances in therapeutic development, diagnostics, and treatment recommendations,[2] there have been setbacks and suggestions that machine learning has reached peak hype.[3] Some criticism is merited, but machine learning continues to offer transformative potential for health and healthcare (see sidebar, "About machine learning").

Unlocking machine learning's full potential, however, requires recognizing and addressing issues raised to date. In this article, we discuss how pharmaceutical and healthcare companies can use machine learning more effectively to exploit its promise of spurring innovation and improving health.

Health and healthcare are different from other sectors

Early successes applying machine learning in other industries may not readily translate when attempting to scale machine learning in healthcare. That underscores a fundamental difference in applying machine learning to health and healthcare. Health—specifically, the understanding of diseases and treatments—is fundamentally different from other areas where machine learning has been used.

First, despite notable gains in the understanding of disease, the most common health conditions remain poorly understood multifactorial processes. The medical school dean's admission to her students—"Half of what we teach you will be wrong; we just don't know which half"—remains true. Second, gains from mechanistic drug development have led to the expectation that we fully understand the impact of interventions. However, the fields of drug, device, and healthcare interventions are replete with examples of interventions that worked mechanistically, improved intermediate or other proxy measures, but failed final evaluation. Finally,

About machine learning

Machine learning is a form of artificial intelligence in which algorithms learn from data, with or without explicit guidance, to improve predictions or classifications of current data. An algorithm, at its simplest, is designed to accomplish a specific task, then trained on data, and revised. The process is repeated until the algorithm achieves optimal performance in terms of fit to the training data. The machine itself generates the algorithm rather than relying on external coding to direct the algorithm's construction. The ability to ingest a broader swath of variables and to explore multiple permutations offers gains over classic approaches to modeling. Deep learning is a form of machine learning that uses multiple layers of neural networks with large quantities of data to optimize a host of algorithms for performing a specific task. Machine learning has great potential for therapeutic development and healthcare, ranging from discovery to diagnosis to decision making.

despite the volume of healthcare data generated during current practice, much information that could meaningfully support insights remains uncollected, siloed, or otherwise inaccessible to researchers. Data that are often used are primarily collected for reimbursement, and it is unlikely that administrative data will offer truly meaningful insights compared with richer clinical data. In short, health represents a distinct challenge for machine learning because of our still-limited understanding of disease, the effects of our interventions, and the lack of integrated data that can effectively capture this information at meaningful scale. Given this more challenging analytical environment, we need to be more thoughtful about how we employ machine learning in health and healthcare.

Six levers can unlock machine learning's potential

Six levers can help position machine learning to realize more of its potential in therapy discovery and healthcare:

- judicious consideration of data sufficiency and representativeness
- commitment to reproducibility, ensuring that insights withstand known challenges with replication
- prioritization of transparency and a move beyond black-box algorithms to ensure that insights are understood and trusted
- credibility, or ensuring that insights are consistent with established science and reflect the input of domain expertise
- recognition of the need to demonstrate the impact or otherwise to quantify the gain (in tangible outcomes) from the use of an algorithm compared with other approaches
- recognition that data reflect the clinical and social context in which healthcare is delivered; left unchecked, algorithms risk inadvertently perpetuating the biases of these environments

The six levers apply to the myriad ways machine learning is deployed in therapeutic development and healthcare delivery. This includes pharma and medical-technology manufacturers using machine learning to inform development, shaping clinical-trial design, and later elements of life-cycle management, such as understanding variations in responses to treatment. Similarly, the levers apply to healthcare uses of machine learning, including informing diagnosis, developing treatment algorithms, and using digital therapeutics. Examples relevant for pharmaceutical, medical-technology, and healthcare companies are included below.

1. Data quality is a critical, yet often overlooked, success factor

Statistical techniques depend on the quality of data available to generate findings. Poor-quality data will not yield meaningful insights, and no analytical method, regardless of its sophistication, can overcome shortfalls in data sufficiency, representativeness, or scale. Regrettably, machine learning is sometimes deployed with the expectation that it may overcome these shortfalls. For example, the omission of clinical-practice guidelines produced a machine learning algorithm suggesting that respiratory infections are a leading cause of chest pain, completely omitting cardiac causes.[4]

Separate but equally important is ensuring that data inputs are representative. Data drawn from selected populations run the risk of not only lacking generalizability, but also generating incorrect conclusions. An algorithm developed to help detect melanoma, for instance, should be derived from a sufficiently diverse sample; when these data aren't used or aren't available, algorithms run the risk of overlooking disease in underrepresented patient

populations.[5] Clinically important variables and endpoints should be present and robust, but this remains a challenge in many data resources.[6]

Finally, machine learning cannot overcome the analytic constraints of a small sample.[7] Here too, reliance on a small sample with few outcomes cannot be overcome with an algorithm that feeds in additional data elements.[8] Paradoxically, this approach makes the algorithm even more beholden to the idiosyncrasies of the smaller sample. In short, machine learning will not overcome limitations in data availability, quality, or generalizability. Investments in data engineering, data-set curation, use of rigorous epidemiological methods to evaluate bias, well-integrated domain expertise to ensure the validity of data, and consideration of Bayesian modeling techniques can reduce data-related challenges.

2. Results must be reproducible

Machine learning algorithms perform as well as or better than a conventional statistical approach with the data set used to develop them. There is a risk that this gain in performance and the resulting new insights may be artifacts of the sample and not indicative of true underlying causal processes. Conventional statistical approaches face the same limitation. Addressing this issue requires more than testing performance in a holdout sample but rather testing in a different population, without recalibration or retraining. Ideally, this evaluation is conducted prospectively or using unified data models that support validation between different data sources. The value in this approach is that it distinguishes between algorithm features that are robust across settings—and thereby truly insightful—and those that contribute marginally. It ensures that the elements of the algorithm are truly informative and thus likely to remain so as additional insights are generated. Most important, demonstrating consistent results across settings helps confirm the algorithm's utility.

3. Algorithms must be transparent

Perhaps the most prominent criticism of machine learning approaches is that they represent a black box and provide no clear understanding of how they generate insights. Unchecked, this raises the risk that machine learning may become a form of alchemy, in which users cannot understand why some algorithms work and others fail or the criteria for choosing between different algorithm structures.[9] Because no algorithm "understands" its inputs or outputs, key healthcare stakeholders—manufacturers, payers, providers, and regulators—are not likely to accept machine learning algorithms that lack transparency.

The absence of transparency also limits the impact of machine learning, as users don't know which part of the algorithm provided a gain over conventional approaches. Machine learning's identification of novel cause is lost. Leading groups that use machine learning recognize this shortfall and have begun to make their algorithms more transparent. Google, for instance, recently published a machine learning algorithm to evaluate retinal images, which included an evaluation of the elements of the model that produced its recommendations.[10] Greater attention to explainable artificial intelligence is a meaningful step in this direction.[11] Clarifying the elements of an algorithm—and their distinctive impact—will be increasingly important if machine learning is to overcome skepticism among healthcare stakeholders.

4. Algorithms must be credible

Machine learning algorithms must offer insights that are credible and aligned with the scientific or clinical consensus. An algorithm that fails to replicate established findings or counters the established body of evidence is more likely an indication of a methodological oversight or a data artifact than a truly novel insight. A pharmaceutical manufacturer recently described a scenario in which a machine learning algorithm concluded that reducing low-density lipoprotein cholesterol after a heart attack

was not associated with cardiac outcomes. This finding does not change 20-plus years of established clinical science, but rather speaks to nuances in the data and analytic structure. Without such a context, machine learning could conclude that cigarette lighters cause lung cancer.

This context is provided by domain-specific expertise. Its absence results in decisions that, while analytically sound, produce algorithms that are not likely to be adopted. For instance, a recent machine learning algorithm to predict cardiovascular events included "lack of data" as a key risk factor.[12] Truly unsupervised clusters may generate clinically incomprehensible combinations. These data-architecture challenges are not specific to machine learning. Ensuring that domain expertise informs the structure of an algorithm and provides feedback during its generation not only minimizes this risk but also guides the algorithm toward a more credible, and ultimately useful, result.

5. Algorithms must demonstrate impact

Machine learning algorithms use gains in performance compared to conventional statistical approaches as grounds for claims of improvement. However, this is not the correct standard. Rather, the standard should be, "Does this machine learning algorithm do something better than we do now, and in doing so, *does this materially change outcomes?*" In short, we should test the impact of a machine learning algorithm much as we would that of any other intervention.

Verily, for instance, developed an impressive algorithm to detect macular degeneration automatically, including the ability to identify lesions that providers miss. But it is unclear if this achievement produces meaningfully better outcomes than screening by physicians. Are the lesions indolent macular degeneration that results in lead-time bias or is the algorithm finding clinically important lesions that physicians overlook? A recent study suggested that machine learning realized a 1 percent increase in discrimination over a conventional statistical model built using known factors (C-statistic 0.80 and 0.79, respectively).[13] Is this gain clinically significant? Does it support the use of a more sophisticated analytical approach given the satisfactory performance of previous ones?

These questions are best answered by measuring impact. A machine learning algorithm focused on optimizing clinical trials, for example, reduced patient-enrollment times by more than 10 percent.[14] Measuring impact is likely the greatest challenge for machine learning algorithms, given that deploying an algorithm and showing its impact require additional resources. Combined approaches such as augmented intelligence, where machine learning algorithms support or enable existing processes rather than replacing them entirely, will likely show greater promise.[15]

6. Algorithms must be fair

Data is not value free but rather a product of the systems that generate and collect healthcare data. Without recognizing this context, advanced analytic approaches may inadvertently perpetuate systemic problems. Machine learning has at times struggled with this problem. For example, early photo-categorization software could not recognize visible minorities, facial-recognition software has viewed Asian subjects as blinking, some video algorithms could not detect people with darker skin tones, and automated résumé screening can adversely affect female candidates.[16]

In more subtle cases, the use of specific data may come with important trade-offs. For instance, using a patient's socioeconomic status as a proxy for external resources may improve prediction, but also inadvertently "excuse" low-quality care provided to vulnerable populations. Even common factors (such as age) should be carefully considered to avoid generating algorithms that inadvertently perpetuate

or exacerbate disparities. These risks are even more consequential in health and healthcare. To reduce these risks, machine learning approaches will need to recognize data context and use approaches that affirm and protect key social values.[17]

Machine learning is poised to transform the way we conduct pharmaceutical and healthcare analytics. To realize its full potential, machine learning approaches will need to address key criticisms, including perceptions that machine learning may be regarded a panacea for all analytic challenges. A vision of augmented intelligence in which machine learning enables other processes rather than entirely replacing them is likely to have more impact. The development of a set of principles or guidelines by stakeholders can help shape best practices. Within the fields of pharmaceuticals, medical technology, and healthcare, understanding when and how to deploy machine learning approaches will be as important as recognizing where other ones may be sufficient. The six levers outlined above provide a way to realize the full potential of machine learning in therapeutics.

[1] Travis May, "The fragmentation of health data," *Medium*, July 31, 2018, medium.com.

[2] Rob Matheson, "Artificial intelligence model learns from patient data to make cancer treatments less toxic," MIT News Office, August 9, 2018, mit.edu; Marwin Segler, "Planning chemical syntheses with deep neural networks and symbolic AI," *Nature*, March 29, 2018, nature.com; David Steiner et al., "Impact of deep learning assistance on the histopathological review of lymph nodes for metastatic breast cancer," *American Journal of Surgical Pathology*, December 2018.

[3] Kasey Panetta, "5 trends emerge in the Gartner Hype Cycle for emerging technologies, 2018," Smarter with Gartner, August 16, 2018, gartner.com.

[4] Martin Muller, "Medical applications expose current limits of AI," *Der Spiegel*, August 3, 2018,spiegel.de.

[5] Adewole S. Adamson and Avery Smith, "Machine learning and health care disparities in dermatology," *JAMA Dermatology*, November 2018, jamanetwork.com.

[6] Cao Xiao et al., "Opportunities and challenges in developing deep learning models using electronic health records: A systematic review," *Journal of the American Medical Informatics Association*, October 2018, pp.1419–28.

[7] Robert A. Harrington and Frank E. Harrell, "Data torture and dumb analyses: missteps with big data," *Medscape*, August 6, 2018, medscape.com.

[8] G. K. Reeves et al., "Incidence of breast cancer and its subtypes in relation to individual and multiple low-penetrance genetic susceptibility loci," *JAMA*, July 28, 2010; N. Risch et al., "Interaction between the serotonin transporter gene (5-HTTLPR), stressful life events, and risk of depression: a meta-analysis," *JAMA*, August 5, 2009.

[9] Argmin blog, "An addendum to alchemy," blog entry by Ali Rahimi and Ben Recht, December 11, 2017, argmin.net.

[10] Jeffrey De Fauw et al., "Clinically applicable deep learning for diagnosis and referral in retinal disease", *Nature Medicine*, 2018, nature.com.

[11] Marco Riberio et al., "Introduction to local interpretatble model-agnostic explanations," O'Reilly, April 12 2016, oreilly.com.

[12] Stephen F. Weng et al., "Can machine-learning improve cardiovascular risk prediction using routine clinical data?", *PLOS One*, April 4, 2017, journals.plos.org.

[13] Andrew J. Steele et al., "Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease" *PLOS One*, August 31, 2018, journals.plos.org.

[14] Vasant Narasimhan, "3 things that will change medicine in 2018," World Economic Forum Future of Health and Healthcare, January 24, 2018, weforum.org.

[15] David Steiner et al., "Impact of deep learning assistance on the histopathological review of lymph nodes for metastatic breast cancer," *American Journal of Surgical Pathology*, December 2018.

[16] Kate Crawford, "Artificial intelligence's white guy problem," *New York Times*, June 25, 2016, nytimes.com; Jeffrey Dastin, "Amazon scraps secret AI recruiting tool that showed bias against female candidates," Reuters, October 9, 2018, reuters.com.

[17] Filippo Raso et al., "Artificial intelligence and human rights," Berkman Klein Center for Internet & Society at Harvard University, September 2018, cyber.harvard.edu.

David Champagne is a partner in McKinsey's London office; **Sastry Chilukuri** and **Jordan VanLare** are partners in the New Jersey office; **Martha Imprialou** is a consultant for QuantumBlack, a McKinsey company; and **Saif Rathore** is a medical director in the Chicago office.

Designed by Global Editorial Services.