Evaluating the Machine Learning Literature: A Primer and User's Guide for Psychiatrists

Adrienne Grzenda, M.D., Ph.D., Nina V. Kraguljac, M.D., William M. McDonald, M.D., Charles Nemeroff, M.D., Ph.D., John Torous, M.D., M.B.I., Jonathan E. Alpert, M.D., Ph.D., Carolyn I. Rodriguez, M.D., Ph.D., Alik S. Widge, M.D., Ph.D.

"Mr. A," a 24-year-old man, presents for evaluation of worsening depression. He describes a history of depression since adolescence, although he notes that he suffered a troubled childhood, including emotional neglect. He believes a recent breakup and having been denied a promotion precipitated this episode. "I'm sleeping all the time, and my body feels heavy," he adds. He also reports increased appetite, weight gain, and "urges to cut, which I have not done in years." However, he remains social and actively involved in several hobbies. He discontinued bupropion and escitalopram in the past because of "terrible headaches and irritability." Initially, you consider starting lamotrigine. However, your office recently implemented a clinical decision support system that recommends a trial of phenelzine. The patient's symptoms remit entirely on the medication suggested by the system. Curious as to how the system decided on this treatment, you download several papers on its development.

Health care is in the midst of a big data explosion (1). An estimated zettabyte (one trillion gigabytes) of clinical data is produced annually, with the rate anticipated to double every 2 years (2). "Big data" consists of data sets of a size (volume), rate of accumulation (velocity), and diversity of data types (variety) that exceed the capacity of traditional storage and processing. Sources of big data in health care include electronic health records, administrative claims, public health records, "omics" output (e.g., whole-genome sequencing), laboratory results, imaging, and sensor recordings. Internet-enabled devices, such as smartphones and wearables (collectively known as the "Internet of things"), and social media (e.g., Twitter, Facebook) generate real-time data streams (3, 4).

Machine learning and natural language processing, subfields of artificial intelligence, offer promising solutions for harnessing big data toward meaningful insights and applications (5). Safe and successful integration of machine learning or natural language processing-based applications or results into clinical practice will require physicians to comprehend new and sometimes complex methods. Because machine learning and natural language processing are not (yet) routinely incorporated into medical school postgraduate training curricula, this change in the practice environment will require development of new forms of data and statistical literacy (6). This primer provides an introduction to the fundamentals of these methods to empower readers in critically evaluating the emerging literature. Interested readers should consult the online supplement for further introductory reading recommendations.

MACHINE LEARNING

Artificial intelligence aims to develop machines or computer programs to perform tasks associated with human intelligence, including perception, learning, and reasoning (7). A convergent discipline, artificial intelligence draws from statistics, mathematics, computer science, data science, and engineering, among other fields. Artificial intelligence is broadly subdivided into symbolic and nonsymbolic approaches based on how knowledge is ingested, represented, and manipulated toward solving a problem (8).

In the symbolic approach, a program is supplied with knowledge in declarative form, that is, using human-readable symbols that represent facts and rules (e.g., if-then rules, or propositional logic) for how the symbols interrelate to solve complex problems. When queried, the program will use search and logic to reach a conclusion. Because the knowledge must be manually entered, symbolic programs are labor-intensive to develop and intolerant of conflicting data, but their reasoning can be clearly understood. Symbolic programs are best suited to narrowly circumscribed,

See related feature: **CME course** *(online and p. 729)*

static problems, such as playing chess, as famously exemplified by IBM's Deep Blue. In the nonsymbolic approach, a program is supplied with copious amounts of data and a general problem-solving strategy, or algorithm. The program searches for the underlying patterns in the data ("learns") and devises its own mathematical representation of the knowledge model. Nonsymbolic programs adapt to new conditions and conflicting data but are data hungry, and their decision making is not always easily understood. Machine learning, artificial neural networks, and deep learning are implementations of the nonsymbolic approach. These techniques are by no means new, but the exploding availability of big data has driven a rapid expansion of their use across domains and industries.

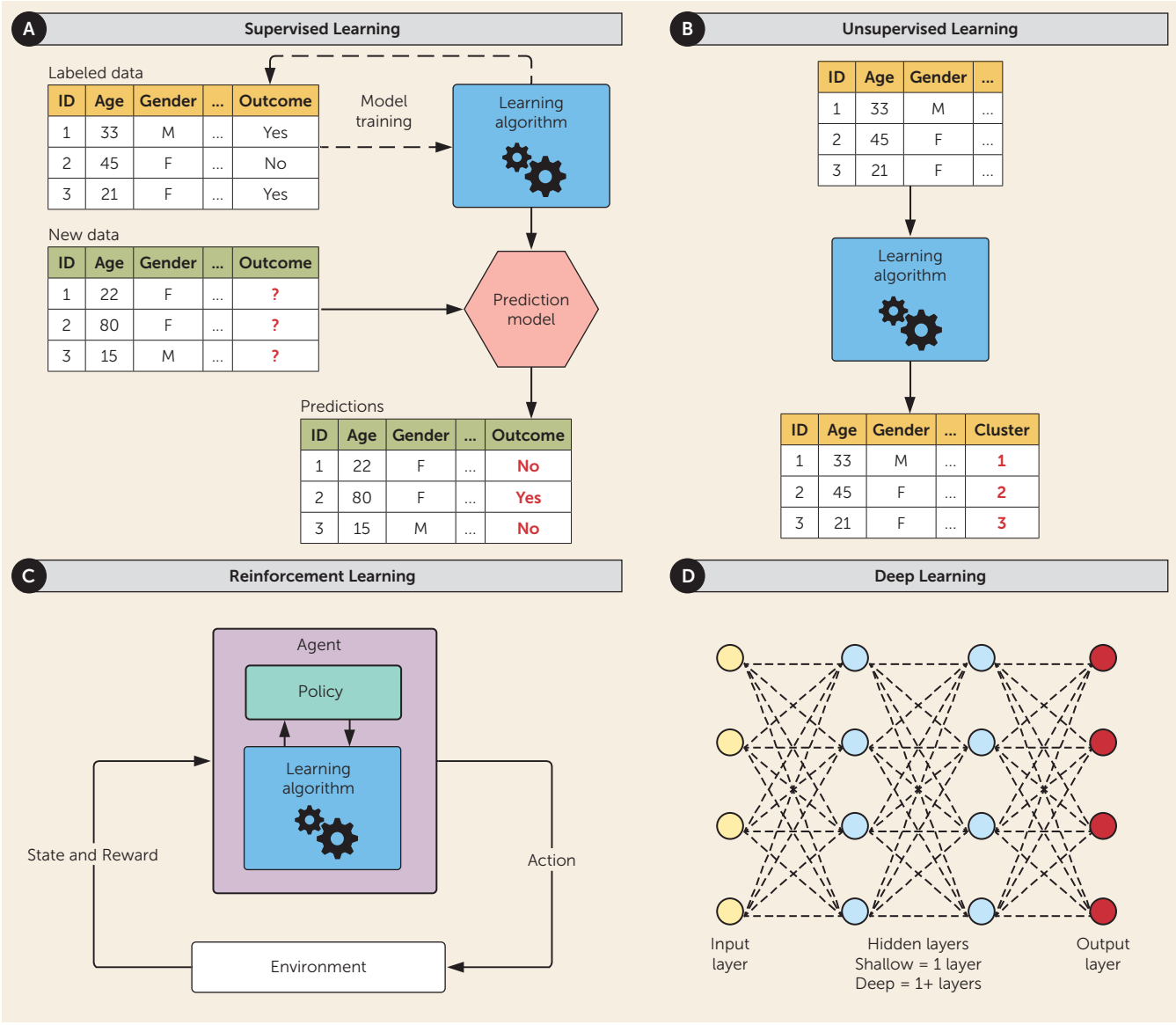
Both machine learning and the traditional statistical approaches commonly used in biomedicine (e.g., regression models) can be used to predict an unknown outcome from known data, although each differs in its capabilities (9). The focus of statistical modeling lies primarily in inference or understanding the nature and significance of the relationships between the outcome and model variables. Traditional models typically require a domain expert to select the variables of interest in advance. A small number of variables will be measured across a large sample, facilitating interpretability. By contrast, machine learning prioritizes predictive accuracy. In "shallow" data sets with many more predictive variables than observations, machine learning produces models capable of generalizing to unseen data even with relatively small sample sizes. To achieve this, machine learning sacrifices variable degrees of interpretability.

There are three general categories of machine learning:

- Supervised learning is task driven (Figure 1A). The machine learning algorithm is provided with labeled training data containing input variables (or predictors) and output variables (or outcomes) for each observation. The algorithm fits a model that estimates the function (or relationship) between the predictors and outcomes (i.e., learns by example), enabling prediction of outcome for new, unseen observations. If the outcome is binary (e.g., 0 or 1, no or yes) or categorical (e.g., multiple labels), the task is known as classification; if the outcome is a real number (e.g., a score between 0 and 27), the task is known as regression. By definition, a supervised algorithm can only be as accurate as the labels provided for training, which raises concerns in psychiatry, where diagnoses and outcomes are highly subjective. Chung et al. (10) used supervised learning to predict brain age from MRI scans; discrepancy between brain age and chronological age was associated with a higher risk for developing psychosis.
- Unsupervised learning is data driven (Figure 1B). No labeled training data are provided. The algorithm instead searches for the underlying patterns among input variables or observations. A common unsupervised learning task is clustering, in which observations are divided into subgroups according to similarity in features (11). Drysdale et al. (12) employed 33,154 functional MRI features to cluster depressed patients into groups with four distinct subtypes of dysfunctional brain connectivity, at least one of which was associated with differential response to treatment with transcranial magnetic stimulation.
- Reinforcement learning is reactive to the environment (Figure 1C). The algorithm (or agent) learns through experience ("trial and error") to make choices to maximize a notion of "reward" while completing a task (13). In a classic reinforcement-learning loop, the initial state, or combination of inputs, is set by the environment (e.g., initial position of chess pieces). The agent will take an action, and the environment will provide feedback on the reward associated with the state change. The agent updates its strategy (or policy) based on this reward and acts again. The loop repeats until the designated termination point is reached (e.g., win or lose). Reinforcement learning most closely mimics human decision making, making it an appealing framework for studying dysfunctional thought processes. Schizophrenia patients, for example, are less likely to explore alternative paths in decision making, even ones associated with better outcomes (14). Reinforcement learning is also promising in clinical decision-making support and the real-time tailoring of interventions. Shortreed et al. (15) employed reinforcement learning for ad hoc analysis of data from the Clinical Antipsychotic Trials of Intervention Effectiveness Study to determine the optimal sequence of antipsychotic medications to reduce psychosis using a low Positive and Negative Syndrome Scale score as "reward." For all subjects, the optimal sequence started with olanzapine (15).

"Deep learning" describes a collection of methods that use artificial neural networks for supervised, unsupervised, or reinforcement learning tasks (Figure 1D). An artificial neural network contains a network of nodes (neurons) with numerous interconnections for the processing of information. The artificial neurons are organized in layers with "visible" input and output layers with additional "hidden" layers of processing (16). Neurons receive information along their connections (or "edges"). Signal strength is determined by weights associated with each edge. Each neuron takes the weighted sum of its inputs, then passes this value through a mathematical function to produce the output signal to send to neighboring neurons. Different layers may perform different types of data transformations. Learning proceeds from the input to the output layer, potentially propagating through the hidden layers multiple times. A "shallow" neural network contains one hidden layer, and a "deep" neural network contains multiple hidden layers. During learning, edge weights are adjusted to reduce error in the assigned task. Deep learning is suited to solving complex,

FIGURE 1. Types of machine learning models[a]



[a] Panel A shows supervised machine learning. The algorithm is supplied with labeled training data that include predictors and the outcomes of interest. The algorithm fits a model to describe the relationship between the features and the outcomes. The trained model can then be applied to new data to make predictions on new, unseen observations. Panel B shows unsupervised machine learning. The algorithm is provided a set of unlabeled data and searches for patterns among features or observations. A common task is clustering, in which observations are subgrouped by similarity among features. Panel C shows reinforcement learning. The algorithm (or agent) learns through "trial and error" by taking actions toward completion of a task while maximizing "reward." Each action alters the environment, resulting in feedback regarding the changed state of the environment and the reward associated with that action. The loop repeats until the designated termination point is reached. Panel D shows deep learning. Artificial neural networks contain nodes (or neurons) arranged in layers with learning proceeding from the input layer to the output layer through a variable number of hidden layers. Learning may be by supervised, unsupervised, or reinforcement means. Shallow networks contain a single hidden layer, while deep networks contain more than one layer. Each node (or neuron) receives inputs (or signals) from other nodes in the network, processes the summed inputs through a mathematical scaling function, and transmits the output to the other neurons. During learning, the strength of the input signal is adjusted (weighted) at each connection to minimize error in the desired task.

nonlinear patterns, such as facial recognition. Lin et al. (17) developed a deep-learning model combining single-nucleotide polymorphisms (SNPs) and baseline clinical features to predict antidepressant response in major depressive disorder. In the original genome-wide association study, none of the 4,241,701 SNPs had been found to be significantly associated with treatment response.

DEVELOPING A PREDICTIVE MACHINE LEARNING MODEL

Understanding the steps in the development and validation of a machine learning prediction model provides a framework for understanding where errors or bias may be introduced that influence the accuracy of predictions and generalization to new data. For simplicity, the steps

below review development of a supervised classifier, but the pipeline and sources of error for other learning types are similar.

Data Collection and Cleaning

Structured data are collected using a predefined input schema and organizational framework (e.g., billing claims) (18). Structured data are typically organized in a tabular format where rows represent the unique observations (e.g., hospital encounters), and columns represent the features, or the measurable attributes of each observation (e.g., date of admission, primary diagnosis, length of stay). Unstructured data, by comparison, is everything else—text, images, video, or audio—stored in native formats with variable internal organization. An estimated 80% of all health care data are unstructured (2). Electronic health records are semistructured, containing unstructured (narrative text) as well as structured (fixed input) elements. Prior to modeling, raw data must be inspected and corrected for errors, such as inconsistencies in data labeling and removal of duplicate observations.

Algorithm Selection

Algorithm selection depends on the desired task, amount of data, and feature types. Trade-offs exist between machine learning algorithms, in sensitivity to outliers and missing data, speed, accuracy, and interpretability of results. Common supervised classification algorithms and their strengths and weaknesses are summarized in Table 1 (19).

Data-Set Splitting

The first, but most critical, step in model construction involves splitting the cleaned data set into "training" and "test" subsets (Figure 2A). The training set is used to develop the model, while the test set is withheld to evaluate the model's performance on unseen data. A model will have fitted parameters determined by the training data (e.g., regression coefficients), as well as adjustable hyperparameters (e.g., branch points in a decision tree). Using the same test set to optimize parameters and evaluate a model will result in overfitting, where the model describes the noise in the data set better than it does the underlying pattern between features. Errors in this step are the most common source of bias and error in the clinical machine learning literature. Failure to keep the test-set data completely independent of all subsequent steps leads to data leakage (discussed below), which may invalidate the results.

A key challenge is that splitting reduces the amount of training data. The performance of many machine learning algorithms is directly related to the amount of training data available. This challenge is most commonly addressed through an approach called cross-validation (20). In *k*-fold cross-validation, the data set is divided into equal *k* subsets, or folds; $k-1$ of the folds are combined to train the model, and the remaining fold is used for testing. The procedure is repeated *k* times, such that every fold serves once as a validation fold, and the results are averaged. If the data set is appropriately selected (i.e., the training and test folds are truly independent of one another), cross-validation provides a good estimate of how the model will perform on completely new data.

A closely related challenge is that many machine learning algorithms have "hyperparameters," a small set of parameters that tune the algorithm's trade-off between different sources of error. For example, a hyperparameter can control whether an algorithm seeks to predict outcomes by assigning a small amount of importance to many predictors or by heavily weighting a small number of very important predictors. Hyperparameters often need to be fine-tuned for a given data set, while also maintaining independence between training and test-set splits. Cross-validation is commonly used to address this challenge too, specifically a variant known as nested *k*-fold cross-validation. The data set is split into *k* outer folds. From each outer fold, *k* inner folds are created for feature selection, model training, and hyperparameter tuning. The outer loop of cross-validation then evaluates the performance of the best model selected from the inner loop. The cycle repeats for all outer folds and the performance metrics are averaged (Figure 2B and 2C).

Preprocessing

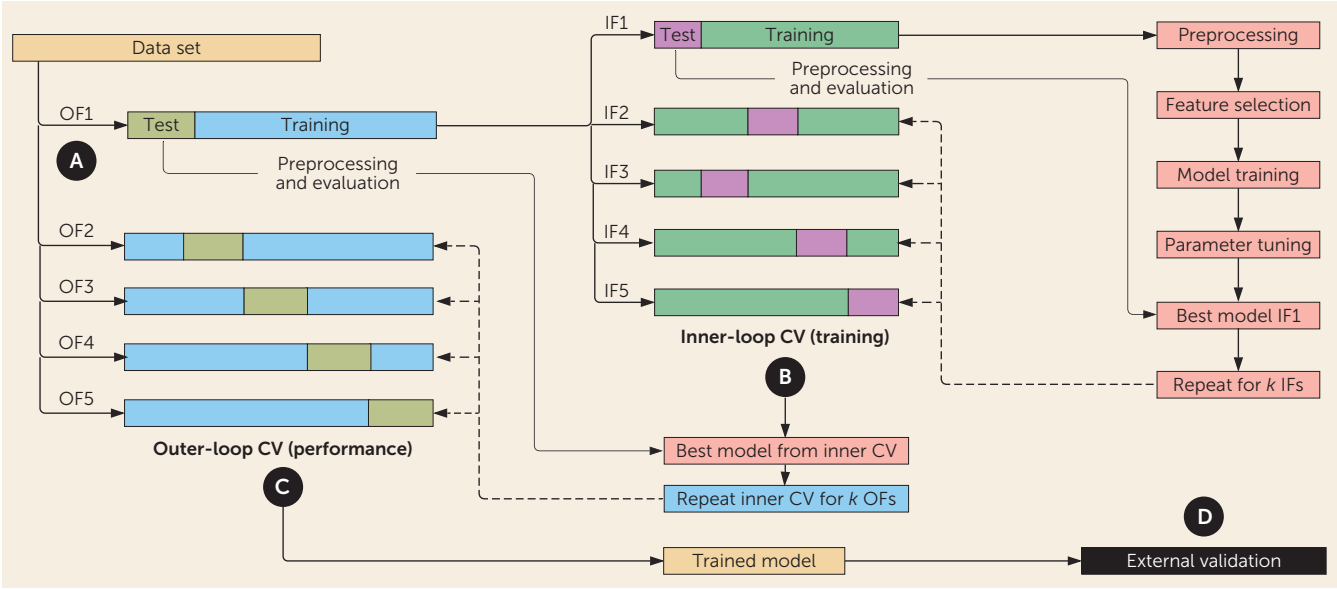
The dimensionality of a data set refers to the number of features (p) multiplied by the number of observations (n). Data sets may be wide or shallow (big p, small n), such as genomics data sets with thousands of genes measured per subject, or tall (small p, big n), such as a clinical trial. For certain machine learning algorithms, when an increasing number of features are included, computation time increases exponentially, but predictive power paradoxically decreases (known as the curse of dimensionality). On the other hand, for some algorithms, high dimensionality can be leveraged toward better performance and faster computation time (blessing of dimensionality) (21). For example, support vector machine algorithms employ a hyperparameter known as the kernel, which can transform the data from a low-dimensional to a high-dimensional space (kernel trick) to help with the linear separation of the data (decision boundary) necessary for classification (22).

For algorithms that are less tolerant of high dimensionality, the number of features needs to be reduced before model fitting, ideally without losing much of the critical information contained within the full feature set. This challenge is often overcome by feature extraction (or feature engineering)—the creation of a smaller number of new features from a larger number of initial raw features. For example, individual brain voxels in MRI scans can be averaged together to describe the signal in a smaller set of brain regions of interest (23). Dimensionality reduction is another group of techniques by which high-dimensional space is mapped to low-dimensional space without losing critical information or the variability of the original data set (24). Principal components analysis, for example, creates a small set of composite

TABLE 1. Common supervised learning algorithms for classification

Classifier	Description	Strengths	Weaknesses
Regression (logistic)	Estimates the probability of a binary outcome (yes/no or class 1/class 0) for each observation, by a linear combination of the predictors using the logistic (or sigmoid) function. If the decision boundary is set to 0.5, probabilities >0.5 will be labeled as class 1, while if <0.05, they will be labeled as class 0.	Easy to implement. Input features do not need to be scaled (no assumptions as to feature distribution). Feature selection can be regularized (to allow fitting with limited training data). Model coefficients are informative as to the relevance of a feature and direction of association (some inference).	Sensitive to outliers and multicollinearity (poor capture of complex relationships between features). Prone to overfitting with a large number of features. Requires a large amount of data.
Support vector machine	Maps each observation in *n*-dimensional feature space. The decision boundary (or hyperplane) separates the outcome classes (0 or 1) while maximizing the marginal distance between the hyperplane and support vectors (observations that rest closest to the hyperplane and therefore are the most difficult to classify).	Less prone to overfitting than logistic regression. Able to detect nonlinear relationships between features. Performs well on semistructured and unstructured data. Regularized feature selection allows good performance on limited training data. Stable to changes in training data.	Sensitive to noise (overlapping outcome classes). Computationally expensive for large data sets. Requires transformation of categorical features to binary dummies (increases dimensionality of data set) and feature scaling. May require selection of a kernel function; poor choices can greatly alter results.
Decision tree	Constructs a decision algorithm based on a series of greater-than/lesser-than comparisons. Starting at the root, the data are sequentially split based on which feature gives the highest information gain (most change in probabilities for a "greater" versus "lesser" answer). The splitting process continues until it reaches a leaf, which contains only one class of outcome labels (class 0 or 1).	Model is easily interpreted. Input does not need to be transformed (e.g., scaling, normalization). Tolerant of missing data (no imputation needed). Automatic feature selection (top branches=highest informative features).	Prone to overfitting, especially with a large number of features. Computationally expensive (increased training time and memory). Instability with even small changes in training data; may not generalize well to new data.
Random forest	Multiple decision trees are created, and the outcome class (0 or 1) is determined by a majority vote of the generated decision trees. Each tree is built from a random subset of the data, reducing the influence of any one feature or data point on the outcome.	Less prone to overfitting than decision trees. Random subsets (ensemble training) make the resulting classifier more generalizable.	Computationally more expensive than decision trees. Often very difficult to interpret because features appear at different levels in individual trees.
k-nearest neighbor	Prediction of outcome class (0 or 1) is based on whether the majority of points that are "near" the new example (in a derived feature space) are from class 0 or 1. *k* is the number of neighbors considered in the majority vote.	Not sensitive to noise or missing data. No underlying assumption about data distribution.	Computationally expensive with increasing number of features. All features given equal importance. Sensitive to outliers. May require feature transformations that distort data. Selection of *k* can greatly influence results. May generalize poorly to new data.
Gaussian naive Bayes	Assumes all predictors are independent and equally important to prediction of the outcome class (0 or 1). Outcome class is determined by the highest posterior probability, a function of the prior probability of a class (class distribution, Gaussian) and the likelihood, or the probability, of a feature given a class.	Scales well to large data sets. Requires less training data. Robust to outliers. Ignores missing values.	Dependency between attributes negatively affects performance. Assumes all features are Gaussian (normally) distributed, which is often not true for clinical variables.
Artificial neural network	Each node (or neuron) receives inputs (or signals) from other nodes in the network, processes the summed inputs through a mathematical scaling function, and transmits the output to the other neurons. During the training process, the strength of the input signal is adjusted (weighted) at each connection to minimize error in predicting the outcome class (0 or 1).	Tolerates nonlinear relationships between features and can use these relationships to improve performance. Can employ multiple learning algorithms. Can represent history-sensitive situations where a predictor's importance depends on what came just before.	Computationally expensive to train, although can be made efficient to apply. Can be very sensitive to how features are preprocessed and extracted; computational expense makes it difficult to explore many options. Largely a "black box," very difficult to understand what influences predictions. Can be very vulnerable to small, even apparently meaningless changes in input data.

FIGURE 2. Development of a machine learning classification model using nested cross-validation[a]



[a] Part A shows data set splitting. The data set is split into *k* outer folds (here, *k*=5). Each fold is divided into a training and a test subset. Part B shows the inner-loop cross-validation (CV). Each outer-fold training data subset is then subdivided further into *k* inner folds (here, *k*=5). The inner folds are used for model training, feature selection, and hyperparameter optimization. The best model from the inner loop of cross-validation is applied to the held-out test of the outer fold to determine model performance. Part C shows the outer-loop cross-validation. The inner loop of cross-validation is repeated for each outer loop, and results are averaged to determine performance. Part D shows external validation. Ideally, the trained model is applied to an unrelated data set (external validation) to further assess performance on unseen data. IF=inner fold; OF=outer fold.

features (principal components) through linear combination of the original features. Feature reduction additionally reduces the computational expense, or the time and computing memory/processing consumption needed to build the model.

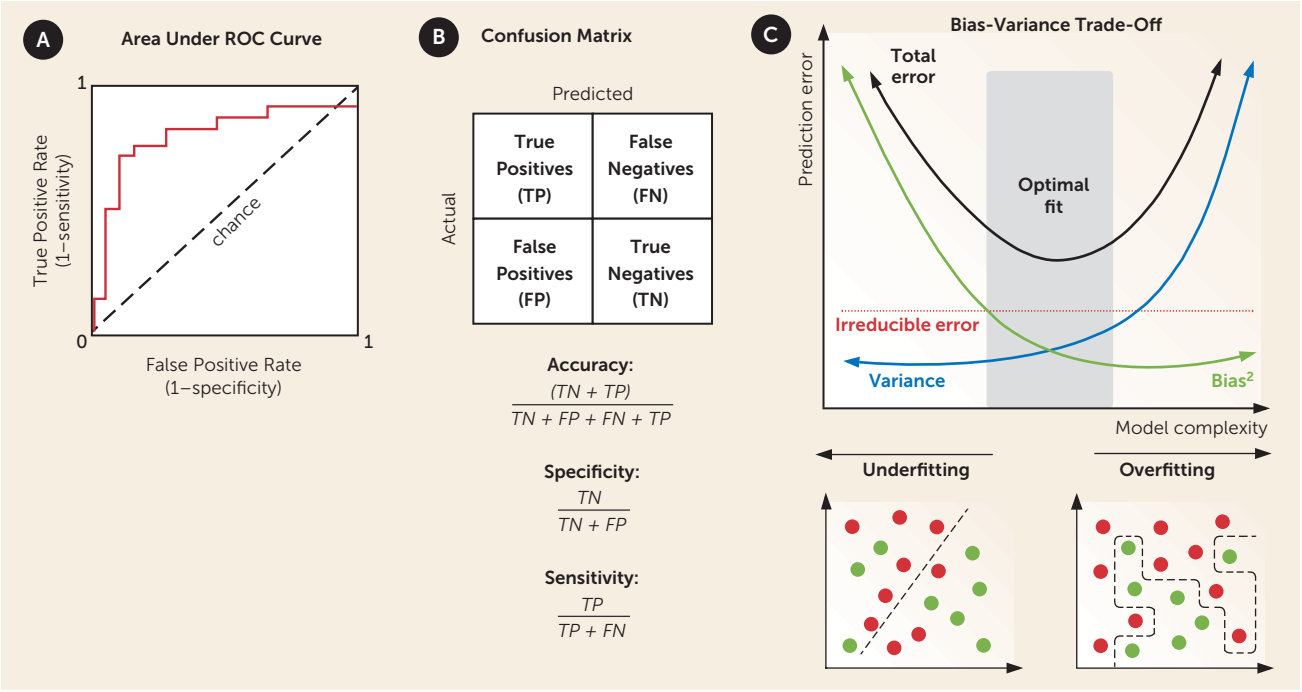
In cases where the best features are not known in advance, feature-selection algorithms can automatically choose an informative set of features for inclusion in a model (25). Filter methods use univariate statistics (e.g., chi-square test, correlation coefficients) to determine whether individual features are associated with the outcome prior to combination in a multifeature model. Wrapper methods remove or add features based on model performance (e.g., recursive feature elimination, sequential feature selection). Embedded methods incorporate feature selection into the learning algorithm. A common example is regularization, in which the algorithm penalizes its internal performance metrics as the number of included features grows. In lasso regression, the coefficients of less informative features are forced to zero, resulting in feature removal. In ridge regression, the coefficients of less informative features shrink toward zero, but all features remain in the model. Elastic net regression combines the lasso and ridge approaches—some features are eliminated, while others remain but are penalized.

Feature transformation involves manipulating feature attributes for optimal modeling, such as altering the range or distribution of features. This is especially important when using machine learning algorithms that employ distance-based metrics (e.g., support vector machines, k-nearest neighbor). A feature scaled from 0 to 100 may receive higher weight compared with one that ranges from 0 to 1. If an algorithm assumes a normal distribution (e.g., linear regression), but the features are not normally distributed, they must be transformed to meet the assumptions. Nonnumerical values (e.g., category labels) must be converted to representative integers (e.g., 0=low, 1=medium, 2=high) and potentially further broken down into individual binary dummy representations, depending on the algorithm used. Managing missing data is another critical feature transformation task. Reasons for missing data must be carefully analyzed to determine whether a systematic relationship exists between the feature and its tendency toward missingness (e.g., data from patients with higher illness severity are more likely to be missing). Deletion of observations missing one or more feature values (complete case analysis) is rarely advised, because it is likely to introduce bias. Estimation of the missing values from other information in the data set, known as imputation, is preferred (26).

Model Selection

Model selection is both iterative and heuristic. A model must balance bias (differences between predicted and correct values) and variance (variability of prediction for a given value), which contribute to the total error of a model. A

FIGURE 3. Evaluation of a machine learning classification model[a]



[a] Model performance can be visualized by a confusion matrix or area under the receiver operating characteristic curve (AUC) plot. Panel A shows the AUC plot, which demonstrates the ability of a model to discriminate between labels and ranges from 0.5 (chance) to 1 (perfect). Panel B shows the confusion matrix, which summarizes the number of true positive (TP), false negative (FN), false positive (FP), and true negative (TN) predictions. Accuracy is defined as the number of correctly predicted labels to all classifications. Specificity is the true negative rate, and sensitivity is the true positive rate. Panel C illustrates model fit. The optimal fit for a model balances model complexity and prediction error, or the trade-off between bias and variance. Bias errors are errors made on the part of the learning algorithm. If the bias is too high, the model will be prone to underfitting, or unable to find a meaningful relationship between the predictors and the outcome of interest. Variance error reflects those errors made as a result of variability in the training set. If the variance is too high, the model will be prone to overfitting, or capturing the noise in the data set rather than the true relationship between predictors and outcome. All models contain some degree of noise (or irreducible error).

high-bias model will tend toward underfitting, while a high-variance model will tend toward overfitting (Figure 3). Feature selection, cross-validation, and hyperparameter tuning assist in addressing the bias-variance trade-off to produce an optimal model. The numerous methods available for tuning hyperparameters are beyond the scope of this primer but were previously reviewed by Hutter et al. (27). The most common methods include grid search, where all values in a predefined set are tested for each hyperparameter, and random search, where a range of values for each hyperparameter are randomly sampled and tested to provide "good enough" coverage. A key point is that for successful cross-validation, both the hyperparameters and the preprocessing method must be developed independently on the training set and then applied to the test set without change (and without any prior testing on the same test set).

Performance Evaluation

There are numerous metrics for scoring models during the selection (inner cross-validation) and testing (outer cross-validation) phases. Accuracy is the proportion of correct classifications out of all classifications (28). A classifier is evaluated by computing the true positive, true negative, false positive, and false negative rates, which are commonly presented as a 2×2 table known as the confusion matrix (Figure 3A). From this information, the sensitivity and specificity of the classifier can be calculated. Confusion matrix data may also be visualized as an area under the receiver operating characteristic curve (AUC) plot, which demonstrates the power of a classifier in discriminating one class from another (Figure 3B). An AUC of 1 indicates perfect precision, while 0.5 is equivalent to chance. The confusion matrix will also highlight issues with class imbalance, a situation where most of the data set does not contain the outcome of interest. Imbalance is a common problem in psychiatry, as we are often interested in predicting a rare event (e.g., suicide completion), resulting in poor model performance (29). In general, all of these metrics should be reported as part of a machine learning analysis, as each conveys different information about a classifier's performance and bias-variance trade-offs (Figure 3C).

External Validation

In external validation (Figure 2D), the final model is applied to a data set entirely unrelated to the data set used in model construction, perhaps gathered from a different clinical site

or period in time. If the model's accuracy declines significantly, overfitting or underfitting is the most likely culprit. Another cause may be related to a phenomenon known as distributional shift, where the underlying process generating the data changes over time or between data sets. For example, if new therapies become available, clinicians may become more likely to make the diagnoses for which those therapies are indicated. This creates a change in the diagnostic mix between the original training data and new incoming clinical data, leading to inaccurate model predictions. Data-set shift can also result from differences in instrumentation, differing patient cohorts, and changing illness prevalence over time (30).

Natural Language Processing

Natural language processing rests at the intersection of artificial intelligence, machine learning, and linguistics, exploring how computers deconstruct, understand, and manipulate natural language toward a variety of problems (31). While this article does not aim to offer a scoping review of this expanding field, understanding its foundation is critical to an overview of the machine learning mental health landscape. Speech recognition and foreign language translation are common applications of natural language processing. The ambiguity and complexity of language make natural language processing among the more difficult tasks for computers to master. In health care, natural language processing has permitted increased use of narrative electronic health record data, historically underutilized because of the labor intensity of manual curation, and enabled insights from novel text-based data sources, such as social media posts (32, 33).

Natural language processing relies heavily on machine learning and deep learning, although it can also be combined with symbolic approaches (34). Natural language processing tasks are generally subdivided into syntactic and semantic analyses (35). In syntactic analysis, a corpus (collection of texts) is systematically parsed and analyzed using formal grammar rules, analogous to the grade-school exercise of constructing sentence diagrams that label the parts of speech (e.g., noun, verb, adjective, article). Semantic analysis explores the relationships between syntactic structures to extract units of meaning, such as whether a corpus contains more negative versus positive emotional expression. Semantic analysis may be applied at any level, from sentences to phrases to whole documents. In a study of suicide risk among veterans, Westgate et al. (36) applied syntactic analysis and found a higher degree of distancing language, such as the use of third-person pronouns and terms (e.g., he, she, vet), in the mental health notes of veterans who would attempt suicide in the following year compared with their nonsuicidal peers.

Development of natural language processing follows a workflow similar to that of the machine learning predictive models covered previously (Figure 4). A raw corpus (or collection of texts) must be cleaned to enable efficient and accurate feature extraction. Cleaning of raw text can include removal of templated text (e.g., standard headers or logos), correction of spelling errors, and removing formatting (e.g., capitalization, punctuation). As part of preprocessing, the cleaned text will be digested into tokens (e.g., words or short phrases). Common words that rarely convey meaning, also known as stop words (e.g., a, the), will often be removed. Tokens may be further transformed, such as by the removal of word endings or stemming (e.g., walk from walking, walked, or walks).

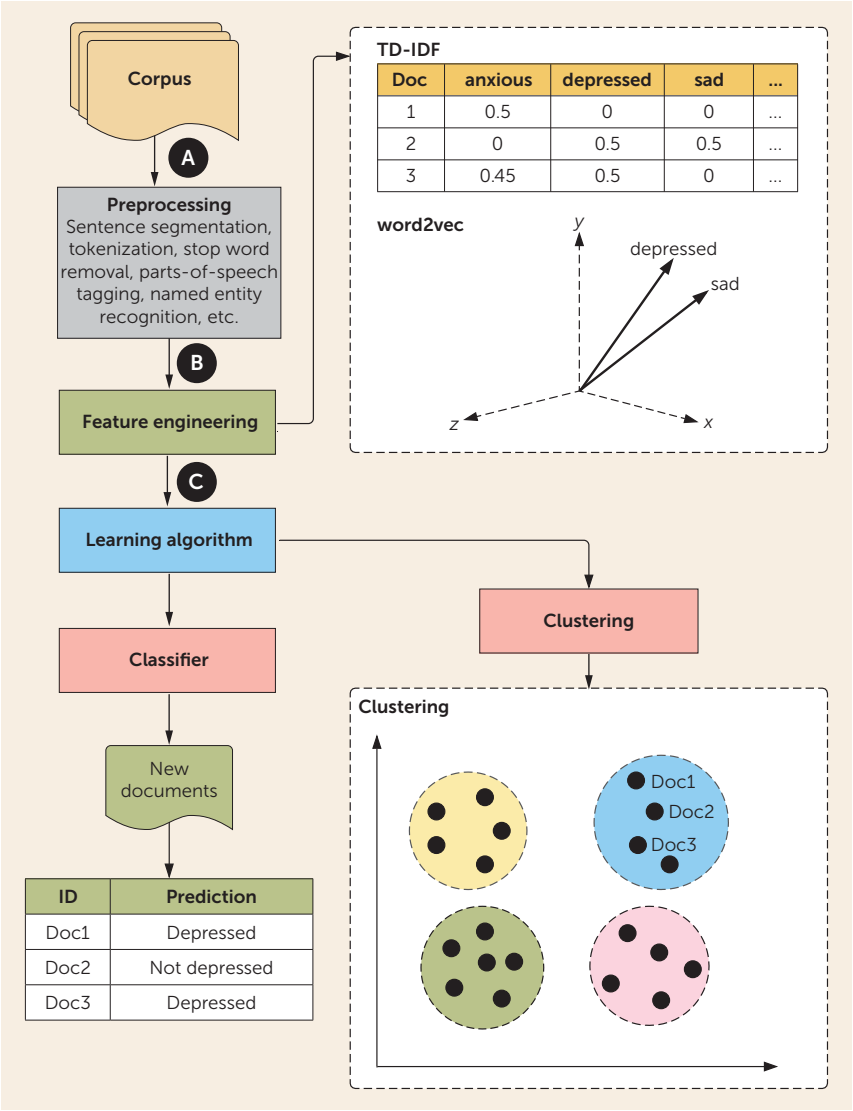
Feature engineering in natural language processing involves the transformation of texts and tokens into numerical representations or word embeddings. As with other machine learning approaches, many methods are available, but a few are particularly popular. The term frequency-inverse document frequency (TF-IDF) method, for example, converts the corpus into a matrix where the relative importance of each token is calculated as a metric that normalizes for its frequency within and between documents. Rare words will have higher TD-IDF values (37). TF-IDF vectorization captures no additional contextual information for words, which can lead to misinterpretations (e.g., if two different clinicians use different words to describe the same symptom). Word2-vec, by comparison, uses a shallow neural network to generate multidimensional vectors for each token that can preserve context with similar tokens sharing vector space (38). Engineered features may then be used in combination with machine learning algorithms for outcome prediction and clustering of related documents, among other tasks (39, Table 2).

PITFALLS

Machine learning is not a panacea for every challenge in psychiatry (40). Decisions in model development and validation can significantly affect generalizability to new data. For example, machine learning-based prediction of antidepressant response using quantitative EEG (QEEG) data is widely reported and commercially marketed. However, a meta-analysis of 81 QEEG biomarkers derived from 76 studies found that none of these classifiers had been externally validated, that most were created using small samples, and that the literature as a whole suffered publication bias (41). The authors concluded that these methodological issues made QEEG markers unsuitable for routine clinical use until larger validation studies are performed. Table 2 offers a framework for evaluating the machine learning/natural language processing literature with several examples that highlight common potential pitfalls to bear in mind, which include the following:

1. Data quality ("garbage in, garbage out"). Model accuracy and validity are highly dependent on the availability of large amounts of high-quality training data, which can be expensive and time-consuming to generate. Small

FIGURE 4. Development of machine learning models using natural language processing[a]



[a] Part A is the preprocessing stage. A collection of texts is known as a corpus, which requires preprocessing and digestion into component subunits (words or short phrases), known as tokens for natural language-processing tasks. Part B is feature engineering. Including all words from a corpus as features in a machine learning model would be computationally expensive, and therefore texts and tokens must be converted into numerical representations or word embeddings. In the term frequency-inverse document frequency (TF-IDF) method, a statistic is calculated for each token to reflect its importance across the corpus. Higher values are associated with rarity of the token. These values are arranged in a sparse matrix in which each row is an individual document and each column the score for an individual token within that document. Another method is word2vec, or the use of shallow neural networks to map each token within each document to a multidimensional vector. Part C is model creation. Engineered features may be used with any variety of supervised or unsupervised classification algorithms for prediction or clustering.

training data sets may yield highly inaccurate predictions when applied to new data. Neuroimaging studies, which often use convenience cases from a single academic medical center, are often critiqued in this regard (42). Disparities in demographic coverage, geographic location, clinical setting, and collection methodology (e.g., proprietary algorithms, different instruments or software) can also negatively influence the generalizability of the training data to the general population and affect real-world model performance (30). Significant care must be taken to determine the quantitative and qualitative equivalence of shared features across time, location, and instrumentation when integrating data sets from different sites or collection protocols (43). Electronic health record and claims data, while both wide and tall, were not collected for research purposes and frequently contain inconsistencies and missing data (44). Clinical practice involves the stratification of patients for referral, treatment, and/or documentation. As such, electronic health record cohorts of a particular diagnosis are unlikely to be true randomly

TABLE 2. Considerations in evaluating a machine learning study and examples of studies[a]

Model features	Considerations	Chekroud et al. (66)	Kessler et al. (67)	Rumshisky et al. (68)
Prediction	Does the prediction have clinical utility? How can the results be used in practice?	Remission of depressive symptoms in response to 12 weeks of citalopram treatment (final Quick Inventory of Depressive Symptomatology score $\leq$5)	Suicide death in 12 months following outpatient mental health visit	30-day psychiatric readmission
Data set	Single or multisite recruitment? Any data collection considerations (e.g., equipment, differing measures)?	Sequenced Treatment Alternatives to Relieve Depression Study, across six primary care sites and nine psychiatric care sites	Historical administrative data system of the Army Study to Assess Risk and Resilience in Servicemembers, 2004–2009	Partners HealthCare electronic health records, including academic and community hospital and clinics in New England, 1994–2012
Subjects	Is this a representative patient population? Are there adequate data for the proposed analysis? Inclusion/exclusion criteria of subjects?	N=1,949; 18- to 75-year-old outpatients with nonpsychotic major depressive disorder and score $\geq$14 on the Hamilton Depression Rating Scale, 2001–2004	N=975,057; male, nondeployed regular U.S. Army soldiers	4,687 patients with inpatient discharge summaries; $\geq$18 years old, with a diagnosis of major depressive disorder; no additional exclusion criteria
Class balance	Is class imbalance present? How is this handled in the analysis?	No class imbalance reported; 51.3% of subjects were nonresponders	569 deaths by suicide with >8,000 control visits per suicide; probability sample of control visits used	470 patients were readmitted within 30 days; no class imbalance correction
Input features	Do feature extraction methods appropriately capture the desired signal? Are included features easily obtained in routine practice? Are the features appropriate to the prediction? Any sources of data leakage?	164 features, including sociodemographic features, DSM-IV-based diagnostic items, depressive severity checklists, eating disorder diagnoses, prior antidepressant history, the number and age at onset of previous major depressive episodes, and first 100 items of the Psychiatric Diagnostic Symptom Questionnaire	Nearly 1,000 features, including outpatient visit clinical factors, prior clinical factors, Army career, prior crime, and contextual factors	Baseline clinical features: age, gender, use of public insurance, and age-adjusted Charlson comorbidity index score; 75 topics extracted by latent Dirichlet allocation from full corpus; top 1,000 words extracted by term frequency-inverse document frequency for each patient
Algorithm	Was the use of a particular algorithm (or algorithms) over others justified? Were other algorithms evaluated and reported? Is the algorithm appropriate for the data and/or problem?	Gradient-boosting machine (ensemble of decision trees); no other algorithms were reported	Naive Bayes, random forest, support vector regression, and elastic net penalized regression were tested	Support vector machine; no other algorithms were reported
Data splitting and resampling	Were cross-validation or other resampling methods used? Were these performed appropriately? Any sources of data leakage?	Ten-fold cross-validation	Cross-validation (type not reported); separate models for suicides with and without prior psychiatric hospitalization	Data set randomly split into training (70%) and testing (30%) data sets; balanced by clinical features; separate models for baseline clinical features, baseline plus 1,000 words, baseline plus 75 latent Dirichlet allocation topics
Imputation	How is missing data handled?	Complete cases (patients with missing data dropped)	Missing data corrected by nearest neighbor or rational imputation	NA
Feature selection	How were features selected? How many features survived?	Elastic net regularization to select top 25 features prior to model building	Univariate association of predictor of suicide compared with other death; significant univariate predictors plus 20 sociodemographic variables and 27 Army-career variables passed to machine learning classifiers; penalized regression for selection in final models	None

continued

TABLE 2. *continued*

Model features	Considerations	Chekroud et al. (66)	Kessler et al. (67)	Rumshisky et al. (68)
Model selection	What metric was used to determine optimal performance (accuracy, AUC, custom metric)? Could this metric bias model selection?	Maximization of AUC	Maximized cross-validated sensitivity in the 5% of visits with the highest predicted suicide risk	Maximization of AUC
Hyperparameter optimization	Any hyperparameters? What metric was used for their evaluation? Was a separate data subset used for hyperparameter optimization?	Same criterion as for model selection	Same criterion as for model selection	Threefold cross-validation on the training data
Performance	Any evidence of overfitting (are the results "too good to be true")? Are the results and proposed model believable? How portable is the model to other contexts? Were any attempts made at model simplification?	AUC=0.70	Elastic net classifier with 10–14 predictors optimized sensitivity; AUC=0.72 (prior hospitalization), 0.61 (no prior hospitalization), and 0.66 (combined) within 26 weeks after visit	Baseline clinical features (AUC=0.618), baseline clinical features plus 1,000 words (AUC=0.682), and baseline clinical features plus 75 latent Dirichlet allocation topics (AUC=0.784)
External validation	Was the model externally validated? Did performance drop significantly in application to new data? If so, is the model still clinically useful? Were reasons for the change in performance explained? Were there any potential hidden confounders or time effects affecting model performance?	Yes; validated in escitalopram treatment group (N=151) of the Combining Medications to Enhance Depression Outcomes trial (accuracy, 59.6%)	Yes; validated by using 2004–2007 data to predict 2008–2009 deaths by suicide; combined AUC (those with and without prior hospitalization) was 0.67–0.72 within 26–5 weeks after visit	No

[a] AUC=area under the receiver operating characteristic curve; NA=not applicable.

selected samples (sampling bias) or reflective of the underlying general population (sampling error) (45). For the purposes of comparing treatments, for example, electronic health record cohorts are highly nonrandomized. A patient receiving one medication versus another is determined by multiple prior decisions (e.g., prior failed medication trials, medical comorbidities).

2. Biased data. The ethical questions surrounding the incorporation of machine learning in health care are numerous, particularly in regard to health equity (46). Racial bias is found in existing health care algorithms, often the result of confounding factors that have not been accounted for. Obermeyer et al. (47) found that a widely used risk algorithm underestimated service needs in Black patients by more than half because it used health care costs as a proxy for illness severity (47). In psychiatry, Black men are disproportionally diagnosed with schizophrenia, even though psychosis symptoms are features in numerous other disorders, including depression (48). A diagnostic model that does not account for flaws in its training data risks propagating bias and reinforcing disparate clinical care. Similar concerns apply to artificial intelligence and machine learning systems in violence risk assessment (49).
3. Algorithm selection ("no free lunch"). Not all machine learning algorithms are created equal, although many behave comparably depending on the data set. Jiao et al. (50) tested four different algorithms to predict autism spectrum disorder diagnosis, using regional cortical thicknesses on structural MRI. The support vector machine classifier required all 66 features, whereas the tree-based classifier only required seven features for similar accuracy. Algorithm selection is often arbitrary or biased toward preliminary performance results. A potential solution is ensemble learning, a meta-algorithmic technique that combines multiple machine learning algorithms to generate a single optimized model (51).
4. Data leakage. One of the most pervasive errors in clinical machine learning occurs when information is inadvertently shared between the training and test data, or data leakage (52). Because a portion of the data is "known," the trained model will perform exceedingly well in testing, like a student who knows the test questions in advance. This leads to "too good to be true" accuracy that drops precipitously when applied to new data as a result of overfitting. A common culprit is preprocessing of the entire data set prior to splitting into training and testing sets. Information about the test data (e.g., its statistical distribution properties) "leaks" into the training set via the preprocessing. Care must also be taken in the cross-validation of time series data. If the time points are simply randomly shuffled, then the model can use data from the future to predict the past. Finally, certain features may also introduce leakage, such as including information about prior prescriptions in a model seeking to predict a patient's new diagnosis. Natural language processing models that use narrative-free electronic health record text can be prone to this type of leakage (e.g., a history of present illness containing information about prior treatments).

5. Model interpretability. Machine learning models, especially deep-learning models, are often criticized for being opaque, "black box" solutions. The meaning of model parameters and feature relationships can be difficult to determine, and in cases where a model errs, it is difficult to determine why. Dramatic examples outside of medicine include "single pixel attacks" in image processing, where the output of a high-stakes task (e.g., determining where the road is for a self-driving car) can be dramatically altered by a visually undetectable change to the input image. Interpretability is an active area of investigation. For example, local interpretable model-agnostic explanation (LIME) is an algorithm that will introduce a small amount of noise into each model feature and then determine how predictions were changed by this perturbation. The predictions are then used to generate a linear model, the coefficients of which can be used for interpretation (53). The ability to explain why a model makes a prediction affects trust in implementation, although tolerance of a "black box" solution may depend on the task. If the aim is to flag anomalies in neuroimaging, sensitivity may be reasonably prioritized over interpretability, whereas in treatment selection, understanding how the recommendation was generated may affect the discussion of alternatives.
6. Model generalizability. External validation of trained models is essential to estimating the reproducibility of the model's behavior in real-world conditions but is rarely done (54). For example, when Dinga et al. (55) attempted to replicate the functional MRI depression subtypes found by Drysdale et al. (12) in an independent sample using an identical analytical pipeline, they could not replicate the key outcome (55). In contrast, several recent schizophrenia investigations have demonstrated the effectiveness of pooled data sets and site-based cross-validation methods in improving the predictive stability of models (56, 57). A related notion is model "portability," or how easily a model can be deployed in other contexts, which is dependent on the number and type of features needed for predictions. If a model requires numerous, expensive, or time-consuming features (e.g., whole-genome sequencing) for adequate performance, clinical utility may be limited.
7. Methodological transparency. The degree to which studies share training data and disclose decisions made in model development varies widely, significantly affecting reproducibility. As open-source and validated machine learning packages and pipelines become more readily available, the lack of transparency in methodological reporting is no longer justifiable. Model transparency includes how the overall model operates, the effects of individual parameters, and the algorithm's learning strategy. User-friendly "automated" machine learning software is emerging, lowering the learning curve for entry into the field but potentially raising risk for studies with methodological flaws. Several multidisciplinary panels have recently proposed guidelines for best practices in reporting machine learning studies (58, 59).

REVISITING THE CLINICAL SCENARIO

In a recent global survey of 791 psychiatrists, only 3.8% of respondents indicated concern that an artificial intelligence system could replace clinicians in the future, but roughly half expressed certainty that artificial intelligence would transform their work (60). Respondents predicted benefits, including reduced administrative burden, continuous monitoring through wearable technologies, and improved accuracy in diagnosis. In the case of Mr. A in the above clinical scenario, the clinical decision support system employed a deep-learning algorithm with information extracted from multiple sources within the clinical record, including past prescription records and free text notes. Trained on data from a wide variety of practice types and geographical locations, the system underwent a multicenter prospective trial that found, among other outcomes, a lower rate of discontinuation of system-recommended medications in cases of high diagnostic uncertainty.

There are numerous ethical, legal, and philosophical challenges to artificial intelligence implementation in health care (61). Deployment in real-world practice will continually challenge artificial intelligence and machine learning systems with unexpected and ambiguous cases. When presented with an anomalous case, humans "err on the side of caution" and attempt to minimize negative consequences. Machine learning systems, however, prioritize performance. Fail-safes, conditions under which a system refuses to act, are necessary (62). Automation complacency, in which clinicians fail to question inconsistent results, must also be addressed (63). Currently, the clinician is the primary responsible party for negligence, even if a negative outcome is based on malfunctioning software. As artificial intelligence becomes increasingly embedded in devices and clinical workflows, legal standards regarding the fiduciary relationship between patient and clinician and malpractice will require revision (64). The U.S. Food and Drug Administration "software as medical device" regulations were not designed for real-time adaptive technologies, an issue acknowledged in a recent revision (65).

CONCLUSIONS

Psychiatry is undergoing a paradigm shift. The heavily biologic approach of recent decades has produced a wealth of specific knowledge but few actionable insights for patients. Machine learning coupled with big data demonstrates promise for eventually generating applications and predictive models of high clinical diagnostic and prognostic utility. The trustworthiness and effectiveness of such applications in real-world practice, however, will necessitate methodological transparency and vigilant scrutiny of results and applications. Best practices for the design and evaluation of machine learning studies are still evolving. However, many of

the pitfalls reviewed here appear disturbingly common in current machine learning studies, most notably failure to use cross-validation, data leakage, and lack of external validation. We feel that this may be indicative of developments in the field outpacing editor or peer-reviewer knowledge. For clinicians, interrogating the most critical elements of the machine learning model development pipeline, as reviewed here, provides a preliminary framework for navigating the literature. Formal, systematic reviews that grade the quality of evidence in the extant literature are currently lacking and represent a critical gap for future investigation.

AUTHOR AND ARTICLE INFORMATION

Department of Psychiatry and Biobehavioral Sciences, David Geffen School of Medicine, University of California, Los Angeles, and Olive View-UCLA Medical Center, Sylmar (Grzenda); Department of Psychiatry and Behavioral Neurobiology, University of Alabama at Birmingham (Kraguljac); Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta (McDonald); Department of Psychiatry, University of Texas Dell Medical School, Austin (Nemeroff); Department of Psychiatry, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston (Torous); Department of Psychiatry and Behavioral Sciences, Albert Einstein School of Medicine, Bronx, N.Y. (Alpert); Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, Calif., and Veterans Affairs Palo Alto Health Care System, Palo Alto, Calif. (Rodriguez); Department of Psychiatry and Behavioral Sciences, University of Minnesota, Minneapolis (Widge).

Drs. Rodriguez and Widge are co-last authors.

Send correspondence to Dr. Widge (awidge@umn.edu).

Dr. Grzenda has received funding from an APA Foundation Research Fellowship. Dr. McDonald has received research grant support from Cervel Neurotherapeutics, the National Institute on Aging, NIMH, the National Institute of Neurological Disease and Stroke, NeoSync, Neuronetics, the Patient-Centered Outcomes Research Institute, Soterix, and the Stanley Foundation; he has a contract with Oxford University Press to co-edit a book on transcranial magnetic stimulation in depression; and he serves as a consultant for Signant Health. Dr. Nemeroff has received grants or research support from NIH and the Stanley Medical Research Institute; he has served as a consultant for Bracket (Clintara), Dainippon Pharma, Fortress Biotech, Intra-Cellular Therapies, Janssen Research and Development, Magstim, Prismic Pharmaceuticals, Sumitomo Navitor Pharmaceuticals, Sunovion, Taisho Pharmaceutical, Takeda, TC MSO, and Xhale; he has served on scientific advisory boards for the American Foundation for Suicide Prevention (AFSP), the Anxiety Disorders Association of America (ADAA), Bracket (Clintara), the Brain and Behavior Research Foundation, the Laureate Institute for Brain Research, Skyland Trail, and Xhale and on directorial boards for ADAA, AFSP, and Gratitude America; he is a shareholder in AbbVie, Antares, BI Gen Holdings, Celgene, Corcept Therapeutics, OPKO Health, Seattle Genetics, and Xhale; he receives income or has equity of $10,000 or more from American Psychiatric Association Publishing, Bracket (Clintara), CME Outfitters, Intra-Cellular Therapies, Magstim, Takeda, and Xhale; and he holds patents on a method and devices for transdermal delivery of lithium (patent 6,375,990B1) and a method of assessing antidepressant drug therapy via transport inhibition of monoamine neurotransmitters by ex vivo assay (patent 7,148,027B2). Dr. Torous has received research support from Otsuka. Dr. Rodriguez has served as a consultant for Epiodyne; she receives research grant support from Biohaven Pharmaceuticals; and she receives a stipend from APA Publishing for her role as Deputy Editor of the *American Journal of Psychiatry*. Dr. Widge has received consulting fees from Circuit Therapeutics, Dandelion Health, LivaNova, and Medtronic; and he is an inventor on multiple patent applications and granted patents related to data analytic techniques in neuroscience and psychiatry. The other authors report no financial relationships with commercial interests.

Supported in part by NIH grants UH3NS100548 and R21MH120785.

The authors thank Diana Clarke, Ph.D., of APA, for critical administrative and technical assistance.

The views and opinions expressed in this article are solely those of the authors and do not reflect the official policy of the U.S. Department of Health and Human Services or any other federal agency.

Received March 4, 2020; revisions received September 3 and October 23, 2020; accepted December 7, 2020; published online June 3, 2021.

Am J Psychiatry 2021; 178:715–729; doi: 10.1176/appi.ajp.2020.20030250

REFERENCES

1. Weissman MM: Big data begin in psychiatry. JAMA Psychiatry 2020; 77:967–973
2. Monteith S, Glenn T, Geddes J, et al: Big data are coming to psychiatry: a general introduction. Int J Bipolar Disord 2015; 3:21
3. Torous J, Onnela JP, Keshavan M: New dimensions and new tools to realize the potential of RDoC: digital phenotyping via smartphones and connected devices. Transl Psychiatry 2017; 7:e1053–e1053
4. Wongkoblap A, Vadillo MA, Curcin V: Researching mental health disorders in the era of social media: systematic review. J Med Internet Res 2017; 19:e228
5. Rutledge RB, Chekroud AM, Huys QJM: Machine learning and big data in psychiatry: toward clinical applications. Curr Opin Neurobiol 2019; 55:152–159
6. Liu X, Keane PA, Denniston AK: Time to regenerate: the doctor in the age of artificial intelligence. J R Soc Med 2018; 111:113–116
7. Wooldridge M, Jennings NR: Intelligent agents: theory and practice. Knowl Eng Rev 2009; 10:115–152
8. Garnelo M, Shanahan M: Reconciling deep learning with symbolic artificial intelligence: representing objects and relations. Curr Opin Behav Sci 2019; 29:17–23
9. Bzdok D, Altman N, Krzywinski M: Statistics versus machine learning. Nat Methods 2018; 15:233–234
10. Chung Y, Addington J, Bearden CE, et al: Use of machine learning to determine deviance in neuroanatomical maturity associated with future psychosis in youths at clinically high risk. JAMA Psychiatry 2018; 75:960–968
11. Saxena A, Prasad M, Gupta A, et al: A review of clustering techniques and developments. Neurocomputing 2017; 267:664–681
12. Drysdale AT, Grosenick L, Downar J, et al: Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nat Med 2017; 23:28–38
13. Gosavi A: Reinforcement learning: a tutorial survey and recent advances. Informs J Comput 2009; 21:178–192
14. Strauss GP, Frank MJ, Waltz JA, et al: Deficits in positive reinforcement learning and uncertainty-driven exploration are associated with distinct aspects of negative symptoms in schizophrenia. Biol Psychiatry 2011; 69:424–431
15. Shortreed SM, Laber E, Lizotte DJ, et al: Informing sequential clinical decision-making through reinforcement learning: an empirical study. Mach Learn 2011; 84:109–136
16. Koppe G, Meyer-Lindenberg A, Durstewitz D: Deep learning for small and big data in psychiatry. Neuropsychopharmacology 2021; 46:176–190
17. Lin E, Kuo PH, Liu YL, et al: A deep learning approach for predicting antidepressant response in major depression using clinical and genetic biomarkers. Front Psychiatry 2018; 9:290
18. Kitchin R, McArdle G: What makes big data, big data? exploring the ontological characteristics of 26 datasets. Big Data Soc 2016; 3:2053951716631130
19. Uddin S, Khan A, Hossain ME, et al: Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak 2019; 19:281

20. Krstajic D, Buturovic LJ, Leahy DE, et al: Cross-validation pitfalls when selecting and assessing regression and classification models. J Cheminform 2014; 6:10
21. Donoho D: High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality. Providence, RI, American Mathematical Society, 2000, pp 1–32
22. Orrù G, Pettersson-Yeo W, Marquand AF, et al: Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. Neurosci Biobehav Rev 2012; 36:1140–1152
23. Khalid S, Khalil T, Nasreen S: A survey of feature selection and feature extraction techniques in machine learning, Proceedings of the Science and Information Conference, London, 2014
24. Van Der Maaten L, Postma E, Van den Herik J: Dimensionality reduction: a comparative review. J Mach Learn Res 2009; 10:66–71
25. Tang J, Alelyani S, Liu H: Feature selection for classification: a review, in Data Classification: Algorithms and Applications. Boca Raton, Fla, CRC Press, 2014, pp 37–64
26. Pigott TD: A review of methods for missing data. Educ Res Eval 2010; 7:353–383
27. Hutter F, Lücke J, Schmidt-Thieme L. Beyond manual tuning of hyperparameters. Künstliche Intelligenz 2015; 29:329–337
28. Sokolova M, Lapalme G: A systematic analysis of performance measures for classification tasks. Inf Process Manage 2009; 45:427–437
29. Batista GEAPA, Prati RC, Monard MC: A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor 2004; 6:20–29
30. Challen R, Denny J, Pitt M, et al: Artificial intelligence, bias and clinical safety. BMJ Qual Saf 2019; 28:231–237
31. Nadkarni PM, Ohno-Machado L, Chapman WW: Natural language processing: an introduction. J Am Med Inform Assoc 2011; 18:544–551
32. Perlis RH, Iosifescu DV, Castro VM, et al: Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. Psychol Med 2012; 42:41–50
33. Perlis RH: A clinical risk stratification tool for predicting treatment resistance in major depressive disorder. Biol Psychiatry 2013; 74:7–14
34. Wu S, Roberts K, Datta S, et al: Deep learning in clinical natural language processing: a methodical review. J Am Med Inform Assoc 2020; 27:457–470
35. Velupillai S, Mowery D, South BR, et al: Recent advances in clinical natural language processing in support of semantic analysis. Yearb Med Inform 2015; 10:183–193
36. Leonard Westgate C, Shiner B, Thompson P, et al: Evaluation of veterans' suicide risk with the use of linguistic detection methods. Psychiatr Serv 2015; 66:1051–1056
37. Havrlant L, Kreinovich V: A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation). Int J Gen Syst 2017; 46:27–36
38. Wang Y, Liu S, Afzal N, et al: A comparison of word embeddings for the biomedical natural language processing. J Biomed Inform 2018; 87:12–20
39. Abbe A, Grouin C, Zweigenbaum P, et al: Text mining applications in psychiatry: a systematic literature review. Int J Methods Psychiatr Res 2016; 25:86–100
40. Vu MT, Adalı T, Ba D, et al: A shared vision for machine learning in neuroscience. J Neurosci 2018; 38:1601–1607
41. Widge AS, Bilge MT, Montana R, et al: Electroencephalographic biomarkers for treatment response prediction in major depressive illness: a meta-analysis. Am J Psychiatry 2019; 176:44–56
42. Schnack HG, Kahn RS: Detecting neuroimaging biomarkers for psychiatric disorders: sample size matters. Front Psychiatry 2016; 7:50
43. Lynch C: Big data: how do your data grow? Nature 2008; 455:28–29
44. Halamka JD: Early experiences with big data at an academic medical center. Health Aff (Millwood) 2014; 33:1132–1138
45. Harford T: Big data: a big mistake? Significance 2014; 11:14–19
46. Vayena E, Blasimme A, Cohen IG: Machine learning in medicine: addressing ethical challenges. PLoS Med 2018; 15:e1002689
47. Obermeyer Z, Powers B, Vogeli C, et al: Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019; 366:447–453
48. Gara MA, Minsky S, Silverstein SM, et al: A naturalistic study of racial disparities in diagnoses at an outpatient behavioral health clinic. Psychiatr Serv 2019; 70:130–134
49. Cockerill RG: Ethics implications of the use of artificial intelligence in violence risk assessment. J Am Acad Psychiatry Law 2020; 48:345–349
50. Jiao Y, Chen R, Ke X, et al: Predictive models of autism spectrum disorder based on brain regional cortical thickness. Neuroimage 2010; 50:589–599
51. Sagi O, Rokach L: Ensemble learning: a survey. WIREs Data Mining Knowledge Discovery 2018; 8:e1249
52. Smialowski P, Frishman D, Kramer S: Pitfalls of supervised feature selection. Bioinformatics 2010; 26:440–443
53. Ribeiro MT, Singh S, Guestrin C: "Why should I trust you?": explaining the predictions of any classifier, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, 2016
54. Foster KR, Koprowski R, Skufca JD: Machine learning, medical diagnosis, and biomedical engineering research—commentary. Biomed Eng Online 2014; 13:94
55. Dinga R, Schmaal L, Penninx BWJH, et al: Evaluating the evidence for biotypes of depression: methodological replication and extension of. Neuroimage Clin 2019; 22:101796
56. Koutsouleris N, Kahn RS, Chekroud AM, et al: Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. Lancet Psychiatry 2016; 3:935–946
57. Zeng LL, Wang H, Hu P, et al: Multi-site diagnostic classification of schizophrenia using discriminant deep learning with functional connectivity MRI. EBioMedicine 2018; 30:74–85
58. Luo W, Phung D, Tran T, et al: Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016; 18:e323
59. Collins GS, Moons KGM: Reporting of artificial intelligence prediction models. Lancet 2019; 393:1577–1579
60. Doraiswamy PM, Blease C, Bodner K: Artificial intelligence and the future of psychiatry: insights from a global physician survey. Artif Intell Med 2020; 102:101753
61. Char DS, Shah NH, Magnus D: Implementing machine learning in health care: addressing ethical challenges. N Engl J Med 2018; 378:981–983
62. Yu K-H, Kohane IS: Framing the challenges of artificial intelligence in medicine. BMJ Qual Saf 2019; 28:238–241
63. Parasuraman R, Manzey DH: Complacency and bias in human use of automation: an attentional integration. Hum Factors 2010; 52:381–410
64. Sheriff K: Defining autonomy in the context of tort liability: Is machine learning indicative of robotic responsibility? Atlanta, Emory University School of Law, 2016
65. U.S. Food and Drug Administration: Artificial intelligence and machine learning in software as a medical device. Silver Spring, Md, US Food and Drug Administration. https://www.fda.gov/media/122535/download
66. Chekroud AM, Zotti RJ, Shehzad Z, et al: Cross-trial prediction of treatment outcome in depression: a machine learning approach. Lancet Psychiatry 2016; 3:243–250
67. Kessler RC, Stein MB, Petukhova MV, et al: Predicting suicides after outpatient mental health visits in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). Mol Psychiatry 2017; 22:544–551
68. Rumshisky A, Ghassemi M, Naumann T, et al: Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. Transl Psychiatry 2016; 6:e921–e921

Continuing Medical Education

You can earn CME credits by reading this article. Three articles in every *American Journal of Psychiatry* issue comprise a short course for up to 1 AMA PRA Category 1 Credit™ each. The course consists of reading the article and answering three multiple-choice questions with a single correct answer. CME credit is issued only online. Readers who want credit must subscribe to the AJP Continuing Medical Education Course Program (psychiatryonline.org/cme), select *The American Journal of Psychiatry* at that site, take the course(s) of their choosing, complete an evaluation form, and submit their answers for CME credit. A certificate for each course will be generated upon successful completion. This activity is sponsored by the American Psychiatric Association.

Examination Questions: Grzenda et al.

1. **A researcher builds a model to predict if a patient will respond to an antidepressant using magnetic resonance imaging (MRI) scans. The model performed well in the held-out test data. However, performance dropped when he applied the trained model to a new set of MRI scans from another institution. What is the most likely reason for the drop in performance?**
   a. Choosing the wrong hyper-parameters for the model fitting algorithm
   b. Not training the algorithm on a separate test set
   c. Changes in the nature of the data produced by the scanner at the second institution
   d. Choosing too many cross-validation splits or folds
2. **A psychiatrist is interested in improving the treatment of schizophrenia. She gathers a dataset of patients with known genotypes and medications tried, and hospitalizations over the following year. She develops a model to predict which medication, matched with which genotype, predicts a reduced risk of hospitalization in the year after starting medication. This is an example of:**
   a. Clustering
   b. Unsupervised learning
   c. Supervised learning with a categorical outcome
   d. Supervised learning with a continuous outcome
3. **A psychiatry resident is interested in predicting psychotherapy outcomes. He gathers a dataset containing numerous possible predictor variables related to patient and therapist characteristics, type of therapy, and outcomes. Which of the following approaches would explore this dataset without leakage?**
   a. Fit five different model types on the whole dataset, and report their average performance
   b. Split the data into an 80% training and 20% testing set. Train a model on the training set, then re-fit it on the testing set and report the parameters of this adjusted model.
   c. Split the data into 60% training, 20% validation, and 20% testing set. Train a model on the training set, then adjust its parameters to maximize performance on the validation set. Run that model on the testing set and report its performance.
   d. Split the data into an 80% training and 20% testing set. Train a model on the full set of data, then report its performance on the training and testing sets separately.