REVIEW ARTICLE

FRONTIERS IN MEDICINE

Machine Learning in Medicine

Alvin Rajkomar, M.D., Jeffrey Dean, Ph.D., and Isaac Kohane, M.D., Ph.D.

A 49-year-old patient notices a painless rash on his shoulder but does not seek care. Months later, his wife asks him to see a doctor, who diagnoses a seborrheic keratosis. Later, when the patient undergoes a screening colonoscopy, a nurse notices a dark macule on his shoulder and advises him to have it evaluated. One month later, the patient sees a dermatologist, who obtains a biopsy specimen of the lesion. The findings reveal a noncancerous pigmented lesion. Still concerned, the dermatologist requests a second reading of the biopsy specimen, and invasive melanoma is diagnosed. An oncologist initiates treatment with systemic chemotherapy. A physician friend asks the patient why he is not receiving immunotherapy.

From Google, Mountain View, CA (A.R., J.D.); and the Department of Biomedical Informatics, Harvard Medical School, Boston (I.K.). Address reprint requests to Dr. Kohane at the Department of Biomedical Informatics, Harvard Medical School, 10 Shattuck St., Boston, MA, 02115, or at isaac_kohane@harvard.edu.

N Engl J Med 2019;380:1347-58.
DOI: 10.1056/NEJMra1814259

What if every medical decision, whether made by an intensivist or a community health worker, was instantly reviewed by a team of relevant experts who provided guidance if the decision seemed amiss? Patients with newly diagnosed, uncomplicated hypertension would receive the medications that are known to be most effective rather than the one that is most familiar to the prescriber.[1,2] Inadvertent overdoses and errors in prescribing would be largely eliminated.[3,4] Patients with mysterious and rare ailments could be directed to renowned experts in fields related to the suspected diagnosis.[5]

Such a system seems far-fetched. There are not enough medical experts to staff it, it would take too long for experts to read through a patient's history, and concerns related to privacy laws would stop efforts before they started.[6] Yet, this is the promise of machine learning in medicine: the wisdom contained in the decisions made by nearly all clinicians and the outcomes of billions of patients should inform the care of each patient. That is, every diagnosis, management decision, and therapy should be personalized on the basis of all known information about a patient, in real time, incorporating lessons from a collective experience.

This framing emphasizes that machine learning is not just a new tool, such as a new drug or medical device. Rather, it is the fundamental technology required to meaningfully process data that exceed the capacity of the human brain to comprehend; increasingly, this overwhelming store of information pertains to both vast clinical databases and even the data generated regarding a single patient.[7]

Nearly 50 years ago, a Special Article in the *Journal* stated that computing would be "augmenting and, in some cases, largely replacing the intellectual functions of the physician."[8] Yet, in early 2019, surprisingly little in health care is driven by machine learning. Rather than report the myriad proof-of-concept models (of retrospective data) that have been tested, here we describe the core structural changes and paradigm shifts in the health care system that are necessary to enable the full promise of machine learning in medicine (see video).

A video overview of machine learning is available at NEJM.org

MACHINE LEARNING EXPLAINED

Traditionally, software engineers have distilled knowledge in the form of explicit computer code that instructs computers exactly how to process data and how to

make decisions. For example, if a patient has elevated blood pressure and is not receiving an antihypertensive medication, then a properly programmed computer can suggest treatment. These types of rules-based systems are logical and interpretable, but, as a Sounding Board article in the *Journal* in 1987 noted, the field of medicine is "so broad and complex that it is difficult, if not impossible, to capture the relevant information in rules."[9]

The key distinction between traditional approaches and machine learning is that in machine learning, a model learns from examples rather than being programmed with rules. For a given task, examples are provided in the form of inputs (called features) and outputs (called labels). For instance, digitized slides read by pathologists are converted to features (pixels of the slides) and labels (e.g., information indicating that a slide contains evidence of changes indicating cancer). Using algorithms for learning from observations, computers then determine how to perform the mapping from features to labels in order to create a model that will generalize the information such that a task can be performed correctly with new, never-seen-before inputs (e.g., pathology slides that have not yet been read by a human). This process, called supervised machine learning, is summarized in Figure 1. There are other forms of machine learning.[10] Table 1 lists examples of cases of the clinical usefulness of input-to-output mappings that are based on peer-reviewed research or simple extensions of existing machine-learning capabilities.

In applications in which predictive accuracy is critically important, the ability of a model to find statistical patterns across millions of features and examples is what enables superhuman performance. However, these patterns do not necessarily correspond to the identification of underlying biologic pathways or modifiable risk factors that underpins the development of new therapies.

There is no bright line between machine-learning models and traditional statistical models, and a recent article summarizes the relationship between the two.[36] However, sophisticated new machine-learning models (e.g., those used in "deep learning" [a class of machine-learning algorithms that use artificial neural networks that can learn extremely complex relationships between features and labels and have been shown to exceed human abilities in performing tasks such as classification of images][37,38]) are well suited to learn from the complex and heterogeneous kinds of data that are generated from modern clinical care, such as medical notes entered by physicians, medical images, continuous monitoring data from sensors, and genomic data to help make medically relevant predictions. Guidance on when to use simple or sophisticated machine-learning models is provided in Table 2.

A key difference between human learning and machine learning is that humans can learn to make general and complex associations from small amounts of data. For example, a toddler does not need to see many examples of a cat to recognize a cheetah as a cat. Machines, in general, require many more examples than humans to learn the same task, and machines are not endowed with common sense. The flipside, however, is that machines can learn from massive amounts of data.[39] It is perfectly feasible for a machine-learning model to be trained with the use of tens of millions of patient charts stored in electronic health records (EHRs), with hundreds of billions of data points, without any lapses of attention, whereas it is very difficult for a human physician to see more than a few tens of thousands of patients in an entire career.

HOW MACHINE LEARNING CAN AUGMENT THE WORK OF CLINICIANS

PROGNOSIS

A machine-learning model can learn the patterns of health trajectories of vast numbers of patients. This facility can help physicians to anticipate future events at an expert level, drawing from information well beyond the individual physician's practice experience. For example, how likely is it that a patient will be able to return to work, or how quickly will the disease progress? At a population level, the same type of forecasting can enable reliable identification of patients who will soon have high-risk conditions or increased utilization of health care services; this information can be used to provide additional resources to proactively support them.[40]

Large integrated health systems have already used simple machine-learning models to automatically identify hospitalized patients who are at risk for transfer to the intensive care unit,[17] and retrospective studies suggest that more complex

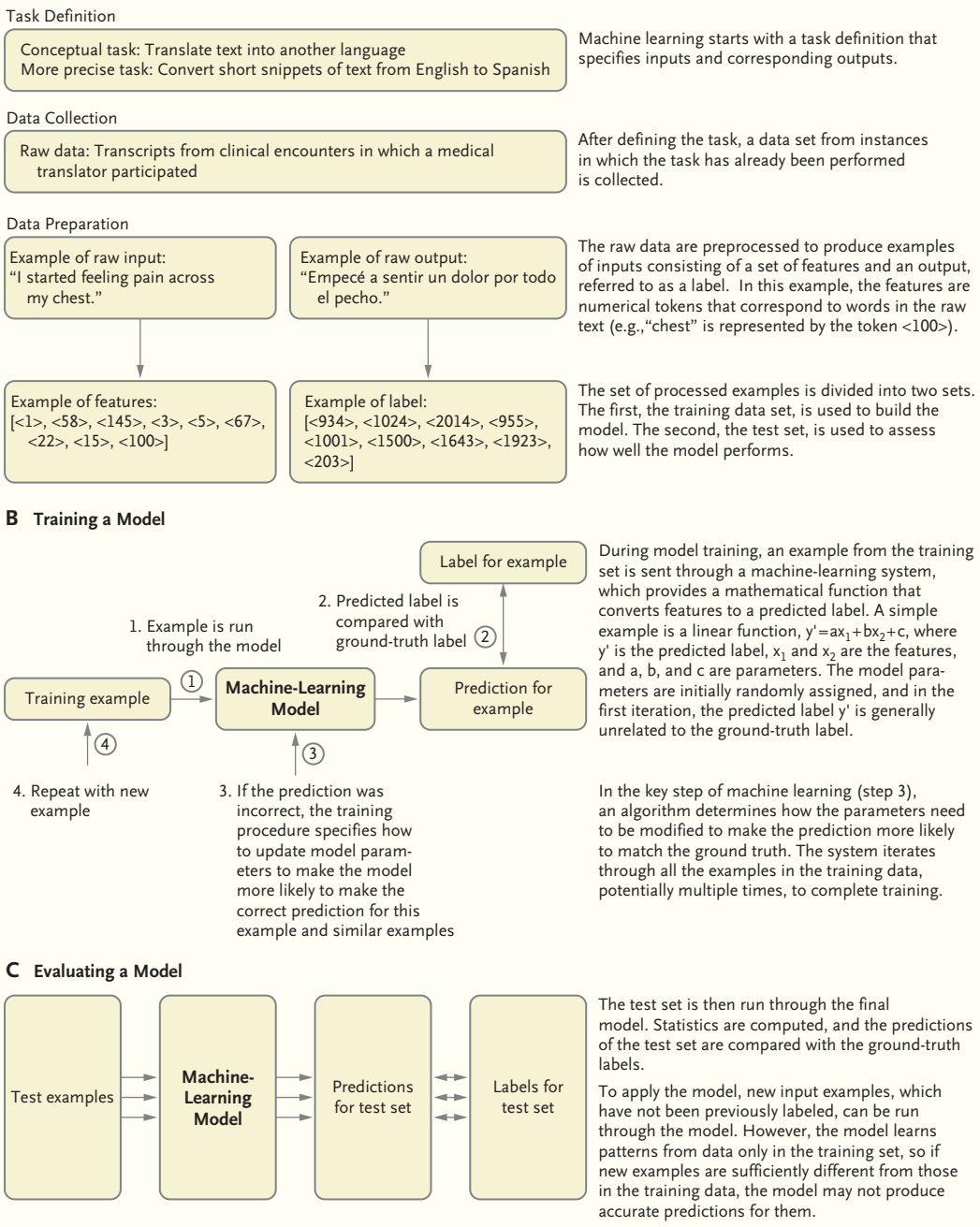

Figure 1. Conceptual Overview of Supervised Machine Learning.

As shown in Panel A, machine learning starts with a task definition that specifies an input that should be mapped to a corresponding output. The task in this example is to take a snippet of text from one language (input) and produce text of the same meaning but in a different language (output). There is no simple set of rules to perform this mapping well; for example, simply translating each word without examining the context does not lead to high-quality translations. As shown in Panel B, there are key steps in training machine-learning models. As shown in Panel C, models are evaluated with data that were not used to build them (i.e., the test set). This evaluation generally precedes formal testing to determine whether the models are effective in live clinical environments involving trial designs, such as randomized clinical trials.

Table 1. Examples of Types of Input and Output Data That Power Machine-Learning Applications.*

Application and Input Data (Feature)	Output Data (Label)	Purpose of Machine Learning	Comments
Commonly encountered machine-learning applications			
Cardiovascular risk factors (e.g., hypertension)	Myocardial infarction	Predicts a patient's cardiovascular risk (e.g., according to Framingham risk score or predicted by pooled cohorts equation[11])	Experts select pertinent risk factors (e.g., hypertension).
Search query (e.g., "What is the pooled cohorts equation?")	Web page that contains the most relevant information	Identifies the best source of information for user (e.g., Internet search engines[12])	Machine learning is a key component in determining the most relevant information to show users. Since there are innumerable types of Web queries, it is impractical to manually curate a list of results (outputs) for each query (input).
Sentence in Spanish (e.g., "¿Dónde está el dolor?")	Sentence in English ("Where is the pain?")	Translates from one language to another (e.g., computer translation[13,14])	
Daily clinical workflow			
Query about a patient's condition (e.g., "Has this patient previously had a venous thromboembolism?")	Information from a patient's chart that answers the question	Identifies key information from a patient's chart	Machine learning could help find information in a patient's chart that physicians want to see.
Audio recording of a conversation between a doctor and a patient	Sections of a clinical note that correspond to the conversation; correct billing codes	Provides automated generation of sections of a clinical note[15] or automatic assignment of billing codes from an encounter[16]	Automated documentation could ameliorate data-entry burdens of physicians.
Real-time EHR data	Clinical deterioration in next 24 hr	Provides real-time detection of patients at risk for clinical deterioration[17]	Real-time detection could serve as a patient-safety mechanism to ensure that attention is paid to patients at high risk.
Genetic sequence of a cancer	Event-free survival	Makes personalized predictions of patients' outcomes[18]	Machine learning, which can help use data that is difficult for humans to interpret, can make meaningful clinical contributions.
Workflow and use of medical images			
Image of the eye fundus	Diabetic retinopathy, myocardial infarction	Automates diagnosis of eye diseases,[19-24] predicts cardiac risk[25]	Automated diagnosis could be used in regions where screening tools exist but there are not enough specialists to review the images; a medical image can be used as a noninvasive "biomarker."
Digitized pathological slide of a sentinel lymph node	Detection of metastatic disease	Reviews pathological slides[26-28]	The efficiency and accuracy of experts in medical imaging can be augmented by machine learning.
CT of the head	Intracranial hemorrhage	Triages urgent findings in head CTs[29]	Abnormal images can be triaged by models to ensure that life-threatening abnormalities are diagnosed promptly.
Real-time video of a colonoscopy	Identification of polyps that warrant or do not warrant biopsy	Improves selection of polyps that require resection,[30] provides real-time feedback if a proceduralist risks injuring a key anatomical structure	Machine learning may enable "optical biopsies" to help proceduralists to identify high-risk lesions and ignore low-risk ones; real-time feedback is similar to lane-assist in cars.

Changes in patient experiences			
Real-time data from smartwatches or other digital sensors	Atrial fibrillation, hospitalization, laboratory test results	Provides outpatient screening of common diseases such as atrial fibrillation,[25,31] remote monitoring of ambulatory patients to detect conditions that warrant hospitalization, and detection of physiological abnormalities without blood draws[32]	Many trials have not shown a benefit for remote monitoring to detect conditions that warrant hospitalization, but with better sensors this may improve; sensors could be used as a noninvasive "biomarker" for traditionally measured laboratory tests.
Robotic movements of a prosthetic arm	Successful use of feeding utensils	Increases functionality of prosthetic equipment	Techniques used in the computer program AlphaGo, such as reinforcement learning, can be used to help robots perform complex tasks such as manipulating physical objects.
Text-messaging chat logs between a consumer and chatbot (i.e., a computer program that converses through sound or text)	Diagnoses	Provides early identification of patients who may need medical attention	Patients may delay seeking care when appropriate or seek care unnecessarily, leading to poor outcomes or waste.
Smartphone image of rash	Diagnosis	Provides automated triage of medical conditions[33]	Unnecessary medical consultation could be avoided, improving care.
Changes in regulations and billing			
Documentation of a medical encounter and "prior authorization" application	Insurance approval	Provides automated and instantaneous approval of insurance payment for medications	Automated approval could be used to cut down on administrative paperwork and reduce fraud.
Real-time EHR data	Future health care utilization	Identifies high-cost patients who may receive resource-intensive care in the future[34]	Proactive identification of patients who are at risk for clinical deterioration or gaps in care creates new opportunities to intervene early.
Real-time analysis of chest x-ray films, ventilation settings, and clinical variables	Diagnosis of ARDS	Provides automated identification of patients with ARDS and assessment of proper treatment	Current quality metrics may be suboptimal, so direct measurement for proper treatment, requiring correct case identification and assessment of treatment, can lead to improved quality measures without further documentation.[35]
Changes in epidemiologic factors			
Swabs of hospital rooms sent for genetic sequencing	Identification of pathogens	Provides real-time identification of possibly multidrug-resistant organisms	Immediate identification of emerging or important pathogens may enable rapid responses.

* Machine-learning models require collection of historical input and output data, which are also called features and labels. For example, a study that determined baseline cardiovascular risk factors and then followed patients for the occurrence of myocardial infarction would provide training examples in which the features were the set of risk factors and the label was a future myocardial infarction. The model would be trained to use the features to predict the label, so for new patients, the model would predict the risk of the occurrence of the label. This general framework can be used for a variety of tasks. ARDS denotes acute respiratory distress syndrome, CT computed tomography, and EHR electronic health record.

Table 2. Key Questions to Ask When Deciding What Type of Model Is Necessary.

How complex is the prediction task?
Simple prediction tasks are defined as those that can be performed with high accuracy with a small number of predictor variables. For example, predicting the development of hyperkalemia might be possible from just a small set of variables, such as renal function, the use of potassium supplements, and receipt of certain medications.
Complex prediction tasks are defined as those that cannot be predicted accurately with a small number of predictor variables. For example, identification of abnormalities in a pathological slide requires evaluation of patterns that are not obvious over millions of pixels.
In general, simple prediction tasks can be performed with traditional models (e.g., logistic regression), and complex tasks require more complex models (e.g., neural networks).
Should the prediction task be performed by clinicians who are entering the data manually, or should it be performed by a computer using raw data?
In addition to classifying a prediction task as simple or complex, consider how the model will be used in practice. If a model will be used in a bedside scoring system (e.g., the Wells score for assessment of the probability of pulmonary embolism), then using a small number of variables curated by humans is preferable. In this case, traditional models may be as effective as more complex ones.
If a model is expected to automatically analyze noisy data without any intervening human curation or normalization, then the task becomes complex, and complex models become generally more useful.
It is possible to write a set of rules to process raw data to a smaller set of "clean" features, which might be amenable to a traditional model if the prediction task is simple. However, it is often very time-consuming to write these rules and to keep them updated.
How many examples exist to train a model?
Simple prediction tasks generally do not require many examples to learn from in order to build a model.
The training of complex models generally requires many more examples. There is no predetermined number of examples, but at least multiple thousands of examples are needed to construct complex models, and the more complex the prediction task, the more data are generally required. Specialized techniques do exist to reduce the number of training examples that are necessary to construct an accurate model (e.g., transfer learning).
How interpretable does a model need to be?
Simple prediction tasks are interpretable because the number of features evaluated by the model is quite small.
Complex tasks are inherently harder to interpret because the model is expected to learn to identify complex statistical patterns, which might correspond to many small signals across many features. Although this complexity allows for more accurate predictions, it has the drawback of making it harder to succinctly present or explain the subtle patterns behind a particular prediction.

and accurate prognostic models can be built with raw data from EHRs[41] and medical imaging.[42]

Building machine-learning systems requires training with data that provide an integrated, longitudinal view of a patient. A model can learn what happens to patients only if the outcomes are included in the data set that the model is based on. However, data are currently siloed in EHR systems, medical imaging picture archiving and communication systems, payers, pharmacy benefits managers, and even apps on patients' phones. A natural solution would be to systematically place data in the hands of patients themselves. We have long advocated for this solution,[43] which is now enabled by the rapid adoption of patient-controlled application programming interfaces.[44]

Convergence of a unified data format such as Fast Healthcare Interoperability Resources (FHIR)[45] would allow for useful aggregation of data. Patients could then control who had access to their data for use in building or running models. Although there are concerns that technical interoperability does not solve the problem of semantic standardization endemic in EHR data,[46] the adoption of HTML (Hypertext Markup Language) has allowed Web data, which are perhaps even messier than EHR data, to be indexed and made useful with search engines.

DIAGNOSIS

Every patient is unique, but the best doctors can determine when a subtle sign that is particular to a patient is within the normal range or indicates a true outlier. Can statistical patterns detected by machine learning be used to help physicians identify conditions that they do not diagnose routinely?

The Institute of Medicine concluded that a diagnostic error will occur in the care of nearly every patient in his or her lifetime,[47] and receiving the right diagnosis is critical to receiving appropriate care.[48] This problem is not limited to rare conditions. Cardiac chest pain, tuberculosis, dysentery, and complications of childbirth are commonly not detected in developing countries, even when there is adequate access to therapies, time to examine patients, and fully trained providers.[49]

With data collected during routine care, machine learning could be used to identify likely diagnoses during a clinical visit and raise awareness of conditions that are likely to manifest later.[50] However, such approaches have limitations. Less skilled clinicians may not elicit the information necessary for a model to assist them meaningfully, and the diagnoses that the models are built from may be provisional or incorrect,[48] may be conditions that do not manifest symptoms (and thus may lead to overdiagnosis),[51] may be influenced by billing,[52] or may simply not be recorded. However, models could suggest questions or tests to physicians[53] on the basis of data collected in real time; these suggestions could be helpful in scenarios in which high-stakes misdiagnoses are common (e.g., childbirth) or when clinicians are uncertain. The discordance between diagnoses that are clinically correct and those recorded in EHRs or reimbursement claims means that clinicians should be involved from the outset in determining how data generated as part of routine care should be used to automate the diagnostic process.

Models have already been successfully trained to retrospectively identify abnormalities across a variety of image types (Table 1). However, only a limited number of prospective trials involve the use of machine-learning models as part of a clinician's regular course of work.[19,20]

TREATMENT

In a large health care system with tens of thousands of physicians treating tens of millions of patients, there is variation in when and why patients present for care and how patients with similar conditions are treated. Can a model sort through these natural variations to help physicians identify when the collective experience points to a preferred treatment pathway?

A straightforward application is to compare what is prescribed at the point of care with what a model predicts would be prescribed, and discrepancies could be flagged for review (e.g., other clinicians tend to order an alternative treatment that reflects new guidelines). However, a model trained on historical data would learn only the prescribing habits of physicians, not necessarily the ideal practices. To learn which medication or therapy should be prescribed to maximize patient benefit requires either carefully curated data or estimates of causal effects, which machine-learning models do not necessarily — and sometimes cannot with a given data set — identify.

Traditional methods used in comparative effectiveness research and pragmatic trials[54] have provided important insights from observational data.[55] However, recent attempts at using machine learning have shown that it is challenging to generate curated data sets with experts, update the models to incorporate newly published evidence, tailor them to regional prescribing practices, and automatically extract relevant variables from EHRs for ease of use.[56]

Machine learning can also be used to automatically select patients who might be eligible for randomized, controlled trials on the basis of clinical documentation[57] or to identify high-risk patients or subpopulations who are likely to benefit from early or new therapies under study. Such efforts can empower health systems to subject every clinical scenario for which there is equipoise to more rigorous study with decreased cost and administrative overhead.[54,58,59]

CLINICIAN WORKFLOW

The introduction of EHRs has improved the availability of data. However, these systems have also frustrated clinicians with a panoply of checkboxes for billing or administrative documentation,[60] clunky user interfaces,[61,62] increased time spent entering data,[63-66] and new opportunities for medical errors.[67]

The same machine-learning techniques that are used in many consumer products can be used to make clinicians more efficient. Machine learning that drives search engines can help expose relevant information in a patient's chart for a clinician without multiple clicks. Data entry of forms and text fields can be improved with the use of machine-learning techniques such as predictive typing, voice dictation, and

automatic summarization. Prior authorization could be replaced by models that automatically authorize payment based on information already recorded in the patient's chart.[68] The motivation behind adopting these abilities is not just convenience to physicians. Making the process of viewing and entering the most clinically useful data frictionless is essential to capturing and recording health care data, which in turn will enable machine learning to help give the best possible care to every patient. Most importantly, increased efficiency, ease of documentation, and improved automated clinical workflow would allow clinicians to spend more time with their patients.

Even outside the EHR system, machine-learning techniques can be adapted for real-time analysis of video of the surgical field to help surgeons avoid critical anatomical structures or unexpected variants or even handle more mundane tasks such as accurate counting of surgical sponges. Checklists can prevent surgical error,[69] and unstinting automated monitoring of their implementation provides additional safety.

In their personal lives, clinicians probably use variants of all these forms of technology on their smartphones. Although there are retrospective proof-of-concept studies of application of these techniques to medical contexts,[15] the major barriers to adoption involve not the development of models but technical infrastructure; legal, privacy, and policy frameworks across EHRs; health systems; and technology providers.

EXPANDING THE AVAILABILITY OF CLINICAL EXPERTISE

There is no way for physicians to individually interact with all the patients who may need care. Can machine learning extend the reach of clinicians to provide expert-level medical assessment without personal involvement? For example, patients with new rashes may be able to obtain a diagnosis by sending a picture that they take on their smartphones,[32,33] thereby averting unnecessary urgent-care visits. A patient considering a visit to the emergency department might be able to converse with an automated triage system and, when appropriate, be directed to another form of care. When a patient does need professional assistance, models could identify physicians with the most relevant expertise and availability. Similarly, to increase comfort and lower cost, patients who otherwise may need to be hospitalized could stay at home if machines can remotely monitor their sensor data.

The delivery of insights from machine learning directly to patients has become increasingly important in the areas of the world where access to direct medical expertise is in limited supply[70] and sophistication. Even in areas where the supply of expert clinicians is abundant, these clinicians are concerned about their ability and the effort required to provide timely and accurate interpretation of the tsunami of patient-driven digital data from sensor or activity-tracking devices worn by patients.[71] Indeed, one of the hopes with regard to machine-learning models trained with data from millions of patient encounters is that they can equip health care professionals with the ability to make better decisions. For instance, nurses might be able to take on many tasks that are traditionally performed by doctors, primary care doctors might be able to perform some of the roles traditionally performed by medical specialists, and medical specialists could devote more of their time to patients who would benefit from their particular expertise.

A variety of mobile apps or Web services that do not involve machine learning have been shown to improve medication adherence[72] and control of chronic diseases.[73,74] However, machine learning in direct-to-patient applications is hindered by formal retrospective and prospective evaluation methods.[75]

KEY CHALLENGES

AVAILABILITY OF HIGH-QUALITY DATA

A central challenge in building a machine-learning model is assembling a representative, diverse data set. It is ideal to train a model with data that most closely resemble the exact format and quality of data expected during use. For instance, for a model that is intended to be used at the point of care, it is preferable to use the same data that are available in the EHR at that particular moment, even if they are known to be unreliable[46] or subject to unwanted variability.[46,76] When they have large enough data sets, modern models can be successfully trained to map noisy inputs to noisy outputs. The use of a smaller set of curated data, such as those collected in clinical trials from manual chart review, is subopti-

mal unless clinicians at the bedside are expected to abstract the variables by hand according to the original trial specifications. This practice might be feasible with some variables, but not with the hundreds of thousands that are available in the EHR and that are necessary to make the most accurate predictions.[41]

How do we reconcile the use of noisy data sets to train a model with the data maxim "garbage in, garbage out"? Although to learn the majority of complex statistical patterns it is generally better to have large — even noisy — data sets, to fine-tune or evaluate a model, it is necessary to have a smaller set of examples with curated labels. This allows for proper assessment of the predictions of a model against the intended labels when there is a chance that the original ones were mislabeled.[21] For imaging models, this generally requires generating a "ground truth" (i.e., diagnoses or findings that would be assigned to an example by an infallible expert) label adjudicated by multiple graders for each image, but for nonimaging tasks, obtaining ground truth may be impossible after the fact if, for example, a necessary diagnostic test was not obtained.

Machine-learning models generally perform best when they have access to large amounts of training data. Thus, a key issue for many uses of machine learning will be balancing privacy and regulatory requirements with the desire to leverage large and diverse data sets to improve the accuracy of machine-learning models.

LEARNING FROM UNDESIRABLE PAST PRACTICES

All human activity is marred by unwanted and unconscious bias. Builders and users of machine-learning systems need to carefully consider how biases affect the data being used to train a model[77] and adopt practices to address and monitor them.[78]

The strength of machine learning, but also one of its vulnerabilities, is the ability of models to discern patterns in historical data that humans cannot find. Historical data from medical practice indicate health care disparities in the provision of systematically worse care for vulnerable groups than for others.[77,79] In the United States, the historical data reflect a payment system that rewards the use of potentially unnecessary care and services and may be missing data about patients who should have received care but did not (e.g., uninsured patients).

EXPERTISE IN REGULATION, OVERSIGHT, AND SAFE USE

Health systems have developed sophisticated mechanisms to ensure the safe delivery of pharmaceutical agents to patients. The wide applicability of machine learning will require a similarly sophisticated structure of regulatory oversight,[80] legal frameworks,[81] and local practices[82] to ensure the safe development, use, and monitoring of systems. Moreover, technology companies will have to provide scalable computing platforms to handle large amounts of data and use of models; their role today, however, is unclear.

Critically, clinicians and patients who use machine-learning systems need to understand their limitations, including instances in which a model is not designed to generalize to a particular scenario.[83-85] Overreliance on machine-learning models in making decisions or analyzing images may lead to automation bias,[86] and physicians may have decreased vigilance for errors. This is especially problematic if models themselves are not interpretable enough for clinicians to identify situations in which a model is giving incorrect advice.[87,88] Presenting the confidence interval in a prediction of a model may help, but confidence intervals themselves may be interpreted incorrectly.[89,90] Thus, there is a need for prospective, real-world clinical evaluation of models in use rather than only retrospective assessment of performance based on historical data sets.

Special consideration is needed for machine-learning applications targeted directly to patients. Patients may not have ways to verify that the claims made by a model maker have been substantiated by high-quality clinical evidence or that a suggested action is reasonable.

PUBLICATIONS AND DISSEMINATION OF RESEARCH

The interdisciplinary teams that build models may report results in venues that may be unfamiliar to clinicians. Manuscripts are often posted online at preprint services such as arXiv and bioRxiv,[91,92] and the source code of many models exists in repositories such as GitHub. Moreover, many peer-reviewed computer science manuscripts are not published by traditional journals

An audio interview with Dr. Kohane is available at NEJM.org

but as proceedings in conferences such as the Conference on Neural Information Processing Systems (NeurIPS) and the International Conference on Machine Learning (ICML).

CONCLUSIONS

The accelerating creation of vast amounts of health care data will fundamentally change the nature of medical care. We firmly believe that the patient–doctor relationship will be the cornerstone of the delivery of care to many patients and that the relationship will be enriched by additional insights from machine learning. We expect a handful of early models and peer-reviewed publications of their results to appear in the next few years, which — along with the development of regulatory frameworks and economic incentives for value-based care — are reasons to be cautiously optimistic about machine learning in health care. We look forward to the hopefully not-too-distant future when all medically relevant data used by millions of clinicians to make decisions in caring for billions of patients are analyzed by machine-learning models to assist with the delivery of the best possible care to all patients.

A 49-year-old patient takes a picture of a rash on his shoulder with a smartphone app that recommends an immediate appointment with a dermatologist. His insurance company automatically approves the direct referral, and the app schedules an appointment with an experienced nearby dermatologist in 2 days. This appointment is automatically cross-checked with the patient's personal calendar. The dermatologist performs a biopsy of the lesion, and a pathologist reviews the computer-assisted diagnosis of stage I melanoma, which is then excised by the dermatologist.

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

REFERENCES

1. Bakris G, Sorrentino M. Redefining hypertension — assessing the new blood-pressure guidelines. N Engl J Med 2018;378:497-9.
2. Institute of Medicine. Crossing the quality chasm: a new health system for the twenty-first century. Washington, DC: National Academies Press, 2001.
3. Lasic M. Case study: an insulin overdose. Institute for Healthcare Improvement (http://www.ihi.org/education/IHIOpenSchool/resources/Pages/Activities/AnInsulinOverdose.aspx).
4. Institute of Medicine. To err is human: building a safer health system. Washington, DC: National Academies Press, 2000.
5. National Academies of Sciences, Engineering, and Medicine. Improving diagnosis in health care. Washington, DC: National Academies Press, 2016.
6. Berwick DM, Gaines ME. How HIPAA harms care, and how to stop it. JAMA 2018;320:229-30.
7. Obermeyer Z, Lee TH. Lost in thought — the limits of the human mind and the future of medicine. N Engl J Med 2017;377:1209-11.
8. Schwartz WB. Medicine and the computer — the promise and problems of change. N Engl J Med 1970;283:1257-64.
9. Schwartz WB, Patil RS, Szolovits P. Artificial intelligence in medicine — where do we stand? N Engl J Med 1987;316:685-8.
10. Goodfellow I, Bengio Y, Courville A, Bengio Y. Deep learning. Cambridge, MA: MIT Press, 2016.
11. Muntner P, Colantonio LD, Cushman M, et al. Validation of the atherosclerotic cardiovascular disease Pooled Cohort risk equations. JAMA 2014;311:1406-15.
12. Clark J. Google turning its lucrative Web search over to AI machines. Bloomberg News. October 26, 2015 (https://www.bloomberg.com/news/articles/2015-10-26/google-turning-its-lucrative-web-search-over-to-ai-machines).
13. Johnson M, Schuster M, Le QV, et al. Google's multilingual neural machine translation system: enabling zero-shot translation. arXiv. November 14, 2016 (http://arxiv.org/abs/1611.04558).
14. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv. September 1, 2014 (http://arxiv.org/abs/1409.0473).
15. Kannan A, Chen K, Jaunzeikare D, Rajkomar A. Semi-supervised learning for information extraction from dialogue. In: Interspeech 2018. Baixas, France: International Speech Communication Association, 2018:2077-81.
16. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning for electronic health records. arXiv. January 24, 2018 (http://arxiv.org/abs/1801.07860).
17. 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:Suppl 1:S18-S24.
18. Grinfeld J, Nangalia J, Baxter EJ, et al. Classification and personalized prognosis in myeloproliferative neoplasms. N Engl J Med 2018;379:1416-30.
19. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25(1):44-56.
20. Wang P, Berzin TM, Glissen Brown JR, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut 2019 February 27 (Epub ahead of print).
21. Krause J, Gulshan V, Rahimy E, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology 2018;125:1264-72.
22. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316:2402-10.
23. Ting DSW, Cheung CY-L, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA 2017;318:2211-23.
24. Kermany DS, Goldbaum M, Cai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 2018;172(5):1122-1131.e9.
25. Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng 2018;2:158-64.
26. Steiner DF, MacDonald R, Liu Y, et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. Am J Surg Pathol 2018;42:1636-46.
27. Liu Y, Kohlberger T, Norouzi M, et al. Artificial intelligence-based breast cancer

nodal metastasis detection. Arch Pathol Lab Med 2018 October 8 (Epub ahead of print).
28. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017; 318:2199-210.
29. Chilamkurthy S, Ghosh R, Tanamala S, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet 2018; 392:2388-96.
30. Mori Y, Kudo SE, Misawa M, et al. Real-time use of artificial intelligence in identification of diminutive polyps during colonoscopy: a prospective study. Ann Intern Med 2018;169:357-66.
31. Tison GH, Sanchez JM, Ballinger B, et al. Passive detection of atrial fibrillation using a commercially available smartwatch. JAMA Cardiol 2018;3:409-16.
32. Galloway CD, Valys AV, Petterson FL, et al. Non-invasive detection of hyperkalemia with a smartphone electrocardiogram and artificial intelligence. J Am Coll Cardiol 2018;71:Suppl:A272. abstract.
33. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115-8.
34. Rajkomar A, Yim JWL, Grumbach K, Parekh A. Weighting primary care patient panel size: a novel electronic health record-derived measure using machine learning. JMIR Med Inform 2016;4(4):e29.
35. Schuster MA, Onorato SE, Meltzer DO. Measuring the cost of quality measurement: a missing link in quality strategy. JAMA 2017;318:1219-20.
36. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018;319:1317-8.
37. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44.
38. Hinton G. Deep learning — a technology with the potential to transform health care. JAMA 2018;320:1101-2.
39. Halevy A, Norvig P, Pereira F. The unreasonable effectiveness of data. IEEE Intell Syst 2009;24:8-12.
40. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff (Millwood) 2014;33:1123-31.
41. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. npj Digital Medicine 2018;1(1):18.
42. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 2018;24:1342-50.
43. Mandl KD, Szolovits P, Kohane IS. Public standards and patients' control: how to keep electronic medical records accessible but private. BMJ 2001;322:283-7.
44. Mandl KD, Kohane IS. Time for a patient-driven health information economy? N Engl J Med 2016;374:205-8.
45. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc 2016;23:899-908.
46. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care 2013;51: Suppl 3:S30-S37.
47. McGlynn EA, McDonald KM, Cassel CK. Measurement is essential for improving diagnosis and reducing diagnostic error: a report from the Institute of Medicine. JAMA 2015;314:2501-2.
48. Institute of Medicine, National Academies of Sciences, Engineering, and Medicine. Improving diagnosis in health care. Washington, DC: National Academies Press, 2016.
49. Das J, Woskie L, Rajbhandari R, Abbasi K, Jha A. Rethinking assumptions about delivery of healthcare: implications for universal health coverage. BMJ 2018; 361:k1716.
50. Reis BY, Kohane IS, Mandl KD. Longitudinal histories as predictors of future diagnoses of domestic abuse: modelling study. BMJ 2009;339:b3677.
51. Kale MS, Korenstein D. Overdiagnosis in primary care: framing the problem and finding solutions. BMJ 2018;362:k2820.
52. Lindenauer PK, Lagu T, Shieh M-S, Pekow PS, Rothberg MB. Association of diagnostic coding with trends in hospitalizations and mortality of patients with pneumonia, 2003-2009. JAMA 2012;307: 1405-13.
53. Slack WV, Hicks GP, Reed CE, Van Cura LJ. A computer-based medical-history system. N Engl J Med 1966;274:194-8.
54. Ford I, Norrie J. Pragmatic trials. N Engl J Med 2016;375:454-63.
55. Frieden TR. Evidence for health decision making — beyond randomized, controlled trials. N Engl J Med 2017;377:465-75.
56. Ross C, Swetlitz I, Thielking M, et al. IBM pitched Watson as a revolution in cancer care: it's nowhere close. Boston: STAT, September 5, 2017 (https://www.statnews.com/2017/09/05/watson-ibm-cancer/).
57. Fiore LD, Lavori PW. Integrating randomized comparative effectiveness research with patient care. N Engl J Med 2016;374:2152-8.
58. Schneeweiss S. Learning from big health care data. N Engl J Med 2014;370: 2161-3.
59. Institute of Medicine. The learning healthcare system: workshop summary. Washington, DC: National Academies Press, 2007.
60. Erickson SM, Rockwern B, Koltov M, McLean RM. Putting patients first by reducing administrative tasks in health care: a position paper of the American College of Physicians. Ann Intern Med 2017;166: 659-61.
61. Hill RG Jr, Sears LM, Melanson SW. 4000 Clicks: a productivity analysis of electronic medical records in a community hospital ED. Am J Emerg Med 2013; 31:1591-4.
62. Sittig DF, Murphy DR, Smith MW, Russo E, Wright A, Singh H. Graphical display of diagnostic test results in electronic health records: a comparison of 8 systems. J Am Med Inform Assoc 2015; 22:900-4.
63. Mamykina L, Vawdrey DK, Hripcsak G. How do residents spend their shift time? A time and motion study with a particular focus on the use of computers. Acad Med 2016;91:827-32.
64. Oxentenko AS, West CP, Popkave C, Weinberger SE, Kolars JC. Time spent on clinical documentation: a survey of internal medicine residents and program directors. Arch Intern Med 2010;170:377-80.
65. Arndt BG, Beasley JW, Watkinson MD, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. Ann Fam Med 2017;15:419-26.
66. Sinsky C, Colligan L, Li L, et al. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. Ann Intern Med 2016;165: 753-60.
67. Howe JL, Adams KT, Hettinger AZ, Ratwani RM. Electronic health record usability issues and potential contribution to patient harm. JAMA 2018;319:1276-8.
68. Lee VS, Blanchfield BB. Disentangling health care billing: for patients' physical and financial health. JAMA 2018;319:661-3.
69. Haynes AB, Weiser TG, Berry WR, et al. A surgical safety checklist to reduce morbidity and mortality in a global population. N Engl J Med 2009;360:491-9.
70. Steinhubl SR, Kim K-I, Ajayi T, Topol EJ. Virtual care for improved global health. Lancet 2018;391:419.
71. Gabriels K, Moerenhout T. Exploring entertainment medicine and professionalization of self-care: interview study among doctors on the potential effects of digital self-tracking. J Med Internet Res 2018;20(1):e10.
72. Morawski K, Ghazinouri R, Krumme A, et al. Association of a smartphone application with medication adherence and blood pressure control: the MedISAFE-BP randomized clinical trial. JAMA Intern Med 2018;178:802-9.
73. de Jong MJ, van der Meulen-de Jong AE, Romberg-Camps MJ, et al. Telemedicine for management of inflammatory bowel disease (myIBDcoach): a pragmatic, multicentre, randomised controlled trial. Lancet 2017;390:959-68.
74. Denis F, Basch E, Septans AL, et al. Two-year survival comparing web-based symptom monitoring vs routine surveillance following treatment for lung cancer. JAMA 2019;321(3):306-7.

75. Fraser H, Coiera E, Wong D. Safety of patient-facing digital symptom checkers. Lancet 2018;392:2263-4.
76. Elmore JG, Barnhill RL, Elder DE, et al. Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study. BMJ 2017;357:j2813.
77. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 2018;178:1544-7.
78. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med 2018;169:866-72.
79. Institute of Medicine. Unequal treatment: confronting racial and ethnic disparities in health care. Washington, DC: National Academies Press, 2003.
80. Shuren J, Califf RM. Need for a national evaluation system for health technology. JAMA 2016;316:1153-4.
81. Kesselheim AS, Cresswell K, Phansalkar S, Bates DW, Sheikh A. Clinical decision support systems could be modified to reduce 'alert fatigue' while still minimizing the risk of litigation. Health Aff (Millwood) 2011;30:2310-7.
82. Auerbach AD, Neinstein A, Khanna R. Balancing innovation and safety when integrating digital tools into health care. Ann Intern Med 2018;168:733-4.
83. Amarasingham R, Patzer RE, Huesch M, Nguyen NQ, Xie B. Implementing electronic health care predictive analytics: considerations and challenges. Health Aff (Millwood) 2014;33:1148-54.
84. Sniderman AD, D'Agostino RB Sr, Pencina MJ. The role of physicians in the era of predictive analytics. JAMA 2015;314:25-6.
85. Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. Health Aff (Millwood) 2014;33:1163-70.
86. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. J Am Med Inform Assoc 2017;24:423-31.
87. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. JAMA 2017;318:517-8.
88. Castelvecchi D. Can we open the black box of AI? Nature 2016;538:20-3.
89. Jiang H, Kim B, Guan M, Gupta M. To trust or not to trust a classifier. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, eds. Advances in neural information processing systems 31. New York: Curran Associates, 2018:5541-52.
90. Cohen IG, Amarasingham R, Shah A, Xie B, Lo B. The legal and ethical concerns that arise from using complex predictive analytics in health care. Health Aff (Millwood) 2014;33:1139-47.
91. arXiv.org Home page (https://arxiv.org/).
92. bioRxiv. bioRxiv: The preprint server for biology (https://www.biorxiv.org/).

IMAGES IN CLINICAL MEDICINE

The *Journal* welcomes consideration of new submissions for Images in Clinical Medicine. Instructions for authors and procedures for submissions can be found on the *Journal*'s website at NEJM.org. At the discretion of the editor, images that are accepted for publication may appear in the print version of the *Journal*, the electronic version, or both.