Demystifying artificial intelligence in pharmacy

Scott D. Nelson, PharmD, MS, Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN

Colin G. Walsh, MD, MA, Department of Biomedical Informatics, Medicine, and Psychiatry, Vanderbilt University Medical Center, Nashville, TN

Casey A. Olsen, PharmD, Advocate Aurora Health, Downers Grove, IL

Andrew J. McLaughlin, PharmD, MBA, Cerner Corporation, Kansas City, MO

Joseph R. LeGrand, PharmD, MS, HealthIT, Vanderbilt University Medical Center, Nashville, TN

Nick Schutz, PharmD, North Memorial Health, Robbinsdale, MN

Thomas A. Lasko, MD, PhD, Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN

Purpose. To provide pharmacists and other clinicians with a basic understanding of the underlying principles and practical applications of artificial intelligence (AI) in the medication-use process.

Summary. "Artificial intelligence" is a general term used to describe the theory and development of computer systems to perform tasks that normally would require human cognition, such as perception, language understanding, reasoning, learning, planning, and problem solving. Following the fundamental theorem of informatics, a better term for AI would be "augmented intelligence," or leveraging the strengths of computers and the strengths of clinicians together to obtain improved outcomes for patients. Understanding the vocabulary of and methods used in AI will help clinicians productively communicate with data scientists to collaborate on developing models that augment patient care. This primer includes discussion of approaches to identifying problems in practice that could benefit from application of AI and those that would not, as well as methods of training, validating, implementing, evaluating, and maintaining AI models. Some key limitations of AI related to the medication-use process are also discussed.

Conclusion. As medication-use domain experts, pharmacists play a key role in developing and evaluating AI in healthcare. An understanding of the core concepts of AI is necessary to engage in collaboration with data scientists and critically evaluating its place in patient care, especially as clinical practice continues to evolve and develop.

Keywords: artificial intelligence, machine learning, medical decision making, neural networks, prediction, medication systems

Am J Health-Syst Pharm. 2020; XX:XX-XX

Artificial intelligence (AI) has captured public interest with promises of improved quality and decreased cost of care.[1] Reports of suggested benefits in the media come in various forms, some overinflating and some underappreciating the technology's capabilities.[2] Although there has been a recent resurgence of this discussion in the media, AI is not a new phenomenon. The term *artificial intelligence* was first used in 1956,[3] and use of AI grew substantially in the 1970s due to increased availability of computing resources, resulting in the first phase of healthcare interest in AI.[3] However, limited computational power, small data sets, and modest results led to stagnant interest in and growth of AI in healthcare.[4] Now, recent advances in data availability, electronic health record (EHR) adoption, and processing power have paved the way for a resurgence of AI adoption in healthcare,[5] although there are plenty of challenges ahead.[6] Topol[7] has written an excellent review of the use of AI in the broader medical field, including a realistic assessment of how much farther the science has to advance before it makes a large impact on patient care. Specific to medication management, AI is already being actively used in research, patient-facing applications, and inventory management.

The goals of this primer are to help pharmacists (and other clinicians) lead in the design, implementation, and ongoing evaluation of AI-related applications and technologies that

Address correspondence to Dr. Nelson (Scott.Nelson@Vanderbilt.edu).

Twitter: @ScottNelsonRx

DOI 10.1093/ajhp/zxaa218

affect medication-use processes. We hope that the primer will teach pharmacists enough of the concepts and terminology of AI that they can identify good AI use cases in practice and can communicate effectively with the researchers and engineers who develop AI applications. We encourage pharmacists to:

1. Proactively engage in and promote communication among clinicians, pharmacists, and computer/data sciences as medication domain experts.
2. Identify problems or challenges that exist in their workplace that could benefit from AI.
3. Critically evaluate AI models and their claims.
4. Identify when a technology advance is likely to enhance patient care.

This primer is not intended to teach readers how to develop AI applications, nor is this an introduction to a machine learning course. Our intent is not to make pharmacists into data scientists but, instead, to help pharmacists contribute to AI projects by acting as domain experts or a translators between clinical practice and the computer sciences. Liu and colleagues[8] recently produced a guide for clinically oriented readers who wish a more technical introduction.

What is AI?

"Artificial intelligence" is a general term used to describe the theory and development of computer systems to perform tasks that normally would require human cognition, such as perception, language understanding, reasoning, learning, planning, and problem solving.[9-11] However, most applications of AI, especially in healthcare, are what some call "narrow AI." A narrow AI application enables a computer to perform a single, well-defined task that would normally require human intelligence, but it does not enable anything beyond that single task. For example, a narrow AI system might learn to recognize the retinal image patterns that represent diabetic retinopathy, but it would not understand the conversation that the clinician had with the patient about the condition, and it may or may not even recognize other retinal diseases. A set of narrow applications may be combined to perform a more complex task, such as driving a car. Several examples of narrow AI are provided in this primer.

KEY POINTS

- Artificial intelligence is already impacting healthcare and has the potential to make it faster and easier for pharmacists to fulfill their clinical responsibilities.
- This article provides definitions, explanations, and examples of key machine learning concepts and terminology.
- As medication-use domain experts, pharmacists play a key role in developing and evaluating artificial intelligence in healthcare.

Following the fundamental theorem of informatics, which states that the person plus the computer is better than either one alone,[2,12] a better term for AI would be "augmented intelligence"—leveraging the strengths of computers and the strengths of clinicians together to obtain improved outcomes for patients, making it faster and easier to fill our clinical responsibilities. The term *augmented intelligence* also connotes a more healthy perspective from which to view AI—an understanding that the technology is there not to replace me but to help me do my job faster and easier or to perform mundane or repetitive tasks, thus freeing me up to perform more interesting and higher-level tasks.[7,13]

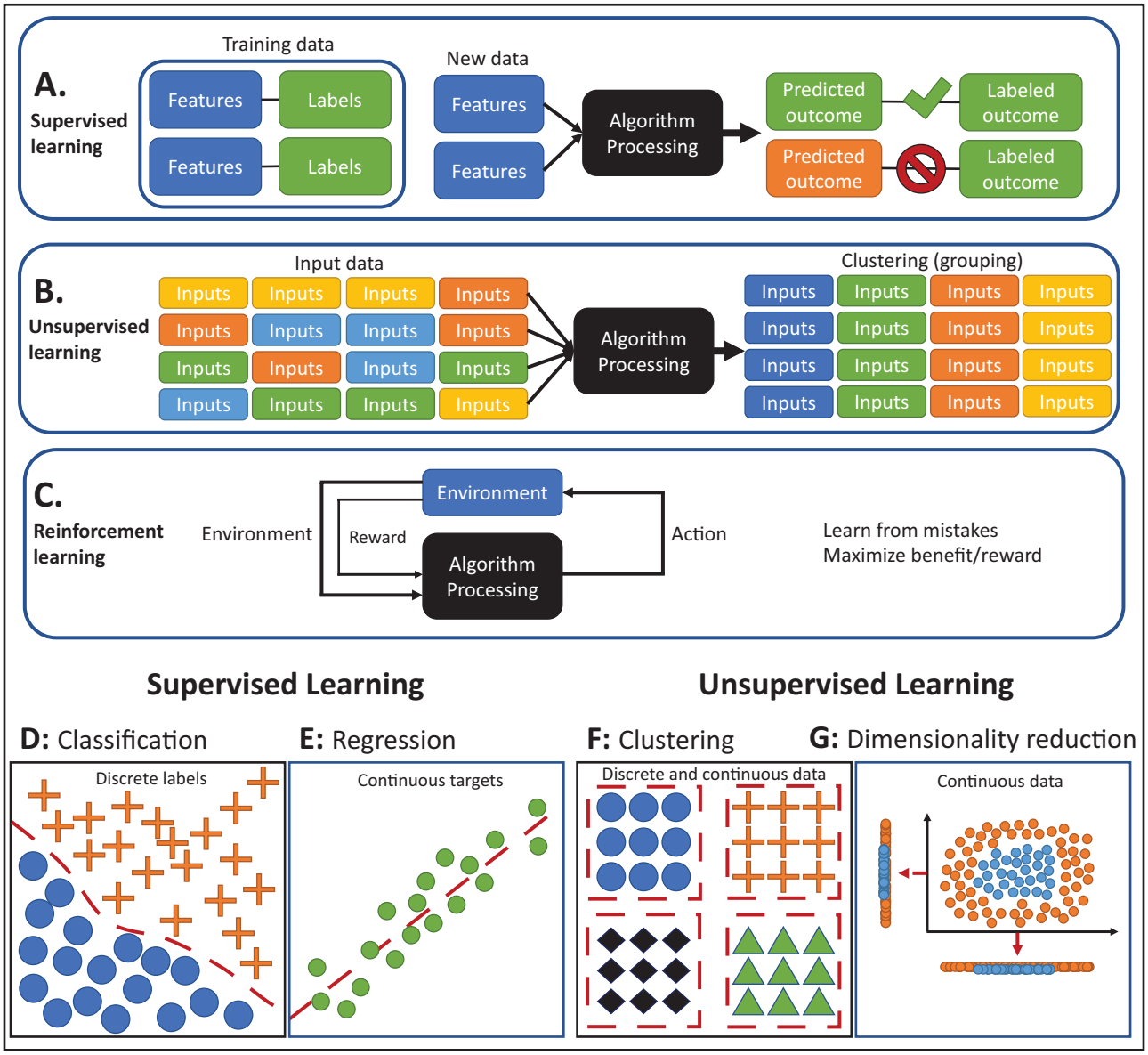
Historically, there has been a spectrum of technical approaches to AI, with what are called expert systems at one end and machine learning at the other. Expert systems are powered by explicitly programmed rules or heuristics. For example, the Internist-1 diagnostic engine used expert-defined rules to specify which diseases are evoked by given signs or symptoms, and defined formulas for combining evidence into a differential diagnosis.[14] In contrast, machine learning is a mathematical approach that is particularly useful when we do not know the rules that govern a particular phenomenon but have data on many examples of it happening. The computer itself learns patterns from the data[15-17] and then uses those patterns to make decisions about previously unseen examples.[16,17] Generally, using more data provides for better decisions,[18] including some that experts may not have considered; however, if the relationships have strong correlation, then smaller data sets may suffice. Furthermore, it is now usually easier and cheaper to collect a lot of data than it is to specify and continually refine many tedious rules, so machine learning approaches to AI have become dominant these days. For example, applications to understand written language (called natural language processing) used to be largely rule-based applications in the past, with heuristics requiring developers to tediously program specific rules, but now they tend to use machine learning instead of enormous data sets.[19] Because of this current dominance, the rest of this article will be about the machine learning component of AI rather than its other, more historical aspects.

Categories of machine learning

There are different categories of machine learning, which differ in terms of the data we have available and what kinds of things we want to learn from the data (Figure 1).

Superevised learning. This is the most common category of machine learning (Figure 1A). It is used when we want to make some kind of prediction based on example data, such as predicting whether a given patient would respond to a given antidepressant. The

Figure 1. Three machine learning models and examples of supervised and unsupervised learning. A. Supervised learning is a machine learning approach where the inputs (features) and outputs (labels or targets) are known. The computer learns the patterns and relationships between features and labels in the training data set and then uses those patterns to predict the labels for previously unseen input features. B. Unsupervised learning is a machine learning approach where the computer learns the patterns and relationships between input variables (features) without knowing the output variables. C. Reinforcement learning is a machine learning approach where the computer learns to make decisions on its own by making lots of decisions, learning from mistakes, and maximizing the benefit or reward. D. Classification is the process of predicting the class or category output (discrete label) of input variables (features). E. Regression is the process of using independent variables (inputs or features) to predict continuous dependent variables (outputs or targets). F. Clustering is the process of grouping inputs (features) based on the similarity and dissimilarity between them. G. Dimensionality reduction is the process of reducing the dimensions or complexity of the inputs (features) by finding a smaller set of composite variables that contain most of the information from the larger set of raw data.



input variables that form the basis of the prediction are called features. The predicted variable is commonly called the label if it is a categorical variable like "adverse event" or "no adverse event" or the target if it is a continuous variable like patient weight. Sometimes labels or targets naturally exist in the data (such as whether an Internet user clicked on an ad), and sometimes they need to be specified by humans (such as whether a patient responded to a treatment). The term *supervised* refers to the fact that the label or target provides analytic guidance in the form of the known

correct answer, and the algorithm looks for relationships between input variables and the provided label. If we have a sufficiently large data set in which both the features and the labels are known for all examples, then the computer can learn the patterns of relationships between them. The data set must be large enough for the true patterns to be distinguished from those present only by chance in the data set. Larger data sets are needed for larger numbers of input variables and if more complex relationships are sought. Once the patterns are identified, AI can be used to predict the output variables for new input examples. Supervised-learning problems are divided into classification problems for predicting categorical labels (Figure 1D) and regression problems for predicting continuous targets (Figure 1E). Pharmacists can play a key role as domain experts for supervised-learning projects by identifying what outcome variables (labels or targets) would be useful to predict, what kinds of input features may inform those predictions, and identifying data sets that may contain the features, labels, or targets. When outcome variables are not naturally present in the original data set, they can be assigned manually by experts, and pharmacy expertise may help do that as well. The Framingham Study, in which a linear model (Section 2.2) was developed to predict sudden coronary death, is a famous example of supervised learning.[20] Other examples include detecting meaningful brain activation in clinically unresponsive intensive care unit patients,[21] classifying arrhythmias in electrocardiogram traces,[22] and predicting hospital readmissions,[23] inpatient medication orders,[24] suicide attempts,[25] or acute kidney injury.[26] These kinds of predictions are usually made from known clinical, demographic, or other data about a patient.

Unsupervised learning. This is used when we want to understand the structure and relationships among input features rather than try to predict an outcome label from them (Figure 1B). The term *unsupervised* refers to the absence of a label for guiding the algorithm toward important relationships among variables. But because for many phenomena we do not yet know enough to assign labels or targets or it may be too labor intensive to do so, unsupervised learning can be an effective approach to understanding those phenomena. Historically, unsupervised learning mainly focused on clustering, or finding groups of similar data points (Figure 1F), and dimensionality reduction, which involves finding a smaller set of composite variables that contain most of the information found in the larger set of raw variables (Figure 1G). But more recently it has come to include much more sophisticated disentangling of complex patterns that would be needed to detect emergent diseases, to discover unknown effects of a medication, or to identify previously unrecognized drug diversion patterns. This new kind of application is sometimes called data-driven discovery.[27] Pharmacists can play a key role in these projects by identifying areas in which data-driven patterns may improve the precision or depth of our understanding, by identifying data sets that may contain those patterns, and by helping to assess and interpret the patterns once they are identified. For example, we might use unsupervised learning to understand drug diversion, because we do not have knowledge of which people are engaging in diversion. The algorithm might identify several different important patterns of variables, perhaps some representing patients in chronic pain, others representing patients in acute pain, and still others representing scenarios that are not immediately obvious. It would take human intelligence and domain expertise to understand what all of the patterns represent in real-world terms, and which of those patterns might correspond to diversion. Another example is use of unsupervised learning to identify groups of patients who could benefit the most from pharmacy interventions. A famous example of unsupervised learning was the discovery of a reliable method to differentiate acute myeloid leukemia vs acute lymphoblastic leukemia from gene expression data.[28] Other examples include identifying spatiotemporal patterns of epileptiform discharges in electroencephalogram (EEG) data,[29] discovering different fingerprints of disease in the temporal patterns of laboratory values,[30] and discovering prognostically relevant subtypes of glioblastoma in cancer gene expression data.[31]

Semisupervised learning. This lies somewhere between supervised and unsupervised learning. It is used when we want to make predictions but only have labels for a small fraction of the data examples. Semisupervised learning uses the unlabeled examples to try to improve the prediction accuracy of a model learned from the labeled examples. A common approach is to learn meaningful patterns about the population in general from the unlabeled data and then somehow use those patterns when learning to predict the labels; this can be helpful when labels are difficult or expensive to obtain but unlabeled data are abundant, with the tradeoff that the results are often not as accurate as when all labels are available. The key role of pharmacists in these projects is similar to that in supervised and unsupervised applications. Examples of semisupervised learning in the clinical domain are less common than the other types of learning, but they include identifying cell boundaries in images,[32] predicting preclinical toxicity of compounds,[33] and detecting adverse drug events described in Twitter posts.[34]

Reinforcement learning. This is used when we want the computer to learn to make decisions on its own, with the consequences of those decisions potentially appearing only long after the decisions are made (Figure 1C). This approach differs from the others described above in that it does not require an existing data set but requires only a context in which the computer can act over time and a way to measure the desirability of the current state. This setup naturally lends itself to computers learning to play games, because the context of the actions and

the rewards can be all contained in software; this has led to computer games becoming the dominant application of reinforcement learning. One particular approach has achieved superhuman performance on complex games like Go and chess by configuring the learning algorithm to play against itself for millions of games, thereby learning from billions of decisions.[35] Learning from a large number of decisions poses a problem for healthcare applications, because clinical decisions are made at a much slower rate than those of a computer playing a game against itself, and we cannot allow real patients to suffer the consequences of the mistakes that are required for the computer to learn. However, a reinforcement learning approach could be successful if a healthcare problem were cast as self-play in a strictly computational environment. For example, reinforcement learning was successfully used to control depth of anesthesia with propofol in healthy volunteers, using EEG signals for feedback, after the algorithm was trained using simulated physiology that modeled pharmacokinetics, pharmacodynamics, and many patient variations.[36] In this case, the simplicity of the control actions (a single variable representing the propofol infusion rate) and the environment state (a single variable representing depth of anesthesia, computed from the EEG), rendered this a tractable problem. More complex clinical problems will require much more complex simulations.

Specific algorithms and models

Every application of machine learning includes using an algorithm to learn informative patterns from a data set and to store those patterns in what is called a model. Differences among algorithms come down to how they learn those patterns from the data and how they represent them in the model. Algorithms to find more complex patterns generally need more data to reliably find them, but simpler patterns can be learned using smaller amounts of data.

Additionally, models that contain complex patterns may be difficult to understand. These models are said to suffer from low interpretability and are sometimes called "black box models." In healthcare, we tend not to like to use such models, but in some circumstances we may accept a lower interpretability in favor of a higher accuracy, especially if we have empirically verified that accuracy in our intended application.[37] This would be analogous to our use of a medication because we know through experiment that it is effective, even though we do not completely understand its mechanism of action.[38,39]

Linear models. Linear models are simple models for supervised learning that capture linear relationships between the inputs (features) and the outputs (labels or targets). They can be used for both regression problems and classification problems (as well as other kinds of problems), because these problems can be reduced internally to a regression problem that predicts a continuous value. How the actual scientific problem gets reduced internally to a regression problem determines the subtype of linear model that is used. If our desired output is a continuous variable, such as forced expiratory volume in 1 second (FEV_1), a measurement of lung function, we have a regression problem and could use a linear regression algorithm to learn how FEV_1 is affected by the input features of age, gender, and smoking status.[40] Linear regression models represent the simplest type of linear model because they directly predict the desired continuous value without any further transformation. If our desired output is a yes/no label, such as the development or nondevelopment of lung cancer, then we have a classification problem, and we would use the logistic regression algorithm to learn how that binary outcome is affected by the same input features.[41] The phrase "logistic regression" is somewhat confusing, because these models perform classification, not regression. The phrase refers to the fact that the models solve a linear regression problem internally, with a transformation (named the logistic transformation) to transform the inner continuous value into probability estimates for the yes/no label values. If we want to understand when an event happens in time, such as death due to lung cancer, then we have defined a time-to-event or survival problem and might use a Cox regression algorithm to learn how the input features affect 1-year survival.[42] Because linear models are relatively simple and therefore easily interpretable, they are commonly used in healthcare for identifying risk factors, such as the determinants of long-term antidepressant use.[43] Their simplicity also lends them for use as easily implementable risk scores,[44,45] such as the CHA_2DS_2VASc score,[46] quick SOFA score,[47] LACE index,[48] and Framingham Risk Score.[49] But their simplicity is also their main limitation: They can only identify linear relationships between input features and targets, and any potential interactions between variables must be explicitly specified ahead of time.

Most machine-learning practitioners consider the algorithms that produce these linear models to be examples of machine learning, because although very simple, the mathematical details of the models are nevertheless learned by the algorithms from data. Some statisticians disagree because these models predate the term *machine learning,* and they would not consider them to fall under the umbrella of AI. As one might imagine, the debate can become quite heated.

Tree-based models. A decision tree is a model that captures the organization of a data set as a tree-like structure, where the data set is split in succession at each branch of the tree, depending on the values of its variables. At one end of the tree (the *root*) the data set is in one piece, and at the other (the *leaves*), it has been split into small similar subsets. These models are usually built by a supervised learning algorithm, with the goal of having the same label or target value for most examples in a given leaf. The advantages

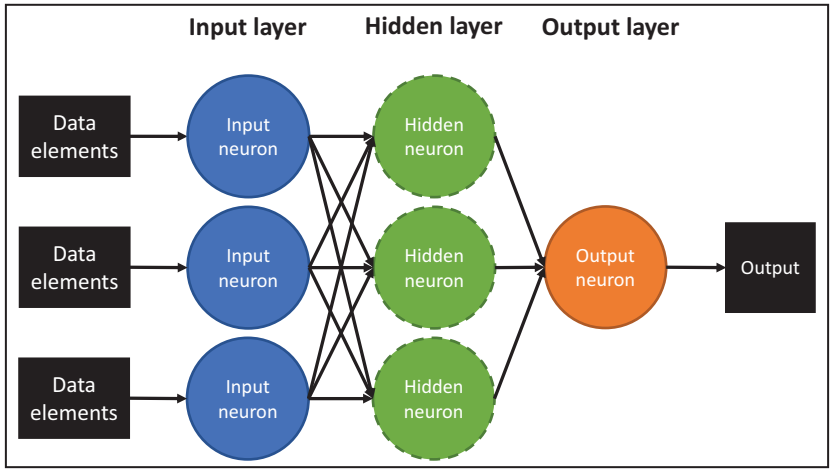
of decision trees are their clear interpretability and their ability to capture complex patterns and variable interactions in the branching points. The biggest downside is that a single tree tends to make relatively inaccurate predictions compared to other approaches. However, the accuracy can be greatly improved by learning an ensemble of many trees that differ in strategic ways (such as which variables are considered at each branch point) and then aggregating their predictions. The most common versions of ensemble tree models are random forests and gradient-boosted trees.[50,51] These models are very popular because, in addition to being able to learn complex patterns, they can handle data sets with many more input features than data examples (which is what most genetic data sets and complex healthcare data sets look like), a scenario that few other approaches can cope with. By analyzing the ensemble, we can also identify variables with important relationships to the target, such as key risk factors associated with early kidney transplant rejection.[52]

Bayesian models. While all models use the language of probability to express the patterns they find or the predictions they make, Bayesian models are explicitly designed to use the laws of probability internally, and much of their effort goes into quantifying the uncertainty around the learned patterns and predictions.[53] Moreover, Bayesian methods can infer values and uncertainties for hidden variables (such as the psychiatric state of a patient) that we cannot directly measure. These inferences can be valuable, but they can take a lot of computational effort. Bayesian methods have therefore become more common in recent years as computational power has increased.[54] An example of their use is modeling complex pharmacokinetics equations, such as those used to estimate the ratio of vancomycin area under the curve (AUC) to minimum inhibitory concentration (AUC/MIC) using a single vancomycin level, along with serum creatinine, age, and weight (as compared to manual AUC/MIC calculations, which require 2 vancomycin levels and thus delay clinical intervention).[55] Algorithms for producing Bayesian models have the advantage of being able to continually improve their estimates as more data are added a little at a time, compared to algorithms for other models that need all of the data together at once. The famous drawback of Bayesian models is their need for the designer to specify what the baseline probability distributions should be for the case when no data about the system are observed (these are called the prior probability distributions). These specifications are important because they have a strong influence on the output when there are only small amounts of data. A fair amount of effort with Bayesian methods also goes into evaluating how well the final model fits the data and how sensitive it is to different prior probability distributions.

Neural network models. Neural network models are thus named because their structure is inspired by the organization of the nervous system.[15,56] The building block of a neural network is the neuron, which is basically a linear model that captures relationships between the neuron's input features and its output value, usually followed by a nonlinear distortion like the transformation used by logistic regression models. The complexity and power of a neural network comes from the fact that the inputs and outputs of these neurons are connected together in layers, with the outputs of neurons in one layer all feeding the inputs of neurons in the next layer. This architecture means the network can learn highly complex, nonlinear relationships between the initial input features and the final output target or label (Figure 2). Like Bayesian models, neural network models can be continually improved by adding data a little at a time. However, neural network models have the expected drawbacks of complex models—they require more data to learn complex relationships, and once learned, the relationships are difficult to visualize or interpret. Existing examples of neural network models include identifying myocardial infarctions in the emergency department[57] and predicting a patient's response to warfarin.[58]

Deep models. For decades, most neural network models had only 1 or 2 hidden layers (Figure 2). Attempts to increase the number of layers (and therefore the power of the models) wound up with the extra layers learning nothing

Figure 2. Example of a neural network. A neural network connects several "neurons" (linear models) together, where the outputs form one layer become the inputs for the next. The input layer processes the raw data inputs and then passes the results to the next layer (hidden layer) for additional processing, with the results then passed to the output neuron to obtain the final result.



meaningful and all of the real learning being done in the final, hidden layer. This had become such an impediment that most of the field had moved on to other machine learning algorithms by the mid 1990s. However, in 2006 a breakthrough allowed for meaningful learning in multiple layers of a neural network (Figure 3), and this opened whole new avenues for experimentation and development.[59] After just a few years, neural networks with many hidden layers, called deep models, deep architectures, or deep learning, had become the dominant approach in several scientific domains, such as computer vision and speech recognition.[60] (It turns out that the same key ideas had been previously described as early as 1965 but were not given a catchy name and were not widely known within the machine-learning community.[61]) Specific innovations in deep architectures have been responsible for leaps forward in supervised, semisupervised, unsupervised, and reinforcement learning. Deep architectures have the same drawbacks as shallow neural networks but even more so; they require even larger data sets, and it can be even more difficult to visualize and interpret what the network has learned. Nevertheless, the leap in predictive accuracy made possible by deep learning is what has powered the current explosion of interest in machine learning and AI. Deep learning also excels at discovering and modeling very complex hidden variables that we are not able to measure directly. It has been applied to EHR data for predicting in-hospital mortality, 30-day unplanned readmission, prolonged length of stay, and final discharge diagnoses.[62] Other examples include interpreting radiology images,[63] predicting acute kidney injury,[26] detecting diabetic retinopathy from images,[64] medication adherence using video to confirm medication ingestion,[65] and drug discovery.[66,67]

Hierarchically, a deep learning model is a specific type of neural network model, which is in turn a subset of all models used in machine learning, which is a subdiscipline of the AI field (Figure 4).

Model development

A key challenge of AI in the medication-use process can be the gap between the clinicians who understand the problems, the developers who create the models, and the administrators who make the decisions about which AI solutions to finance. Clinicians should understand where an AI solution could be a good fit and, conversely, where there are simpler rules or equations that already exist and would perform just as well and with less complexity. As stated above, a pharmacist's key role in model development is that of domain or subject-matter expert. Pharmacists understand the workflows, the problems to be solved, and the data sets available to help solve those problems. They can help steer model development in a direction that will augment their existing clinical skills. Developers need input and requirements from clinicians who both understand pertinent complex clinical problems and can understand the basic needs of developing a model. By empowering clinicians to communicate effectively with developers, administrators can gain clearer insight into which solutions would have the greatest benefit. Without all of these groups working together, we could end up with a highly funded data science effort that results in the discovery of something that is not clinically useful or is already commonly known (such as that fever is associated with pneumonia). In this section, we will look at this process of model development, with a focus on how to identify problems that can benefit from machine learning.

Define the task or problem to be solved. The first step is identifying and defining a task or problem to be solved. There are plenty of areas in a pharmacist's everyday work that could benefit from machine learning and plenty of areas where it cannot or should not be used. For example, calculating a renal dose adjustment

Figure 3. Example of a deep neural network, a subtype of neural networks with multiple hidden layers between the input and output layers. The outputs from one layer of neurons becomes the inputs for the next layer of neurons.

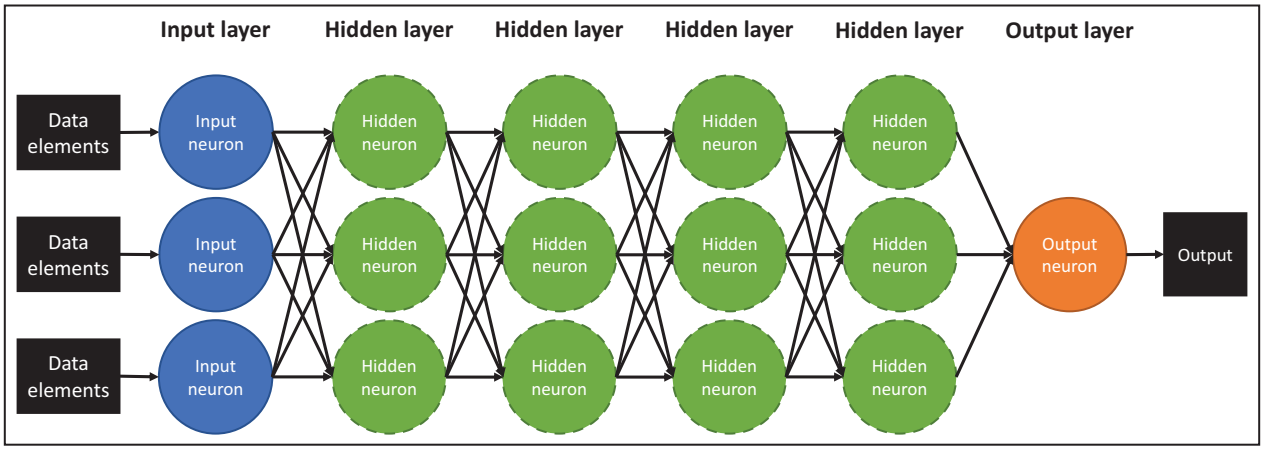
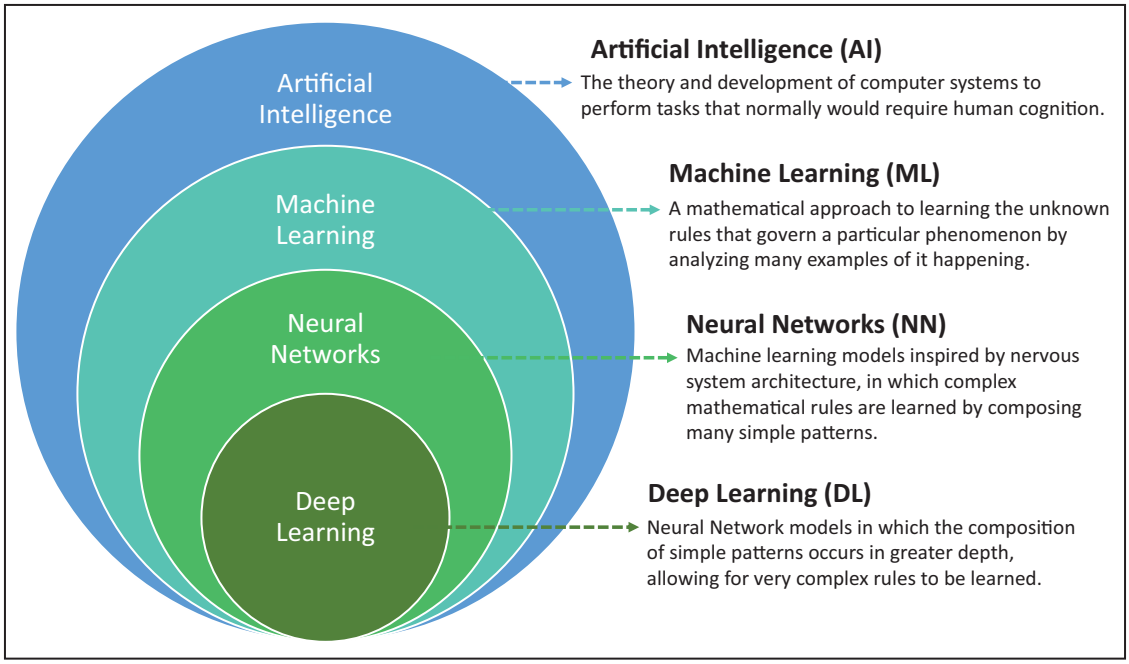


Figure 4. How deep learning fits into the hierarchy of artificial intelligence (AI). Deep learning models represent a specific subset of neural network models, which are in turn a subset of all models used in machine learning, which is a subdiscipline of the AI field.



for sertraline would be a better application for standard software than for machine learning, because we already understand the processes that govern that calculation. On the other hand, if there were unknown factors that affected the half-life of sertraline, identifying those factors would be a good application for supervised learning, as long as we had many examples where both the potential factors and the sertraline level were measured. A good task definition includes answering certain design questions. These include:

1. Are we proposing a classification task (estimating the probability of a yes/no label), a regression task (estimating a continuous target value), or an unsupervised task (finding patterns in the absence of a target)?
2. Do I care more about knowing the relationships between the inputs and the outputs, or do I care more about the accuracy of the estimate? This impacts the tradeoff between simple and complex models.
3. What do we already know about those relationships? If we know the relationships are simple correlations, with few interactions between variables, then we can work with a smaller data set and a simpler model.
4. How accurate would the predictions or patterns need to be to be clinically useful? The amount of data available often determines how accurate the prediction can be. If we need very accurate predictions but have very little data, then this would not be a viable project.
5. Is the task something that a human could do with sufficient accuracy if given the available data and enough time? If not, then a computer is unlikely to be able to do it either. This thought process leads us to either look for additional data that make the task possible or move on to a different problem.

There are good, in-depth discussions elsewhere of these and other questions that help guide the development of clinical models.[8,68-70]

Collecting and preparing the data. Once a specific, useful, and realistic task is defined, the next question is whether there are enough data available to train a model to perform the task. Obtaining access to data and ensuring the quality of the data used for creating a model are among the most important and time-consuming aspects of model development and use. The quantity and quality of the data used to train the model will ultimately determine the model's quality and credibility. While data availability and preparation are challenging in machine learning in general, they are even more so in healthcare. Much of the data in the EHR comes from manual clinician documentation that is prone to gaps and errors, such as a medication administration not being documented at all, being documented in the EHR later than it actually occurred, or even being documented in the record of the wrong patient. Additionally, huge portions of EHR data are stored in an unstructured format, such as free text or images.[71] For instance, provider notes and study interpretations are typically entered as

free-text narratives that most learning algorithms cannot directly use. Instead, natural language processing tools are used to extract specific, structured items from the text, and that structured data is then used as a data source for a model, such as extracting and structuring left ventricular ejection fractions from free-text reports[72] or capturing medication infusion–related data from free-text infusion notes.[73] (Incidentally, the development of these natural language processing tools is another large field of AI). While there are ways to incorporate elements of discrete data capture within provider notes where appropriate (for example, picklists built into various parts of the note that can store discrete data), the availability and reliability of these mechanisms can vary on a case-by-case basis. And, of course, a model cannot learn anything from data it does not have access to or that simply do not exist.[2]

For all of the machine learning approaches we mention here, the input data set takes the form of a rectangular matrix, which looks like a spreadsheet with 1 row per example and 1 column per input variable. If our data cannot be represented that way, even after preprocessing, then it may not be appropriate to apply a machine learning approach. Good questions to ask at this stage include the following:

1. Exactly what information would a human need in order to complete the task, whether it is a supervised task of predicting a variable or an unsupervised task of understanding a phenomenon from its data? Can that information be represented as a few dozen carefully crafted variables, or is it distributed among hundreds or thousands of variables (such as image pixels or the set of all International Classification of Diseases codes)? If a human could do the task using a small number of variables, that suggests that this is a simple task that can be performed by a linear model trained with a relatively small amount of data, maybe a few hundred examples. If the task requires thousands of variables, that implies a more complex model and much more data, possibly millions of examples.
2. If the task is to predict the value of a particular variable, is that value available or computable from what is already in the data set? If the value is a categorical label that is not inherently part of the data set, we can sometimes use clinical expertise to assign labels to examples one by one, although this can get very expensive.
3. How noisy or incomplete are the data? If there are more than a small fraction of empty cells in the data matrix or if many of the variables are known to be noisy, this increases the complexity of the project (although not necessarily the complexity of the model), and decreased accuracy should be expected.
4. When are data available as part of the workflow? Would we be able to act on the output of the model in time to affect the outcome? This is important for considering when and where in the workflow the results of the model should be computed and whether there are sufficient data available at or before that time in the process for the model to use after implementation.

Training the model. Once collected into a matrix form, the data can be used to train a model. This training usually involves applying a computational algorithm that tries different settings for the various parameters of the model (of which there can be millions in a deep model) and updating the parameters incrementally, gradually improving the model's performance. This is the step that most people consider to be where the real machine learning happens, and the people involved in this step need a deep understanding of the specifics of the model and its training algorithm. This step is also where most of the research literature is concentrated. However, when evaluating whether machine learning is a good match for a given project, the details discussed here probably represent the least important of all steps.

Evaluating the model. After a model has been trained, it needs to be evaluated by testing it on new data, because the model may have come to rely on relationships between variables that are only present by chance in the training data. If a machine learning algorithm is sufficiently powerful, it can produce a model that uses those random relationships to essentially memorize the training set. That model can achieve close to perfect performance on the training set, but then it will perform poorly on previously unseen data. When this happens, even to a small degree, it is called overfitting.

Understanding whether a model reported in the literature was ever tested on data it had never seen before is key in assessing whether it is likely to perform well in the real world. Some reports leave out this step, which leaves the model open to the possibility of overfitting. A common mistake is to build several models, try each of them on a held-out validation set, and then choose the model that performs best on that set. However, if any choice of models is made (by a human or an algorithm) based on the performance of a model on a data set, that data set becomes part of the training set and is no longer useful for validation. A true validation set is only used to test performance after the final model has been selected.

Various schemes to test for overfitting provide different ways to provide new test data that the model was not trained on. The simplest way to do that is to randomly set aside what is called a holdout test set and use that set to test the model after training. The main drawback to this approach is that holding back test data decreases the amount of data available for training. Other, more involved methods, such as k-fold cross validation and bootstrap resampling, partially overcome this drawback by averaging the performance of many different models, each trained by leaving out a different set of data each time.

For example, we may consider a hypothetical model to predict whether use of a given antidepressant would result in an adverse effect in a given patient. For such a model, we might use 10,000 patient records representing a diverse group of individuals who received various antidepressants and any related adverse effects recorded as outcomes. We would then train the model using 8,000 patient records in training data and hold out the remaining 2,000 records for model testing. The learning algorithm would produce a predictive model that captures the relationship between the input features and outcome labels for the 8,000 records. For validation, we would test the model by using the input features from the 2,000 holdout records and assess the degree to which the predicted labels agree with the labels of the holdout set. This type of validation is sometimes called internal validation and is usually the first step in determining how well the model can perform when exposed to data it has not encountered before.

More sophisticated validation strategies to assess model accuracy exist, and the key test for clinical use is whether there has been prospective validation, meaning that the model has been tested on an entirely new data set, collected as it was created (as opposed to collected retrospectively from an existing data store), under the same conditions that the model will encounter in practice. The best case is to perform this validation at the site intended for its use. When the validation site is different from the development site (or at least the data collection circumstances are different from the original training set), this is called external validation. The external, prospective collection of data can avoid the serious problem of information from the target variable inadvertently leaking into the input data, which causes the model to appear to be much more accurate than it really would be in practice. This information leakage problem (and several other sources of model error) can be very difficult to detect without external and prospective validation. Additional information and considerations about validation are available in the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines.[74,75]

All validation depends on being able to assess the accuracy of a model's predictions. In general, the predictions or outputs of a classifier model, such as a model for predicting antidepressant-associated adverse effects, are not a simple yes/no value. Instead, they produce a value between 0 and 1 that represents the classifier's confidence that the answer is yes. If a model produces the number 0.72 for a given patient, it can be interpreted as the predicted probability that this patient will have the outcome (ie, an adverse effect). We usually talk about this number as a probability, but it is a probability in the sense of the "belief" or the "confidence" that the classifier (model) has in its answer. There are 2 ways in which we want this number to be accurate, and they are evaluated by measuring the discrimination and the calibration of the model.

Discrimination is a measure of how well the model ranks and separates the populations of those who will and those who will not experience an effect. It is usually calculated and reported as the area under the receiver operating characteristic (AUROC) curve.[76] An AUROC curve value of 0.5 is produced by random guessing, and a value of 1.0 indicates perfect discrimination. However, keep in mind that the AUROC curve does not indicate clinical utility.[7]

While discrimination indicates how well a model works in terms of populations, calibration indicates how well it works in terms of individuals; in the example discussed above, that means assessing the accuracy of the model's prediction of a 72% probability of an adverse effect. One way to measure this is to consider all cases associated with the 72% risk prediction and calculate how often the predicted outcome occurred (ie, how many predictions were actually positive). If 72% of the predictions are actually positive, calibration is perfect for that point. If only 60% are actually positive, or if 80% are actually positive, then the calibration is off and the model will be less useful in clinical practice. If the model indicates a 72% probability of an adverse effect, we would like to be able to count on that number as a real probability and weigh it with confidence against other factors in making a final prescription decision. A value of 72% would mean that the adverse effect is more likely than not but there is a nontrivial chance it will *not* happen; in that case, prescribing an antidepressant may be worth the risk if other factors argue in favor of use of the medication. However, this kind of reasoning is not possible if the model is not well calibrated and we do not have confidence in the actual number. Historically, researchers have not had enough data to assess calibration with this type of fine resolution, but approximations can be made by binning (rounding predictions into small groups or bins, such as every 5%) or smoothing the data (such as with a running windowed average). Recently, more rigorous efforts at assessing calibration have been developed.[77] It is still relatively uncommon for calibration measures to be reported in the literature, but this information is critically important when considering the use of a model in clinical practice.

In some applications, models are configured to give a discrete yes/no answer rather than giving a probability or risk estimate. This configuration requires us to take the predictions (which fall between 0 and 1) and turn them into discrete yes or no predictions (usually called positive or negative). Once we make this transformation, we can then judge whether the discrete predictions are true or false. Making the transformation requires us to choose a decision threshold and compare it to the predicted probability. So, we might choose a threshold of 60%, and then our 72% prediction would be considered positive because 72% is greater than 60%. If the adverse event actually occurred, then this would be a true positive; otherwise it is a false positive. This gives

us 4 possible buckets for each prediction, which are commonly arranged in a 2 x 2 matrix: true positives, false positives, true negatives, and false negatives. Various measures that emphasize different things can be computed using the counts in each bucket. These measures include sensitivity (also known as recall), specificity, positive predictive value (also known as precision), and negative predictive value. Explaining each of these measures is beyond the scope of this primer, but good tutorials are available in the literature.[78-80]

We can imagine choosing a decision threshold that takes into account the consequences of making a false positive vs a false negative answer. A higher threshold tends to give fewer false positives at the expense of more false negatives, and vice versa, so setting the decision threshold usually takes into account the different costs of false negative vs false positive errors. Setting a decision threshold can make life simpler when interpreting model predictions, but it is hard to set an appropriate threshold before a specific use or population has been selected. And it is even harder when, for example, one patient cares more about false negatives and another cares more about false positives. In these cases, it can be more helpful to leave the prediction as a probability and address the costs of errors on an individual basis.

Implementation and monitoring

After the model has been trained and validated, it is important to assess the model's readiness for production use and monitoring.[81] As in other software implementations, it is important to identify the stakeholders and users who will be affected by the application of the model. Including early adopters early in the process can help ensure the model outputs can be seamlessly integrated, understood, and actionable as part of their workflows; they may also champion the capabilities of the model and eventually become subject matter experts. As such, they can help ensure success by assisting with onboarding and ongoing education to staff on the capabilities, limitations, and desired workflows associated with the model. Additionally, many clinicians may not know how to interpret a given predictive score from a model, so establishing thresholds and actions to be taken when a score reaches that threshold are useful.

Information technology and data science experts will be valuable in understanding how the software associated with the model is installed and the degree to which it can be configured. In the absence of dedicated staff to fulfill these roles, it may be necessary to work with the vendor to either provide ongoing support or train staff to ensure successful sustainment of the model. These individuals should also be responsible for ensuring that input features for the model can be readily provided on an ongoing basis.

After a model has been implemented into practice, it must be continually evaluated and monitored. Patterns that the model relies on for making its predictions may change over time in the real world. There may be changes in the prevalence of disease or patient population characteristics (such as an increased prevalence of obesity over time); medications may be added and removed from practice; and there may changes in clinical practice patterns or data collection procedures. Any of these changes may result in the model becoming less accurate over time, a phenomenon called performance drift. To cope with performance drift, the model's performance must be actively monitored and there must be an explicit policy for when model updates will happen, which may result in considerable costs to ensure the model is updated to maintain its accuracy.[82-84] Applications whose users care more about calibration (such as identifying individual-level risk of readmission) as opposed to discrimination (such as identifying a population with a given disease) are more likely to require recalibration of models over time.[82,84]

Additionally, an interesting feedback loop happens when a model is in place to predict events (such as sepsis) that clinicians will then act to prevent. If they are successful at preventing the event, then the monitor will consider the prediction a false positive. One way to prevent this is to build the intervention into the model so that models can adjust for the actions recommended by the model's predictions.[85] Pharmacists play a key role in knowing the types of interventions that could potentially impact a model, and they could be a crucial part of ensuring the model's long-term validity. This is an interesting area of active research.

Watching for unanticipated consequences. There are many biases and unanticipated consequences that could result from using AI in healthcare, which are beyond the scope of this primer; however, care should be taken to avoid overreliance on AI, which could lead to automation bias. It has been shown that traditional clinical decision support (CDS) tools such as drug interaction alerts can lead to increased errors of commission (acting incorrectly) and omission (failing to act when we should) compared to having no decision support at all,[86] and the same issues can translate to CDS built using AI. While AI is able to evaluate more complex problems than traditional CDS, it usually does not have perfect accuracy, and clinical verification should validate any recommendations from the system.

Evaluating the claims of a model. It is necessary to critically evaluate the claims made by a model's developers in order to determine if a particular model can provide value in a specific situation.[8,87] The key is whether the model predictions are useful in a particular situation. In general, this evaluation process could be similar to those used for evaluating a diagnostic test.[88,89] Pharmacists can play a key role in this process as domain experts, using their training, experience, curiosity, and pragmatism. Regardless of prior experience with AI models, a pharmacist can begin to evaluate the claims of a model by asking questions similar to the following:

1. How would this model provide value in our organization?
2. What would people do with the model's outputs or predictions?
3. How accurate is the model (consider discrimination and calibration)?
4. Is that accuracy clinically significant and useful?
5. Does the population used to train the model reflect the population served by our organization?
6. How thoroughly was the model validated? What did prospective external validation show?
7. How well does the model maintain accuracy with changing inputs (drift)?
8. How can we collect the data necessary to use this model in production?
9. Do we need near real-time data feeds to use the model in production, and can we get them?
10. Is the model technically feasible in our setting? Who would implement and maintain the technical infrastructure?
11. What could be the potential unintended consequences of using the model at our organization?

Hypothetical example of model evaluation. As a hypothetical example of model claims evaluation, a technology vendor may approach your organization asserting that it has recently developed a supervised machine learning model that helps improve depression management by guiding a prescriber to select the most appropriate antidepressant for a specific patient. On the surface, this sounds like an appealing proposal to provide value and improve patient care at your organization, but that will be so only if the model works as well in your organization's patient population as it is purported to perform in the marketing materials.

Suppose that the vendor provides some internal validation examples and reports an AUROC curve value of 0.85, which sounds pretty good to you. However, the fact that accuracy was reported as an AUROC curve value implies that the medication selection problem has been approached as a yes/no classification problem. And indeed, you read in the vendor-supplied publications that what was actually measured in validation was whether a recommended medication decreased depressive symptoms (as measured by a validated instrument)—not whether the recommended drug was actually better than other therapeutic options—and that overall symptom improvement in treated patients was counted as a success if the decrease was statistically significant but not necessarily clinically meaningful. So, now you are a little more skeptical of the marketing claims. Additionally, you would like to see the outcomes that other clients have achieved in using the model, especially the results of prospective external validation in a patient population similar to your organization's. You compare aspects of those client populations to your organization's population—the rates of depression, the demographics of the patients, and the clinical practice patterns that may affect the accuracy of the model. You discover that development of the model involved a younger, urban population, whereas your organization's patient population is a primarily older, rural population (although one of the clients is reported to have used the model successfully in a rural setting). You decide that your organization would need to pilot test the model on a recent sample of data before committing to a production installation.

As you discuss the model with prescribers at your organization, you discover that they have tended to gravitate towards a small set of medications that they understand well and that they would be hesitant to use recommendations produced by a model. However, they could perhaps be persuaded if there were enough evidence of the model's accuracy. You decide that a pilot trial with a few willing clinicians would be needed to collect evidence of model accuracy in your patient population, with the understanding that the clinicians could override the recommendation of the model if they considered that appropriate.

When assessing the data requirements for the model, you determine that you have access to most of the data needed as input features and that you are able to provide these data from the EHR at the time that a treatment decision is made. You also are able to assess that most input data are reliable. However, you note that the model specifies that smoking status be expressed in terms of packs per day and alcohol use be expressed in terms of drinks per day as input features, but your organization records information on smoking only as narrative text in progress notes—and then only when the clinician decides to ask about it. You realize that implementing the model in your organization will require using natural language processing tools to extract data on smoking and alcohol use before feeding it to the model.

All in all, you consider that the costs of adapting your organization's data stream to the input needs of the model—and the likely clinical resistance to the model's recommendations until its utility has been proven—are not worth the modest gains in clinical outcome likely to result from using the model in the targeted population.

Conclusion

AI is not a new phenomenon and is becoming more common in healthcare as data sets and computing power continue to grow. Augmented intelligence in healthcare leverages the strengths of computers and the strengths of clinicians together to achieve improved outcomes for patients, making it faster and easier to perform clinical activities. As medication-use domain experts, pharmacists play a key role in developing, evaluating, and implementing AI in healthcare. An understanding of the core concepts of AI and machine learning is necessary to engage in collaboration with data scientists, to critically evaluate a model's place in patient care, and to solve real-world problems related to the medication-use process.

Disclosures

The authors have declared no potential conflicts of interest.

Additional information

Dr. Nelson and Dr. Lasko designed and directed the writing project. All authors assembled, discussed, wrote, and revised the article manuscript.

References

1. Forbes Media LLC. Ai and healthcare: a giant opportunity. https://www.forbes.com/sites/insights-intelai/2019/02/11/ai-and-healthcare-a-giant-opportunity/#11059a8d4c68. Published February 11, 2019. Accessed September 2019.
2. Chen JH, Asch SM. Machine learning and prediction in medicine — beyond the peak of inflated expectations. *N Engl J Med.* 2017;376(26):2507-2509.
3. Patel VL, Shortliffe EH, Stefanelli M, et al. The coming of age of artificial intelligence in medicine. *Artif Intell Med.* 2009;46(1):5-17.
4. Kulikowski CA. Beginnings of artificial intelligence in medicine (AIM): computational artifice assisting scientific inquiry and clinical art – with reflections on present aim challenges. *Yearb Med Inform.* 2019;28(1):249-256.
5. Morris KC, Schlenoff C, Srinivasan V. A remarkable resurgence of artificial intelligence and its impact on automation and autonomy. *IEEE Trans Autom Sci Eng.* 2017;14(2):407-409.
6. Stead WW. Clinical implications and challenges of artificial intelligence and deep learning. *JAMA.* 2018;320(11):1107-1108.
7. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25(1):44-56.
8. Liu Y, Chen P-HC, Krause J, et al. How to read articles that use machine learning: Users' guides to the medical literature. *JAMA.* 2019;322(18):1806-1816.
9. Russell SJ, Norvig P. *Artificial intelligence: a modern approach.* Kuala Lumpur, Malaysia: Pearson Education Limited; 2016.
10. Bellman RE. *An introduction to artificial intelligence: can computers think?* San Francisco, CA: Boyd & Fraser Pub. Co; 1978.
11. Winston PH. *Artificial intelligence.* 3rd ed. Reading, MA: Addison-Wesley Pub. Co.; 1992.
12. Friedman CP. A "fundamental theorem" of biomedical informatics. *J Am Med Inform Assoc.* 2009;16(2):169-70.
13. Holden M, Smith JPa. Washington DC: National Science and Technology Council; 2016.
14. Miller RA, Pople HE Jr, Myers JD. Internist-1, an experimental computer-based diagnostic consultant for general internal medicine. *N Engl J Med.* 1982;307(8):468-476.
15. Murphy KP. *Machine learning: a probabilistic perspective.* Cambridge, MA: MIT Press; 2012.
16. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA.* 2018;319(13):1317-1318.
17. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med.* 2019;380(14):1347-1358.
18. Banko M, Brill E. Scaling to very very large corpora for natural language disambiguation. Paper presented at Proceedings of 39th Annual Meeting of the Association for Computational Linguistics; July 2001; Toulouse, France.
19. Halevy A, Norvig P, Pereira F. The unreasonable effectiveness of data. *IEEE Intell Syst.* 2009;24(2):8-12.
20. Kannel WB, Doyle JT, McNamara PM, et al. Precursors of sudden coronary death. Factors related to the incidence of sudden death. *Circulation.* 1975;51(4):606-613.
21. Claassen J, Doyle K, Matory A, et al. Detection of brain activation in unresponsive patients with acute brain injury. *N Engl J Med.* 2019;380(26):2497-2505.
22. Hannun AY, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med.* 2019;25(1):65-69.
23. Zhou H, Della PR, Roberts P, et al. Utility of models to predict 28-day or 30-day unplanned hospital readmissions: an updated systematic review. *BMJ Open.* 2016;6(6):e011060.
24. Rough K, Dai AM, Zhang K, et al. Predicting inpatient medication orders from electronic health record data [published online ahead of print March 5, 2020]. *Clin Pharmacol Ther.* doi:10.1002/cpt.1826.
25. Walsh CG, Ribeiro JD, Franklin JC. Predicting risk of suicide attempts over time through machine learning. *Clin Psychol Sci.* 2017;5(3):457-469.
26. Tomasev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature.* 2019;572(7767):116-119.
27. Stalzer M, Mentzel C. A preliminary review of influential works in data-driven discovery. *Springerplus.* 2016;5(1):1266.
28. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science.* 1999;286(5439):531-537.
29. Baud MO, Kleen JK, Anumanchipalli GK, et al. Unsupervised learning of spatiotemporal interictal discharges in focal epilepsy. *Neurosurgery.* 2018;83(4):683-691.
30. Lasko TA, Denny JC, Levy MA. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PLoS One.* 2013;8(6):e66341.
31. Young JD, Cai C, Lu X. Unsupervised deep learning reveals prognostically relevant subtypes of glioblastoma. *BMC Bioinformatics.* 2017;18(suppl 11):381.
32. Ramesh N, Liu T, Tasdizen T. Cell detection using extremal regions in a semisupervised learning framework. *J Healthc Eng.* 2017;2017:4080874.
33. Luechtefeld T, Marsh D, Rowlands C et al. Machine learning of toxicological big data enables read-across structure activity relationships (rasar) outperforming animal test reproducibility. *Toxicol Sci.* 2018;165(1):198-212.
34. Lee K, Qadir A, Hasan SA, et al. Adverse drug event detection in tweets with semi-supervised convolutional neural networks. Paper presented at Proceedings of the 26th International Conference on World Wide Web; April 2017; Perth, Australia.
35. Silver D, Hubert T, Schrittwieser J, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science.* 2018;362(6419):1140-1144.
36. Moore BL, Pyeatt LD, Kulkarni V, et al. Reinforcement learning for closed-loop propofol anesthesia: a study in human volunteers. *J Mach Learn Res.* 2014;15(1):655-696.
37. London AJ. Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Cent Rep.* 2019;49(1):15-21.
38. Malhi GS, Tanious M, Das P, et al. Potential mechanisms of action of lithium in bipolar disorder. Current understanding. *CNS Drugs.* 2013;27(2):135-153.
39. Ban TA. The role of serendipity in drug discovery. *Dialogues Clin Neurosci.* 2006;8(3):335-344.
40. Schneider A, Hommel G, Blettner M. Linear regression analysis: part 14 of a series on evaluation of scientific publications. *Dtsch Arztebl Int.* 2010;107(44):776-782.
41. Sperandei S. Understanding logistic regression analysis. *Biochem Med (Zagreb).* 2014;24(1):12-18.

42. Sedgwick P. Statistical question cox proportional hazards regression. *Bmj-British Medical Journal.* 2013; 347.
43. Meijer WE, Heerdink ER, Leufkens HG, et al. Incidence and determinants of long-term use of antidepressants. *Eur J Clin Pharmacol.* 2004;60(1):57-61.
44. Sullivan LM, Massaro JM, D'Agostino RB Sr. Presentation of multivariate data for clinical use: The Framingham Study risk score functions. *Stat Med.*2004;23(10):1631-1660.
45. Wu C, Hannan EL, Walford G, et al. A risk score to predict in-hospital mortality for percutaneous coronary interventions. *J Am Coll Cardiol.* 2006;47(3):654-660.
46. Lip GY, Nieuwlaat R, Pisters R, et al. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: The Euro Heart Survey on atrial fibrillation. *Chest.* 2010;137(2):263-272.
47. Seymour CW, Liu VX, Iwashyna TJ, et al. Assessment of clinical criteria for sepsis: for the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA.* 2016;315(8):762-774.
48. van Walraven C, Dhalla IA, Bell C, et al. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *CMAJ.* 2010;182(6):551-557.
49. D'Agostino RB Sr, Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care: The Framingham Heart Study. *Circulation.* 2008;117(6):743-753.
50. Breiman L. Random forests. *Mach Learn.* 2001;45(1):5-32.
51. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat.* 2001;29(5):1189-1232.
52. Shaikhina T, Lowe D, Daga S, et al. Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. *Biomed Signal Process Control.* 2019;52:456-462.
53. Gelman A, Carlin JB, Stern HS, et al. *Bayesian data analysis*. Boca Raton, FL: Chapman and Hall/CRC; 2013.
54. Spiegelhalter DJ, Myles JP, Jones DR, et al. Methods in health service research. An introduction to Bayesian methods in health technology assessment. *BMJ.* 1999;319(7208):508-512.
55. Pai MP, Neely M, Rodvold KA, et al. Innovative approaches to optimizing the delivery of vancomycin in individual patients. *Adv Drug Deliv Rev.* 2014;77:50-57.
56. Hartnell N, MacKinnon NJ. Neural networks: from science fiction to pharmacy. *Am J Health-Syst Pharm.* 2003;60(18):1908-1909.
57. Baxt WG. Use of an artificial neural network for the diagnosis of myocardial infarction. *Ann Intern Med.* 1991;115(11):843-848.
58. Narayanan MN, Lucas SB. A genetic algorithm to improve a neural network to predict a patient's response to warfarin. *Methods Inf Med.* 1993;32(1):55-58.
59. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science.* 2006;313(5786):504-507.
60. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521(7553):436-444.
61. Schmidhuber J. Deep learning in neural networks: an overview. *Neural Netw.* 2015;61:85-117.
62. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med.* 2018;1:18.
63. McBee MP, Awan OA, Colucci AT, et al. Deep learning in radiology. *Acad Radiol.* 2018;25(11):1472-1480.
64. 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(22):2402-2410.
65. Labovitz DL, Shafner L, Reyes Gil M, et al. Using artificial intelligence to reduce the risk of nonadherence in patients on anticoagulation therapy. *Stroke.* 2017;48(5):1416-1419.
66. Lavecchia A. Deep learning in drug discovery: Opportunities, challenges and future prospects. *Drug Discov Today.* 2019.
67. Schneider G. Automating drug discovery. *Nat Rev Drug Discov.* 2018;17(2):97-113.
68. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an abcd for validation. *Eur Heart J.* 2014;35(29):1925-1931.
69. Lee YH, Bang H, Kim DJ. How to establish clinical prediction models. *Endocrinol Metab (Seoul).* 2016;31(1):38-44.
70. Royston P, Moons KG, Altman DG, et al. Prognosis and prognostic research: developing a prognostic model. *BMJ.* 2009;338:b604.
71. Kong HJ. Managing unstructured big data in healthcare system. *Healthc Inform Res.* 2019;25(1):1-2.
72. Garvin JH, DuVall SL, South BR, et al. Automated extraction of ejection fraction for quality measurement using regular expressions in unstructured information management architecture (UIMA) for heart failure. *J Am Med Inform Assoc.* 2012;19(5):859-866.
73. Nelson SD, Lu CC, Teng CC, et al. The use of natural language processing of infusion notes to identify outpatient infusions. *Pharmacoepidemiol Drug Saf.* 2015;24(1):86-92.
74. Collins GS, Reitsma JB, Altman DG, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis OR Diagnosis (TRIPOD): The TRIPOD statement. *Ann Intern Med.* 2015;162(1):55-63.
75. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med.* 2015;162(1):W1-W73.
76. Lasko TA, Bhagwat JG, Zou KH, et al. The use of receiver operating characteristic curves in biomedical informatics. *J Biomed Inform.* 2005;38(5):404-415.
77. Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Stat Med.* 2019;38(21):4051-4065.
78. Maxim LD, Niebo R, Utell MJ. Screening tests: a review with examples. *Inhal Toxicol.* 2014;26(13):811-828.
79. Altman DG, Bland JM. Diagnostic tests. 1: sensitivity and specificity. *BMJ.* 1994;308(6943):1552.
80. Altman DG, Bland JM. Diagnostic tests 2: predictive values. *BMJ.* 1994;309(6947):102.
81. Breck E, Cai S, Nielsen E, et al. *2017 IEEE International Conference on Big Data*. Piscataway, NJ. IEEE; 2017:1123-1132.
82. Davis SE, Lasko TA, Chen G, et al. Calibration drift in regression and machine learning models for acute kidney injury. *J Am Med Inform Assoc.* 2017;24(6):1052-1561.
83. Davis SE, Greevy RA, Fonnesbeck C, et al. A nonparametric updating method to correct clinical prediction model drift. *J Am Med Inform Assoc.* 2019;26(12):1448-1457.
84. Kramer AA. Predictive mortality models are not like fine wine. *Crit Care.* 2005;9(6):636-637.
85. Lenert MC, Matheny ME, Walsh CG. Prognostic models will be victims of their own success, unless. *J Am Med Inform Assoc.* 2019;26(12):1645-1650.
86. Lyell D, Magrabi F, Coiera E. Reduced verification of medication alerts increases prescribing errors. *Appl Clin Inform.* 2019;10(1):66-76.

87. Shah NH, Milstein A, Bagley SC. Making machine learning models clinically useful. *JAMA*. 2019;322(14):1351-1352.
88. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? *JAMA*. 1994;271(5):389-391.
89. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? *JAMA*. 1994;271(9):703-707.