ANNUAL REVIEWS

Annual Review of Clinical Psychology

Machine Learning Approaches for Clinical Psychology and Psychiatry

Dominic B. Dwyer, Peter Falkai, and Nikolaos Koutsouleris

Department of Psychiatry and Psychotherapy, Section for Neurodiagnostic Applications, Ludwig-Maximilian University, Munich 80638, Germany; email: dominic.dwyer@med.uni-muenchen.de, peter.falkai@med.uni-muenchen.de, nikolaos.koutsouleris@med.uni-muenchen.de

Annu. Rev. Clin. Psychol. 2018. 14:91–118

First published as a Review in Advance on January 29, 2018

The *Annual Review of Clinical Psychology* is online at clinpsy.annualreviews.org

https://doi.org/10.1146/annurev-clinpsy-032816-045037

Keywords

machine learning, personalized medicine, translational psychiatry, artificial intelligence, mental health, clinical psychology, psychiatry

Abstract

Machine learning approaches for clinical psychology and psychiatry explicitly focus on learning statistical functions from multidimensional data sets to make generalizable predictions about individuals. The goal of this review is to provide an accessible understanding of why this approach is important for future practice given its potential to augment decisions associated with the diagnosis, prognosis, and treatment of people suffering from mental illness using clinical and biological data. To this end, the limitations of current statistical paradigms in mental health research are critiqued, and an introduction is provided to critical machine learning methods used in clinical studies. A selective literature review is then presented aiming to reinforce the usefulness of machine learning methods and provide evidence of their potential. In the context of promising initial results, the current limitations of machine learning approaches are addressed, and considerations for future clinical translation are outlined.

Contents

INTRODUCTION

Despite a century of considerable progress in clinical psychology and psychiatry, diagnoses are still unclear (Freedman et al. 2013, Hirschfeld et al. 2003), prognoses remain uncertain (Wunderink et al. 2009), and psychotherapeutic or pharmaceutical treatments are often effective in only 30–50% of patients (Hofmann et al. 2012, Rush et al. 2006, Wong et al. 2010). Thus, tailoring psychiatric care strategies to the needs of the individual patient relies strongly on repeated interactions that yield the patient's long-term diagnosis, prognosis, and optimal treatment regime over months or even years (Wunderink et al. 2009). This clinical model may be ultimately efficacious, but it unnecessarily prolongs suffering and wastes resources that could be spent in a more targeted way if patients were stratified to treatments that maximized their sustained recovery likelihood early in the illness course. This stratification is a central goal of translational clinical psychology and psychiatry, but research conducted over the past 50 years has not substantially improved the prevailing business model.

Generalizability: the performance of a decision function on data from new cases or contexts (e.g., temporal, geographic, genetic, cultural, or disease related)

While there are many reasons why greater advances have not been made (Insel & Cuthbert 2015, Insel et al. 2010, Kapur et al. 2012), in this article, we consider the issue in terms of the ways in which we approach problems, design research, and analyze data through the lens of the dominant statistical framework. We then suggest a different, machine learning approach to the problem of improving clinical care that specifically focuses on optimizing generalizability at a single-subject level using computational methods. In the latter part of the article, we review selected literature, address limitations, and outline future directions. The aim of the article is to provide an understanding of the machine learning approach for clinical researchers and to foster the motivation to use and improve these methods in future research. Achieving this aim is a critical step toward the facilitation of optimal translational research that could lead to the implementation of computational diagnostic and prognostic risk stratification aides in psychiatric and psychological care.

Overfitting: fitting a model to noise and idiosyncratic characteristics of a sample to obtain high predictive power, resulting in low generalizability

WHY HAS MORE RESEARCH NOT BEEN TRANSLATED?

At the turn of the twentieth century, great progress was made in the nascent fields of clinical psychology and psychiatry with the introduction of formal measures to quantify variance, error, and uncertainty of measurements when assessing groups of individuals. A combination of different techniques, including *p*-value testing, effect size measurement, and power analysis, became popular as an integrated approach to designing experiments and assessing the importance of results; this approach ultimately aimed to generalize from a sample of individuals to a hypothesized population (Bzdok & Yeo 2017, Bzdok et al. 2016, Nuzzo 2014). For example, by randomly sampling from the hypothesized populations of two groups (e.g., healthy and mentally ill), inferences could be made about the theoretical possibility of seeing a difference between the group distributions by chance. This approach can be labeled as the classical inferential paradigm and is the dominant method of conducting modern psychological and psychiatric research.

In recent years, the classical inferential approach has been increasingly scrutinized due to issues with replication and reproducibility. Serious questions regarding the interpretation and emphasis placed on *p*-value testing have led to what has been labeled a replication crisis (Ioannidis 2005, Schooler 2014), with replication rates estimated to be as little as 11% for preclinical studies (Begley & Ellis 2012). One reason for this low replication rate is that a *p*-value does not specifically measure the possibility of replication or reproducibility (Goodman 1992), and for many problems, a *p*-value of 0.05 actually means that there is a possibility of replication of about 50% (Goodman 1992, Nuzzo 2014). At the same threshold in neuroimaging research, empirical estimates indicate that there is a 70% chance that significant results are false positives even if they are statistically corrected for multiple comparisons (Eklund et al. 2012, 2016). These results suggest widespread overfitting, where the statistical models only reflect the noise and peculiarities of the current sample (Whelan & Garavan 2014), and have led to results that are statistically significant but not reproducible (Goodman et al. 2016; Ioannidis 2005, 2016; Nuzzo 2014).

Even if results are reproducible, statistically significant findings are often not clinically meaningful (Ioannidis 2016). With large enough samples, statistical significance can be found even in cases where the size of the difference between the two groups or correlation within groups (i.e., the effect size) is marginal (e.g., genome-wide association studies). When considering larger effect sizes, there are difficulties when this group-level result is applied to an individual. For example, effect sizes defined as huge (Cohen's $d = 2.0$) (Sawilowsky 2009) would result in correct identification of a case only approximately 64% of the time in enriched psychiatric samples with a 30% prevalence of disorder (Abi-Dargham & Horga 2016, Fusar-Poli et al. 2013). In non-enriched samples (e.g., 1% prevalence), the effect sizes that would be required to positively identify an individual even 50% of the time are in a range that is so high that it is unlikely to be found with appropriate sample sizes (i.e., Cohen's $d \sim 4.0$). As such, effect sizes are typically too low for research translation—an observation that has not been more widely recognized before because meaningful and intuitive measures of predictive accuracy are often not provided.

Effect sizes are also likely to be attenuated because mental health disorders are complex conditions involving interactions within and among environmental, behavioral, cognitive, emotional,

FACILITATION OF RESEARCH TRANSLATION

The dominant mode of statistical thinking influences the design and analysis of psychiatric and psychological research studies, despite evidence of serious problems that are limiting the translation of results. There is a need to use methods that can facilitate

- analysis of complex multivariate relationships related to high-dimensional data with known interdependencies, especially when these interdependencies are evidenced biologically (e.g., brain connections);
- empirical estimation and optimization of generalizability with clear reporting of probabilities; and
- model application to individuals rather than groups.

and biological systems related to each individual (Borsboom & Cramer 2013, Fornito et al. 2015, Molenaar & Campbell 2009, van de Leemput et al. 2014). Modeling these systems using techniques that independently address small components at a group level would reveal similarly restricted effect sizes because critical interactions (Borsboom & Cramer 2013, Woo et al. 2017) and individual differences (Molenaar & Campbell 2009) would be excluded. This is especially the case in the neuroimaging field, where known interactions between brain regions are ignored by univariate techniques that sacrifice fidelity for localizability (Lessov-Schlaggar et al. 2016), despite the strong hypothesis that mental health disorders are principally disconnection syndromes (Deco & Kringelbach 2014, Fornito et al. 2015). These limitations have contributed to the lack of neuroimaging biomarker discoveries despite the existence of thousands of research articles (Kapur et al. 2012), but they also apply to translational psychological research when it is dominated by group-based concepts (e.g., latent factors) whose validity is questionable when applied to individuals (Molenaar & Campbell 2009).

There are solutions to these problems within the dominant, classical statistical framework. Silver-bullet causative factors may still be found with highly controlled, hypothesis-driven, experimental studies or in mega-analysis of massive samples ($n > 10{,}000$) using univariate (Schizophr. Work. Group 2014) or multivariate techniques (Miller et al. 2016). Other ideas aimed at fostering good research practices have been proposed, such as including pretest probability estimates (Ioannidis 2005), reporting confidence intervals (Cumming 2014), conducting meta-analyses (Cumming 2014), or preregistering exploratory results in a database that is then used to replicate results (Nuzzo 2014). However, these suggestions are mostly additions to the same classical paradigm that fundamentally still applies to groups rather than individuals, involves questionable inferences regarding the generalizability of results, does not assess the ability to clinically translate results, and was not designed to be used with complex, multimodal, and massively multivariate data (Goodman et al. 2016). Thus, consideration of different approaches may be required to overcome these limitations (see the sidebar titled Facilitation of Research Translation).

MACHINE LEARNING AND WHY IT IS HERE TO STAY

Machine learning is broadly defined as a computational strategy that automatically determines (i.e., learns) methods and parameters to reach an optimal solution to a problem rather than being programmed by a human a priori to deliver a fixed solution. It is considered a subfield of artificial intelligence (AI) because this learning process putatively simulates a facet of human intelligence and can be used for ostensibly intelligent ends (e.g., speech, writing, face recognition, self-driving cars, or medical decision aides). To further define the term and provide concrete examples for other important concepts, we find it useful to begin with a historical machine called a perceptron that arose at the same time as advances were being made in digital computing in the 1950s (Bishop 2006).

The aim of the creators of the perceptron was to build a machine that could develop its own formulae to solve problems through experience (Rosenblatt 1958). Specifically, it was designed to identify pictures of triangles from 400 light sensors that would detect patterns of light and shade representing various shapes (e.g., squares and triangles). Rather than programming the machine in a top-down fashion and telling it what a square or a triangle looked like, the idea was to build it so that it would learn to correctly label the pattern from a bottom-up process of trial and error called training. The process was supervised by the experimenter, who labeled triangles as correct responses and the other shapes as errors. Over many trials, the machine would modify the statistical weights associated with each light sensor using an internal equation, or function. The response from some sensors (i.e., those in a triangle pattern) would be very heavily weighted during training, whereas others (i.e., those in a square pattern) would not, and this weighting would determine the prediction. The performance of this learning process was evaluated by showing an unlabeled set of images and testing the accuracy of the internal algorithm. In this way, the machine learned directly from the messy variations of triangle patterns to create an internalized, statistically based representation of the pattern—i.e., the machine learned a fuzzy concept of the pattern. Thus, the beginnings of machine learning centered on learning from real data, making limited prior assumptions about what that data looks like, and assessing the machine's performance in real-life circumstances.

Training: learning optimal decision rules or formulae within a subset of data with the aim of applying the functions to new instances

Function: a mathematical relation or expression involving one or more variables, usually with inputs and outputs

The perceptron helped to change the way computers were considered, from calculators that had to be programmed a priori with formulae and rules that were already known (e.g., prediction of planetary movements) to something that could learn fuzzy rules on its own (e.g., identification of patterns). This early success generated a large amount of hype that largely involved imaginative generalizations of the underlying logic to other problems where there are no explicit rules—mostly in psychology and the life sciences. However, the bubble burst when the hopes did not materialize because of limitations of early algorithms and the computational power available at that time. Some of these limitations can now be addressed because of increases in computational power that allow machines (which are now instantiated as software in digital computers) to learn from sometimes very complex data in cases where traditional models and algorithms have not performed optimally (e.g., face recognition).

Machine learning algorithms are now incorporated into our daily lives in the form of Internet searches and product recommendations, translation services, speech recognition services, and self-driving cars (Jordan & Mitchell 2015). In health care, the machine learning approach has performed equally as well as, or better than, clinicians in tasks that involve pattern recognition in images, such as the detection of skin cancer (Esteva et al. 2017), lung cancer (Yu et al. 2016), and eye disease (Long et al. 2017). Medical imaging companies have also already integrated machine learning algorithms to ultrasound devices to detect breast cancer (e.g., Samsung, RS80A). These advances are fueling the rapid growth of commercial machine learning health care solutions, with major oncology companies already using the technology (e.g., **http://nanthealth.com/**), investment in some companies exceeding $200 million USD (iCarbonX; **https://www.icarbonx.com/en/**), and new growth occurring in the psychiatric field (e.g., Spring; **https://www.springhealth.com/**). The early success of the new wave of machine learning has fueled considerable popular speculation that deserves to be treated with a degree of skepticism (Chen & Asch 2017), but it is clear that machine learning techniques will continue to be used because they perform better than classical approaches for many problems. This review will demonstrate that a bottom-up approach with modern machine learning methods can contribute greatly to clinical psychology and psychiatry by changing the way that problems are considered, addressing multidimensional interrelated data, and making generalizable predictions at the single-subject level.

IMPORTANT MACHINE LEARNING METHODS USED IN CLINICAL PSYCHOLOGY AND PSYCHIATRY

This section outlines key machine learning techniques for clinical psychological and psychiatric researchers. It is divided into six sections that progress from the types of problems and data to explanations of advanced approaches. For further details, there are several well-established and highly regarded comprehensive methodological guides to machine learning that include formal statistical nomenclature (Bishop 2006, Hastie et al. 2009, James et al. 2015).

Types of Problems and Data

The problems of translational clinical psychology and psychiatry that can be optimally addressed with machine learning fall into four main categories: diagnosis, prognosis, treatment prediction, and the detection and monitoring of potential biomarkers. Within this context, the ultimate aim of translational machine learning is to generate procedures that would be beneficial for clients, general practitioners, and in specialized hospital settings to improve patient outcomes, for example, a decision support aide that could use clinical or biological signatures to suggest a diagnosis, future prognosis, optimal treatments, and perform biological signature monitoring as an objective surrogate of treatment success. The predictions could be in the form of classifications (e.g., the person will benefit from treatment X) or regression frameworks to deliver continuous estimates (e.g., the patient will benefit from a specific dose of medication X). Predictions could also be enhanced with machine learning techniques to identify subgroups of individuals (e.g., clustering) or to index individuals against a population norm (Koutsouleris et al. 2014, Marquand et al. 2016). Although not a focus of this review, machine learning is also used in the enhancement of computer-aided psychotherapy (Bohannon 2015).

Importantly, the translational machine learning approaches discussed in this paper explicitly aim to produce models that are sufficiently meaningful, accurate, and generalizable to be integrated into clinical care. Accuracy can be assessed with a wide variety of intuitive measures of model performance derived from classical statistical approaches, such as sensitivity, specificity, and accuracy (**Table 1**), and further work has provided measures that enhance interpretability by integrating clinical decision theory into model optimization (Vickers & Elkin 2006). For example, net benefit analysis indexes the accuracy of a model against predefined clinical criteria reflecting the cost–benefit ratio of a positive prediction (e.g., needing to assess ten individuals when only one will have the disorder). To achieve accurate predictions, any type of quantitative data can be used for analysis, and there are fewer assumptions (e.g., normality, homogeneity of variance), in part because the estimates of model performance are empirically determined. Techniques in machine learning are also specially designed for the multivariate analysis of data sets with high dimensionality (i.e., many variables), even when the ratio of cases to variables is limited (Cortes & Vapnik 1995).

Generalizability and Cross-Validation

Generalizability can be broadly defined as the extent to which a statistical model generated in one group performs accurately in new groups or individuals. It can be assessed in a hierarchy that includes retrospective and prospective analyses, as presented in **Figure 1**. For all sample designs, generalizability can be estimated by directly applying the models to a new sample, with computer simulations, or with a combination of both techniques (Filzmoser et al. 2009, Koutsouleris et al. 2016, Stone 1974). The application of statistical models built in one sample and applied to another

Table 1 Performance metrics used to interpret results and optimize predictions

Measure	Description
Sensitivity	The proportion of affected cases with a positive test result (i.e., a true positive) in reference to all affected cases
Specificity	The proportion of nonaffected cases with a negative test result (i.e., true negative) in reference to all nonaffected cases
Accuracy	The fraction of correctly predicted cases in reference to all cases
Balanced accuracy	The accuracy in terms of true positive and negative cases balanced by the sample size of each positive and negative group; used to optimize models with unbalanced sample sizes
Positive predictive value	The probability that cases with a positive test result are actually positive; e.g., "60% of individuals with a positive test result will have the disease"
Negative predictive value	The probability that cases with a negative test result are actually negative; e.g., if an individual is classified as test negative, then there is a 77% chance that the test is correct
Positive likelihood ratio	The probability of a true positive test result divided by the probability of a false positive result, with 1 as the lowest limit
Negative likelihood ratio	The probability of a false negative test result divided by the probability of a true negative test result, with 1 as a lower limit
Diagnostic odds ratio	A ratio of the probability that the test is positive in subjects who are positive for the condition relative to that for negative results
Area under a curve	Area representing the discriminative power of a test between 0.5 (no discrimination) and 1 (perfect discrimination)
Youden's index	The addition of sensitivity and specificity minus 1 with a range between −1 (no discrimination) and 1 (perfect discrimination)
Net benefit	The proportion of true positives minus false positives, indexed by a predefined proportion of false positive cases (e.g., 1 out of 10 identified falsely)

is the gold standard in all translational science (Cannon et al. 2016, Carrion et al. 2016), but machine learning approaches also include the ability to simulate this process (**Figure 1**).

Within a machine learning framework, generalizability is estimated, and can be optimized, using simulations that resample data (e.g., bootstrapping). Of these techniques, cross-validation (CV) is the most robust because it separates the data in which the models are learnt from the data in which they are tested—i.e., similar to real-world circumstances where models need to be applied to new individuals. A simple example of CV is the leave-one-out scheme, where a test individual is withheld from a sample, a model is produced in the remaining training subjects, and then the model is applied to the left-out test case subject. This is repeated for all subjects in the sample, and the average accuracy is computed as an estimate of the out-of-sample generalizability. However, despite its simplicity, the leave-one-out scheme is not recommended due to the high variability of the predictions, the possibility of biased results, and the long computational time (Hastie et al. 2009, Varoquaux et al. 2017).

An alternative approach is *k*-fold CV, wherein the sample is divided into subsets of individuals called folds (**Figure 1*a***) (Stone 1974). An entire test fold is then left out in this procedure, a model is learned on the rest of the training data, and the model is tested on the left-out individuals (e.g., leave 10% of subjects out and then train on the rest). This process is then repeated for a prespecified number of *k* folds and results in more stable estimates of generalizability outside the sample because the training groups are more variable and there are more individuals in the left-out test sets (Hastie et al. 2009). A common question concerns the number of folds that are

Cross-validation (CV): a resampling technique to empirically assess and maximize the accuracy and generalizability of statistical models

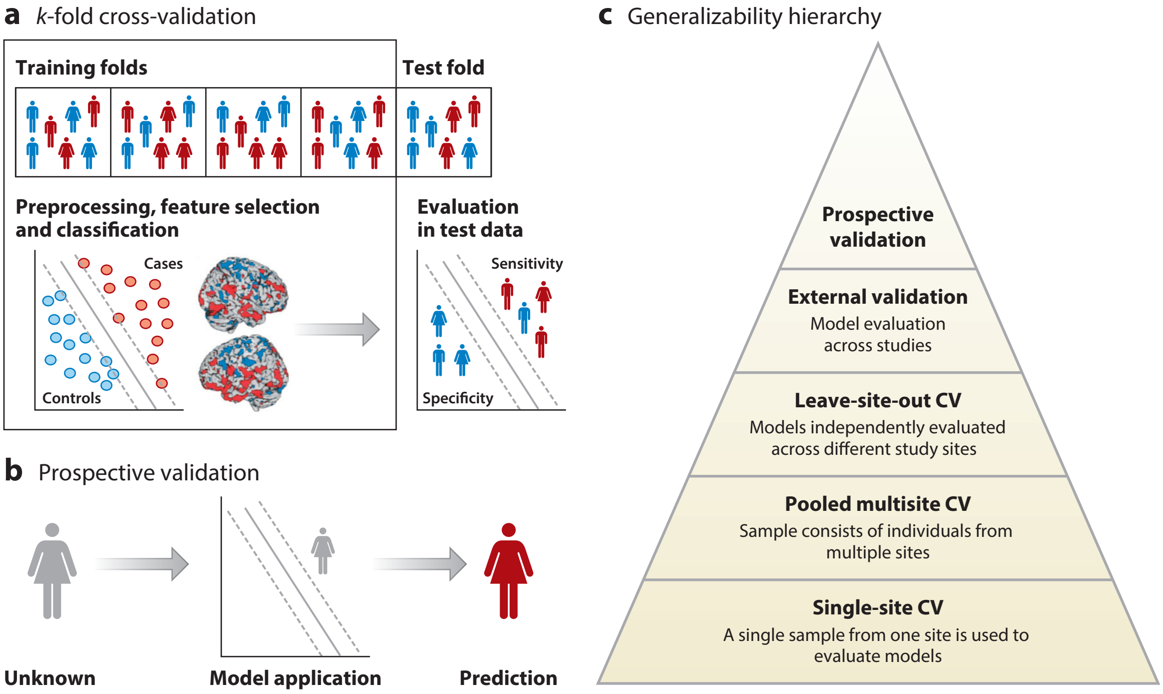
a *k*-fold cross-validation
Training folds
Test fold
Preprocessing, feature selection and classification
Cases
Controls
Evaluation in test data
Sensitivity
Specificity
b Prospective validation
Unknown
Model application
Prediction
c Generalizability hierarchy
Prospective validation
External validation
Model evaluation across studies
Leave-site-out CV
Models independently evaluated across different study sites
Pooled multisite CV
Sample consists of individuals from multiple sites
Single-site CV
A single sample from one site is used to evaluate models

Figure 1

Generalizability assessment using cross-validation (CV). (*a*) Generalizability can be assessed and optimized using *k*-fold CV, which simulates the application of the model to new individuals by separating a sample into folds consisting of training and test sets. In the simplest scheme, one test fold is left out, and models are trained in the remaining data. The models are then applied to the left-out individuals to finally assess the sensitivity and specificity of the model performance. This process is repeated for all folds. (*b*) Once the models are created in retrospective, labeled data, they can be prospectively applied to new individuals to make a prediction. (*c*) Generalizability can be assessed in a hierarchy where models are applied in progressively more diverse selections of individuals, contexts (e.g., temporal and geographical), and populations (e.g., genetic, cultural, diagnostic, or otherwise). Single-site CV (*bottom*) is most common and involves the creation and assessment of models using CV in a single location (e.g., a single hospital or clinic), thus measuring the degree of generalizability across different individuals in the respective catchment area. Pooled multisite CV involves combining subjects from multiple sites to create and assess models and is a simple test of model generalizability across individuals and contexts. Leave-site-out CV is a much more stringent test of context generalizability that involves explicitly separating experimental sites during the training and testing process. In this framework, one site is left out as a test site, models are trained in the remaining sites, and the models are assessed in the left-out site, thus more effectively assessing the degree to which a model will generalize to new contexts and populations. External validation is a technique used when models have been previously generated (using CV or otherwise) and are applied to new data from a different study (e.g., different individuals, contexts, and often study protocols). Prospective validation involves the application of pre-existing models to new individuals either in a clinical trial or in real-life circumstances.

used, and while authors recommend 5- or 10-fold CV (Breiman & Spector 1992) or statistical criteria (Hastie et al. 2009, James et al. 2015), it largely depends on the sample size, the number of variables, the machine learning algorithms used, and whether other procedures are being used (e.g., feature selection).

Feature: a single variable (e.g., age) usually from a larger feature set

A *k*-fold design is flexible and adaptations to the technique have allowed the optimization of generalizability (Filzmoser et al. 2009, Konig et al. 2007, Stone 1974). The current gold-standard scheme is nested (or double) CV, which includes a CV cycle within another, superordinate CV cycle that is ultimately used to assess the generalizability of the models (Filzmoser et al. 2009, Stone 1974). The nested CV design is powerful because parameters or features that optimize generalizability to test subjects in the inner, nested CV1 cycle can be learnt before the models are ultimately applied to the completely held-out subjects of the outer, CV2 cycle (Varma & Simon 2006). Within multicenter consortium-based studies (e.g., PRONIA; **https://www.pronia.eu**), it is also possible to assess the degree to which models generalize across different geographic sites using a leave-site-out design (Konig et al. 2007, Koutsouleris et al. 2016) (**Figure 1*c***). In this design, one site is left out, models are trained on the remaining sites, and the predictions are applied to the left-out site; this design thus provides a measure of between-site generalizability that can also be optimized in a nested CV design, which is a critical step in building models that specifically enhance generalizability. Further applications involve the assessment and optimization of multiple forms of generalizability by selecting the left-out folds (e.g., temporal, geographic, genetic, cultural, diagnostic, or experimental).

Preprocessing

For CV to be effective, the complete separation of training and test data is critical for all analysis steps. This applies to any preprocessing statistical operations that need to be performed on the data prior to classification or regression—e.g., scaling, imputation, nuisance variance removal, feature selection, or dimensionality reduction. These steps are embedded into a CV pipeline wherein parameters of the preprocessing steps can be optimized during training based on performance criteria (see, e.g., **Table 1**) and applied to test sets. If preprocessing steps are conducted outside of CV, then information about the test subjects could be included during the learning process, and this information leakage undermines the assessment of generalizability. A common example of information leakage is the case in which features are selected based on their predictive ability in the entire sample (e.g., choosing brain regions from a whole-brain comparison of two groups), and CV is then conducted using these variables. Estimates of generalizability are invalid and inflated under these circumstances because the features have already been optimized for the test subjects in the sample (Varma & Simon 2006).

The choice of preprocessing techniques depends on the data, the model that will be used, and the question being addressed. Some methods follow steps that would be conducted for any statistical analysis, such as the removal of variance associated with nuisance covariates (e.g., age or sex), the removal of variables with minimal variance, or the imputation of missing values. An important facet of preprocessing for machine learning contexts with large data sets is dimensionality reduction, which involves transforming the data from a high-dimensional space (i.e., many variables) to a lower-dimensional space (i.e., fewer variables) while still maintaining the information contained in the data (also known as data compression) (Van Der Maaten et al. 2009). Dimensionality reduction techniques include common statistical methods, such as principal components analysis, that aim to explain the variance in a data set by mathematically determining sets of uncorrelated variables called components (Hotelling 1933). Feature reduction is important in circumstances where the dimensionality of the data set undermines the generalizability of the models and obscures understanding of the results (e.g., neuroimaging or genetics). Other techniques to reduce the number of features include filters that select variables based on a predefined statistical test, as discussed below (e.g., correlation with the target variable).

The strength of preprocessing within a cross-validated machine learning pipeline is that the statistical parameters for each step can be automatically determined to balance the importance of achieving maximum accuracy during training with the ability of the models to generalize to the test sample. For example, rather than fixing the amount of principal components that are retained following an analysis, the optimal number to enhance generalizability can be learned during the training process (or, equivalently, within the nested CV1 cycle on test subjects). This facilitates an

Feature selection: the selection of optimal variables (i.e., features) from a larger data set with the aim of increasing accuracy and generalizability

Preprocessing: processing steps prior to the analysis of data using a predictive algorithm such as scaling, imputation, filtering, and dimensionality reduction

Pipeline: a series of statistical processing steps required for prediction and included in a cross-validation design

Information leakage: unintentional transference of information about the test data into the training data that invalidates generalizability claims

Filter: a procedure to identify optimally predictive features during preprocessing

Target: a variable that is predicted using the features empirical approach to the determination of parameters that are often guided by rules of thumb, default settings, or norms within scientific fields, which can be arbitrary and are usually associated with specific problems and analysis techniques (e.g., smoothing parameters for imaging, *p*-value criteria for genomics, or elbow criteria for principal components). However, there are restrictions to the optimization process because the possibility of overfitting to training and test data increases as a function of the number of parameter combinations being tested (i.e., the parameter space). As such, the choice of CV scheme and sample size is critical in this process, whereby stronger tests of generalizability are required in the presence of a larger parameter space, and the ways in which the best-performing parameters are chosen must be carefully considered (Eberhart & Kennedy 1995, Snoek et al. 2012). Ultimately, the parameter space and optimization process are often guided by Occam's razor to favor parsimony above complexity in order to achieve maximal generalizability (except for so-called deep learning models as discussed below).

Label: the target descriptor for the prediction (e.g., group label or continuous variable such as illness severity)

Hyperparameter: a modifiable setting of an algorithm that can be altered to obtain optimal prediction accuracy and generalizability

Support vector machine: a margin-based statistical technique to classify individuals into two or more groups by establishing a hyperplane on the basis of specific cases called support vectors

Supervised learning: labeling the cases to facilitate learning specific rules that can be later applied to unlabeled cases

Machine Learning Algorithms

A central component of the machine learning approach is the algorithm that is used to perform classification, regression, clustering, or normative modeling. Broadly, these can be separated into supervised techniques where the cases are labeled (e.g., into diagnostic groups), unsupervised techniques where the aim is to divide an unlabeled sample into groups of related cases, and semi-supervised techniques containing labeled and unlabeled cases. The algorithms available have often been developed within the machine learning field, but as is the case for preprocessing algorithms, other popular methods are also commonly used across statistical cultures (e.g., *k*-means).

The algorithms used in machine learning are generally united by their ability to optimize hyperparameters that modify the rules associated with a function. The ultimate aim is to learn hyperparameters that optimize a model to achieve a prespecified goal (e.g., accuracy in **Table 1** and generalizability in **Figure 1**). For example, in a classical regression framework using the method of least squares, there is one solution, one line is fitted to the data, and the coefficients are fixed. However, further machine learning developments of regression have added regularization hyperparameters that can be modified to determine a model that optimizes accuracy and generalizability (e.g., Ridge, LASSO, Elastic Net) (Zou & Hastie 2005). The specific values for these hyperparameters can be learned using CV during the training process and then applied to the test data, while the core parameters for the function (e.g., the coefficients) are determined and modified based on the hyperparameters chosen (e.g., minimizing an objective function using gradient descent). Regularization hyperparameters transform least squares regression into a flexible method that can be tuned to maximize the balance between optimally fitting a function to training data and generalizing to new single subjects (for an application of regularized regression in psychiatry, see Chekroud et al. 2016).

Many machine learning algorithms have been developed, and it is outside the scope of this review to cover them all (e.g., random forest, neural networks, and decision trees). However, a further example that is most widely used in psychiatry is the support vector machine (SVM) (Arbabshirani et al. 2017, Kambeitz et al. 2015, Orru et al. 2012). The SVM is a multivariate (or multivariable in regression terminology) supervised learning technique to sensitively classify individuals into groups within a margin-based statistical framework (James et al. 2015). This technique is important to cover because it has its origins in early multivariate pattern recognition algorithms (Fisher 1938) that aimed to automatically discover regularities in multivariate data to fulfill a goal (e.g., to classify individuals into groups or predict outcomes) (Bishop 2006). Like regression techniques, statistical pattern recognition techniques have been optimized for machine learning contexts.

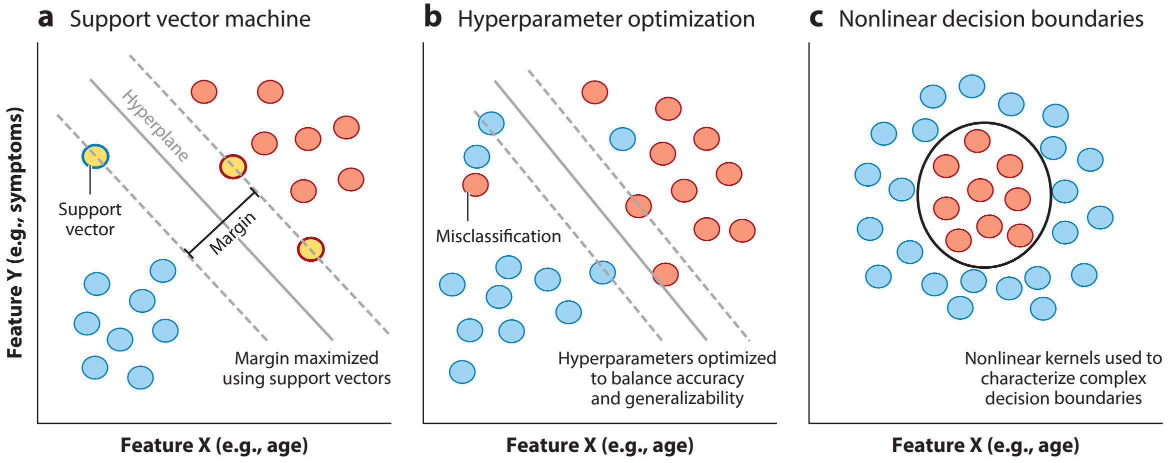
a Support vector machine
Hyperplane
Support vector
Margin
Margin maximized using support vectors
Feature Y (e.g., symptoms)
Feature X (e.g., age)
b Hyperparameter optimization
Misclassification
Hyperparameters optimized to balance accuracy and generalizability
Feature X (e.g., age)
c Nonlinear decision boundaries
Nonlinear kernels used to characterize complex decision boundaries
Feature X (e.g., age)

Figure 2

Conceptual representation of the support vector machine (SVM) approach to supervised classification. (*a*) The SVM algorithm works by identifying cases on the inner boundary of group distributions (*red* and *blue*) to construct a margin that maximally separates cases with different labels based on a hyperplane. The hyperplane determines the group membership within the training and testing samples. (*b*) Hyperparameters of a modern soft-margin SVM (Cortes & Vapnik 1995) can be tuned to balance accuracy in the training set and generalizability in the test data. In this case, modification to the C parameter has led to an optimally generalizable decision boundary that results in misclassification of two cases. (*c*) Nonlinear kernels can be used to characterize complex decision boundaries, such as the circular decision boundary depicted in this panel.

The SVM approach is illustrated in **Figure 2**, where cases are represented in a two-dimensional space, and the aim is to determine a linear boundary, or hyperplane, that can be used to optimally classify current cases and generalize to future (unlabeled) cases. Instead of using all cases to calculate the placement of the hyperplane, the SVM algorithm only uses cases on the closest external borders of the distributions—these cases are called support vectors. A margin can then be defined using the support vectors with the goal of maximizing the distance between the margins and the hyperplane. Parameters that determine the size of the margin and the degree of misclassification can be manipulated to balance the goals of correct classification in the training set and generalizability to the test set. For example, a hard margin can be employed that tightly fits the training data by not allowing any cases to be misclassified, but this margin may not generalize well. Alternatively, a soft margin can be used that allows cases to be within the margins or misclassified with the aim of increasing generalizability—counter-intuitively, allowing misclassifications increases generalizability by reducing overfitting during training (Cortes & Vapnik 1995) (**Figure 2**). The margin and misclassification allowance are controlled via the C hyperparameter, which can be optimized, or tuned, during the CV process to achieve a goal (e.g., the highest accuracy in the test subjects).

Aspects of the SVM algorithm have developed over time to include the ability to characterize nonlinear hyperplanes by using a data transformation implemented by a kernel function (Hastie et al. 2009, James et al. 2015). For nonlinear problems, the data is transformed, or mapped, via a kernel into a space with an added dimension (e.g., a third dimension that warps the data based on Gaussian distribution). In this space, a linear boundary can be used to separate the points, but when the data is back-projected to its original dimensions, a complex nonlinear decision boundary is apparent (**Figure 2**). For example, a polynomial kernel can be used to obtain a curved decision boundary in the original space, or a radial basis function kernel can establish a circular decision boundary for complex problems. The parameters for these kernels that define the degree of nonlinearity can also be tuned during the training process with the goal of optimizing generalizability in the test sample. Like adaptations to regression, the process of learning what the best parameters are to balance the fit of a model with its generalizability is a central component of machine learning.

Curse of dimensionality: a decrease in generalizability (mediated by overfitting) with increasing features relative to cases

Wrapper: an iterative procedure incorporated with the predictive algorithm to select the highest-performing features from a feature set

Feature Selection

Feature selection is the selection of specific variables from a larger set to enhance accuracy and generalizability. The problems in machine learning are typically very high dimensional (e.g., brain images, audio, or video files), and the feature space (i.e., the multidimensional representation of each case in terms of its value on each feature) often needs to be reduced to obtain generalizable predictions due to the curse of dimensionality—where an increasing number of features relative to cases results in lower accuracy and generalizability. In the machine learning context, feature selection can be conducted and optimized as part of preprocessing, as discussed above, or it can be combined with the machine learning algorithm itself.

An example of embedded feature selection is when machine learning algorithms employ mathematical regularization terms that reduce the relative contribution of specific features to zero (e.g., LASSO regression), effectively removing their influence and leaving the most predictive and nonredundant features. Another method is to use a computational approach called a wrapper that systematically employs separate analyses with different feature sets to find the most generalizable combination of a predefined size (e.g., 60% or 70% of all features). For example, greedy forward wrappers fundamentally use two steps: (*a*) Separate predictive algorithms (e.g., SVMs) are run for each feature, the best feature is chosen according to a predefined rule (e.g., accuracy), and it is then added to a feature pool; and (*b*) each remaining feature is added separately to the pool, separate SVMs are run again, and the best combination is selected. The second step is then repeated until it reaches the criterion set by the experimenter to obtain a parsimonious multivariate feature set (for an application of this procedure, see Koutsouleris et al. 2016). These procedures are conducted within CV in order to reduce overfitting and increase generalizability.

Advanced Approaches

The power of a machine learning approach is the automatic determination of analysis (hyper)parameters, but also the fact that multiple models with differing parameters can be used to maximize accuracy and generalizability. By using CV, one can ultimately produce multiple models using the training data that may contain differing optimal hyperparameter settings or feature sets. In the case of a nested CV design with a subordinated inner CV cycle, these potentially differing models are ultimately applied to the test set to determine the prediction for a given individual. This process is called ensemble learning (Polikar 2006) and is based on the reasoning that a measure of central tendency is likely to give a better assessment of the most generalizable level of accuracy (Galton 1907).

By considering the predictive problem in terms of finding a set of models that together produce the most accurate and generalizable solution, the standard approach can be extended. For example, in a process called late fusion, pipelines for different algorithms (e.g., linear and nonlinear SVMs) can be optimized such that the scores representing the degree to which an individual is included within a group (i.e., the decision scores) are then fused together to produce an average score that is used to make a final decision. Conceptually, this is similar to taking votes from a specialist committee of experts in different areas (e.g., clinical, cognitive, neuroimaging, and genetics) to obtain a final predictive outcome.

A conceptually related technique known as stacked generalization involves training different models, combining their decision scores within the training population, and then using another learning algorithm from these decision scores to obtain a final prediction (Wolpert 1992). This is also more broadly termed meta-learning because the models are being learned from a layer of outputs from other models. Using the above analogy in a clinical setting, this is similar to having a committee leader who learns from the suggestions of all specialist members before coming to a final decision independently. As with late fusion, stacking can be applied to different data modalities or to models that have used different algorithms (e.g., linear and nonlinear models).

Combined, these fusion and stacking methods are useful because they allow specialized learning of multivariate patterns within each putative system (e.g., brain, genes, cognition, or emotion), in addition to learning across the systems, to obtain optimal accuracy and generalizability. The idea of meta-learning is an important concept in fields such as deep learning (LeCun et al. 2015), where a machine learns from multiple layers that represent the data in different ways. In the analogy above, the deep learning process is more akin to having specialists for smaller components of the problem (e.g., one for each brain area or clinical question) and then combining these specialists' recommendations to form a final decision. Deep learning has not been widely employed in psychiatry yet because large amounts of data are required to effectively combat the curse of dimensionality and overfitting in the context of potentially thousands of hyperparameter combinations—i.e., increased fine-grained detail in the learning process requires more data.

SELECTIVE REVIEW OF RESEARCH LITERATURE

Multivariate formulae have been used for diagnostic, prognostic, and treatment decisions since the early twentieth century, when nomograms were critically involved in the treatment of polio (West 2005), and are now widely available online (**https://www.mskcc.org/nomograms**). However, multivariate formulae for outcomes or treatment decisions have been less widely used in clinical psychological or psychiatric practice (but see Cannon et al. 2016, Perry et al. 1998), possibly because of the complexity of the predictions that are required. The following general and selective review focuses on how machine learning might facilitate research translation for the optimization of diagnoses, prognoses, and treatment outcome predictions and for the detection of biomarkers that could be used as an index for all predictions. For comprehensive literature reviews in specific fields, the reader is directed toward comprehensive articles by Arbabshirani et al. (2017), Gabrieli et al. (2015), Kambeitz et al. (2015, 2016), Orru et al. (2012), and Woo et al. (2017).

Diagnosis

Early machine learning studies mostly explored whether diagnostic divisions between individuals could be recapitulated using high-dimensional data, in particular structural and functional neuroimaging (Arbabshirani et al. 2017, Orru et al. 2012). Studies have predominantly focused on Alzheimer's disease (Kloppel et al. 2008), depression (Fu et al. 2008), and schizophrenia (Csernansky et al. 2004, Davatzikos et al. 2008), but have more recently expanded to increasingly cover the diagnostic spectrum, including anxiety disorders (Lueken et al. 2015), anorexia (Lavagnino et al. 2015), substance abuse (Whelan & Garavan 2014), and specific phobias (Visser et al. 2016). Research has also expanded into using nonimaging modalities, such as genetic (Pettersson-Yeo et al. 2013), metabolomic (Setoyama et al. 2016), and proteomic data (Diniz et al. 2016). Overall, extant literature suggests that machine learning can be used to identify individuals with psychiatric disorders on the basis of brain data with accuracies above 75% (Arbabshirani et al.

2017; Kambeitz et al. 2015, 2016)—however, Schnack & Kahn (2016) provide a critical review of the accuracies reported by this literature.

The signatures learned from one diagnosis can be used to clarify clinical questions. One promising research direction is focused on clinical utility in cases where diagnoses are unclear (i.e., for differential diagnosis) and when current assessments are complex, time consuming, and costly, such as in the cases of mild cognitive impairment (Davatzikos et al. 2008) or the at-risk mental state for psychosis (Koutsouleris et al. 2009). For example, given that up to 75% of bipolar cases are misdiagnosed with unipolar depression (Fajutrao et al. 2009, Hirschfeld et al. 2003), Redlich et al. (2014) used machine learning to learn the brain signature that best separated cases when the diagnosis was established (Redlich et al. 2014). This signature was then used to separate cases in a completely different sample with a misdiagnosis rate of only 31%. In another case, Koutsouleris et al. (2015) learned a signature that separated depression from schizophrenia and used this signature in samples with uncertain diagnoses, such as first-episode psychosis and the high-risk state for psychosis. Similar work has been conducted to separate different dementias (Kloppel et al. 2008). These studies indicate that the neuroanatomical signatures learned in diagnostic studies can be used as clinical decision aides in unclear diagnostic circumstances.

Diagnostic machine learning studies address problems with mass univariate testing (Abi-Dargham & Horga 2016, Lessov-Schlaggar et al. 2016, Whelan & Garavan 2014) and descriptive group-based analyses (Gabrieli et al. 2015) by providing multivariate signatures that are valid at the single-subject level and could be used as biomarkers to monitor the progress of illness or the effectiveness of a treatment (Woo et al. 2017). However, predictions are limited by the symptomatic and neuroanatomical heterogeneity introduced by broad clinical definitions, which is increased with increasing sample size (Schnack & Kahn 2016). Ongoing efforts thus aim to resolve the heterogeneity of psychiatric groups by using unsupervised machine learning to automatically detect subgroups of individuals based on similar profiles of cognitive (Wu et al. 2016), genetic (Arnedo et al. 2015), brain functional (Clementz et al. 2016, Drysdale et al. 2017, Du et al. 2015), or brain structural data (Fair et al. 2012, Marquand et al. 2016, Varol et al. 2016). Other approaches involve parsing heterogeneity with reference to a reference population (Marquand et al. 2016). Although results are mixed, there are indications that this subtyping results in increased predictive accuracy in identifying individuals with mental illness when compared to healthy control subjects (Drysdale et al. 2017, Fair et al. 2012).

Prognosis

Determining an individual's prognostic outcome is critically important in clinical psychology and psychiatry for management, psychoeducation, and the provision of preventative psychotherapeutic and pharmacological interventions. Currently, the best that can be done with objective evidence is to assume that an individual fits into a diagnostic group based on clinical symptoms and signs (e.g., schizophrenia) and then refer to population averages (e.g., 50% chance of remission) (Harrison et al. 2001, Hegarty et al. 1994, Wunderink et al. 2009). Over time, these predictions become easier (e.g., if the patient has had a chronic and unremitting course for many years, then this is likely to continue) (Hegarty et al. 1994), but at the start of the illness, prognoses are inaccurate (Wunderink et al. 2009). A lack of accurate stratification results in an inflation of the numbers of individuals treated unnecessarily (i.e., the number needed to treat) (Alvarez-Jimenez et al. 2011, Siskind et al. 2016, van der Gaag et al. 2013). Having stratified prognostic predictions would thus be helpful for treatment planning to determine key junctures during the course of an illness, such as transition from a high-risk state to a criterion episode, relapses or remissions, and changes in symptom severity, daily functioning, and quality of life.

Research has demonstrated promising results in predicting illness course across disorders. There are at least 27 neuroimaging studies predicting transition from mild cognitive impairment to Alzheimer's disease, as reported in recent reviews (Arbabshirani et al. 2017), which show an average predictive accuracy above 70%. Studies of patients who transition to psychosis also show comparable predictive rates using neuroimaging (Koutsouleris et al. 2009, 2012, 2015), electrophysiology (Ramyead et al. 2016), and clinical measures (Mechelli et al. 2017). In depression, Schmaal et al. (2015) used a machine learning approach with data fusion of structural and functional task-based magnetic resonance imaging (MRI) to characterize depression trajectories (chronic, improving, and fast remission) over a period of 2 years and had similar predictive rates. Other studies in depression have also used self-reported clinical questionnaires to predict illness course (Kessler et al. 2016) or case records to stratify individuals based on suicide risk (Modai et al. 2002, Tran et al. 2014). Individuals have also been stratified based on models that predict future substance misuse using neuroimaging data (Bertocci et al. 2017), and using combinations of demographic, clinical, cognitive, neuroimaging, and genetic data (Whelan et al. 2014). These studies highlight the ability to stratify individuals into groups to optimize prognostic assessments.

As an example, Koutsouleris et al. (2016) predicted the functional outcomes of individuals with a first episode of psychosis using 189 questionnaire items that were collected across 44 mental health centers. A leave-center-out CV design (**Figure 1**) was used to empirically assess the geographic and contextual translatability of the models within a pipeline consisting of scaling, imputation to fill missing data values, and a feature selection process that used a wrapper (greedy-forward search). The results of this analysis highlighted that outcomes could be predicted with accuracies above 70% over 4-week and 12-month time periods and that leave-center-out CV did not significantly reduce this accuracy. The feature set could also be reduced to just 10 of the top-performing variables to predict positive outcomes at an accuracy of 72%, which was over 40% better than the base rate of individuals who recover as a proportion of the sample size. Thus, the study highlights the strength of a machine learning approach to measure generalizability and create parsimonious sets of features that could be used for the creation of new questionnaires to measure patient outcomes.

Other research directions include the prediction of continuous measures of symptom severity and outcomes for psychosis (de Wit et al. 2017, Tognin et al. 2014) and obsessive-compulsive disorder (Askland et al. 2015, Hoexter et al. 2013). Using pipelines similar to those addressed above but with regression algorithms (e.g., support vector regression), these studies have found modest predictive value. Another promising direction includes the use of electronic health records. A recent study assessed over 70,000 individuals to predict their future health status over a 1-year period using a pipeline focusing on dimensionality reduction and random forest algorithms (Miotto et al. 2016). Results indicated that attention deficit hyperactivity disorder and schizophrenia could be predicted with an area under the curve of approximately 0.85 (i.e., high accuracy).

Treatments

Currently, there are no objective, personalized methods to choose among multiple options when tailoring optimal psychotherapeutic and pharmacological treatment. Treatment choices are often initially guided by recommendations based on broad symptom classifications, such as the experience of depression, anxiety, or psychosis, and become personalized over time through a process of trial and error (Rush et al. 2006, Wong et al. 2010). This experimental medical approach is problematic in the context of research demonstrating that symptom remission following initial treatment with antidepressants can be as low as 11–30% of individuals with depression (Rush et al. 2006, Wong et al. 2010), and the response rate to cognitive behavioral therapy for conditions such as anxiety is only 46% (Hofmann et al. 2012). Especially when beginning treatment, better techniques are required for choosing among established pharmaceutical and psychotherapeutic techniques, and also for novel techniques such as noninvasive brain stimulation.

Treatment decision augmentation with machine learning has been conducted since the 1990s, with early studies focused on making predictions from clinical case records for psychotic and depressed inpatients (Modai et al. 1993, 1996). These studies focused broadly on making suggestions for multiple forms of treatment, including pharmacotherapy, psychotherapy, or community therapy, and performance was reported to be similar to clinical decisions (Modai et al. 1993, 1996). However, a limitation of early analyses in psychiatry, but also those across medical fields, was that they were limited to small, single-center samples outside of randomized clinical trials (Lisboa 2002). Additionally, due to computational limits, they rarely used biological data, which are commonplace today (e.g., MRI or genetics) and may provide objective biomarkers to guide treatment decisions (Insel & Cuthbert 2015). Also, algorithms that have become standards in psychiatric machine learning due to their performance (e.g., SVM) were still under development at the time (Cortes & Vapnik 1995).

More recent machine learning research has used the power of large-scale, multisite databases and advanced biological data sources to aid treatment decisions. Research into pharmacological decision support aides has been particularly useful in the case of depression, where studies have used large samples ($n > 1{,}000$) to predict response to different drugs (e.g., escitalopram, sertraline, venlafaxine, citalopram) (Chekroud et al. 2016, Etkin et al. 2015). For example, Chekroud et al. (2016) conducted a pattern recognition study in a sample of 1,949 individuals within a clinical trial (Star-D) (Rush et al. 2006) of citalopram. Using a pipeline consisting of k-fold CV with elastic net regression, they detected a parsimonious predictive pattern of 25 clinical questionnaire items from a total of 164 patient-reportable variables. These variables predicted clinical remission at a rate of 65%, which was more than 30% above the base rate of predicted efficacy for the drug. The generalizability of the models was further demonstrated by their ability to predict remission in a completely different sample, and the validity of the models was highlighted by their specificity when they did not generalize to other pharmaceutical treatments. Notably, the high generalizability of the predictions, together with further research (Chekroud et al. 2017), facilitated the rapid translation of the models into a web-based application designed to provide decision support to primary health care providers and clients (**https://www.springhealth.com**). This machine learning service is now being prospectively trialed in hospital settings, thus further highlighting the possibility of direct research translation.

While clinical assessments may provide useful decision support aides, a complementary direction is to incorporate brain structure and functioning. Emerging studies in this field have found that electroencephalographic measurements have been useful in predicting treatment response to drugs for depression (Khodayari-Rostamabad et al. 2013) and schizophrenia (Khodayari-Rostamabad et al. 2010). Biological assessments may be particularly important in cases where the treatment is invasive, such as in the use of electroconvulsive therapy, where a recent study found that brain structure can predict treatment response with an accuracy of 78% (Redlich et al. 2014). Predictive models based on brain data are also useful for non-invasive interventions, with brain functional MRI being used to predict response to cognitive behavioral therapy for anxiety spectrum conditions with accuracies above 75%—i.e., 30–40% better than the base rate of response to treatment (Ball et al. 2014, Doehrmann et al. 2013, Hahn et al. 2015, Whitfield-Gabrieli et al. 2016). Future research in this field is also expected to benefit when the treatment is directly associated with the biological measurement, for example, when brain structure is used to predict response to brain stimulation (Hasan et al. 2017). However, future studies using biological measures will require larger sample sizes and external validation to fulfill requirements for generalizability (**Figure 1**).

While choosing the right treatment at the right time (Insel & Cuthbert 2015) for individuals is important, this choice should also be motivated by a cost–benefit ratio that weighs the possible effects of treatments on symptoms against adverse risks, side effects, invasiveness, possible treatment resistance, and the necessary time and financial cost. Ultimately, these possibilities could be independently learned and incorporated into a clinical decision tool to balance the treatment decision. Burgeoning research in this field is investigating treatment side effects, such as metabolic syndrome following antipsychotic treatment (Van Schependom et al. 2015), and treatment resistance (Perlis 2013), but more research is required in this area.

LIMITATIONS AND FUTURE DIRECTIONS

Psychiatric machine learning studies have been conducted for more than 20 years (e.g., Modai et al. 1993), with the most recent wave of studies taking place in the past 10 years due to advances in algorithms and computing power. Given such a history, the question arises as to why the methods are not being used more widely in clinical practice in the context of their stated potential for translation. Aside from cultural norms associated with clinical practice, other reasons for this lack of research translation may include the validity of diagnostic and prognostic labels, representativeness of training data, mechanistic understanding of detected patterns, robust quantification of generalizability, specification of models' benefit–risk ratios, and practical implementation.

Validity of Diagnostic and Prognostic Labels

The prediction targets used thus far have been predominantly based on research samples that are clinically defined by criteria based on subjective symptoms and signs (Insel & Cuthbert 2015). These samples include a heterogeneous range of individuals who are diagnosed based on a wide range of potential combinations of symptoms, signs, and comorbidities. Equally, prognostic labels that are based on symptoms or functional outcomes can have multiple different interpretations (e.g., general assessment of functioning) (Torgalsboen & Rund 2002). These labels, and the heterogeneity that they create, may be hampering attempts to discover mechanisms and biomarker signatures using machine learning methodologies and, ultimately, the best predictive models for clinical care (Insel et al. 2010, Schumann et al. 2014). Rather than using summary measures and broad categorizations, future machine learning research may follow other initiatives in predicting more specific criteria such as symptom domains (e.g., cognitive functions such as executive functioning) (Gabrieli et al. 2015, Insel et al. 2010) or objective biological markers that cut across diagnostic boundaries (Insel & Cuthbert 2015, Woo et al. 2017). The discovery of objective biomarkers using machine learning would be especially valuable for the development of new, targeted treatments (e.g., pharmacological or psychotherapeutic) and for providing objective measurements for treatment effects or illness progression.

Representativeness of Training Data

Experimental research studies commonly employ a design that seeks to obtain a pure estimate of the population mean—e.g., finding quintessential examples of individuals with schizophrenia without the influence of other factors, such as comorbid illness, medication effects, or different clinical raters. However, if the aim is to create a generalizable prediction algorithm, then the sample needs to be representative of the heterogeneous population to which it will be applied in real life. Partly for this reason, larger sample sizes in neuroimaging machine learning studies generally result in lower predictive accuracy due to increases in heterogeneity (Arbabshirani et al. 2017, Schnack & Kahn 2016). For example, if the study is solely conducted at one location, then it may be the case that even a very large sample will only generalize to future cases from that location because of idiosyncrasies of methods (e.g., rating scales), materials (e.g., a specific scanner type), or participants (e.g., cultural homogeneity). Future research therefore needs to explicitly consider generalizability when designing or reporting research with translational aims. For any strong generalizability claim, collection of data across multiple different sites or tests across different centers is necessary (see **Figure 1**) and can be conducted as part of a single consortium-based research study (see PRONIA, **https://www.pronia.eu/**; NAPLS, **https://campuspress.yale.edu/napls/**; PSYSCAN, **http://www.psyscan.eu/**) by combining the data from multiple consortia or by using data sharing initiatives, such as ADNI (Mueller et al. 2005), ENIGMA (Thompson et al. 2014), or SchizConnect (Wang et al. 2016). Support provided by large data sharing initiatives will also be essential for the creation of generalizable predictions, such as the European Open Science Cloud (see **https://ec.europa.eu/research/openscience**).

Mechanistic Understanding of Detected Patterns

A limitation of current techniques is that the single-subject, predictive patterns are more opaque than more simple group-based statistics and methods (e.g., nomograms). For this reason, machine learning models are thought to be black boxes because interpreting how a model works, or especially why a subject is classified, is difficult (Hart & Wyatt 1990). This is a critical limitation, but it is notable that existing methods do actually allow a degree of interpretation (e.g., variable importance) at a global level that provides insight into how a model works (Koutsouleris et al. 2015, 2016). While at a single-subject level, recent methodological developments are now allowing insight into the reasons for specific classifications of individuals (Bastani et al. 2017, Fong & Vedaldi 2017, Guo et al. 2017, Koh & Liang 2017, Tulio Ribeiro et al. 2016), with proof-of-concept studies already conducted in medical settings (Katuwal & Chen 2016, Yang et al. 2016). These adaptations and additions to existing techniques will provide the specific variable profile that led to that prediction for an individual (e.g., a weighted single-subject brain map or specific clinical profile), which can then be used for increased mechanistic understanding of the classification model, in clinical care planning, and for communication to clients and their caregivers.

Robust Quantification of Generalizability

For any generalizability claim to be valid, there must not have been any unintentional human learning about the test data—e.g., when the optimal parameter ranges are defined based on results from previous cross-validated analyses of the same sample. To guard against this possibility within consortia or data sharing collectives, the labels need to be managed by an independent entity (i.e., in an honest broker approach) to foster transparency of external validation results. Once the models are generated, they could then be forwarded to the central data repository and applied to the held-out data. Given the heterogeneity of software and techniques, this could be achieved by using services such as Docker (**https://www.docker.com/**) or ViPAR (Carter et al. 2015) to align research with current standards of machine learning transparency (Brown et al. 2012, Silva et al. 2014). Repositories of machine learning models (i.e., that include all steps required for an analysis) could also be built to allow any researcher to test generalizability in their own sample and answer new questions, such as how brain signatures of depression, schizophrenia, and bipolar disorder apply to individuals with personality disorders.

Practical Implementation

The first phase in the use of any novel technique is testing and discovery, where the overall aim is to provide proof-of-concept evidence that the technique works and to explore potential avenues of investigation (Woo et al. 2017). However, from a translational perspective, many early studies either address a question that is not immediately clinically useful (e.g., diagnosis of chronic schizophrenia) or are difficult to practically implement on a wide scale (e.g., task-based functional MRI designs). Irrespective of the generalizability of a technique, many methods are unlikely to be implemented in health care settings simply because they are unavailable, they require too much training, or most importantly, they are not commercially viable. Moving forward, it is therefore important to consider the real-world practicality of translating a technique rather than the theoretical possibility that this might be possible at some (undefined) future time. This necessarily involves the contributions of clinicians and clinical researchers, so that the methodological advances are coupled with similarly advanced clinical reasoning to improve the efficacy of current clinical pathways.

A second major barrier in the translation of machine learning methods is that the techniques are unnecessarily difficult to understand and implement for many clinicians and clinical researchers. This review has attempted to provide an understanding of critical techniques, and there are some excellent textbooks that can assist researchers with statistical knowledge (e.g., James et al. 2015). An important additional direction of research is the creation of software that does not require advanced programming skills (e.g., PRoNTO; **http://www.mlnl.cs.ucl.ac.uk/pronto/**). One such software package that was specifically developed to help clinical psychologists and psychiatrists answer questions introduced in this review is available from our group and is called NeuroMiner (**https://www.pronia.eu/neurominer/**). Using this tool, researchers can employ various CV schemes, implement preprocessing, choose a machine learning algorithm, and interpret the models with a graphical interface. This software has demonstrated its usefulness in a number of studies as an in-house tool (Koutsouleris et al. 2009, 2015, 2016), across a large-scale consortium-based project (PRONIA; **https://www.pronia.eu/**), and in collaborative research with other laboratories (Opel et al. 2017).

FUTURE CLINICAL CARE

Given the recent successes of machine learning methods across multiple fields using varied data types (Esteva et al. 2017, Koutsouleris et al. 2016, Long et al. 2017, Yu et al. 2016), this article attempts to demonstrate that these methods have the potential to lead to the translation of research into advanced diagnostic, prognostic, treatment selection, and biomarker detection procedures in clinical psychology and psychiatry. Looking ahead, with the increasing availability of data from massive datasets (e.g., UK BioBank) (Miller et al. 2016) and the ongoing transition to computerized health care (Wachter 2015), the possibility of implementation in psychiatry and psychology is growing. This implementation could take many forms, including the use of publicly available algorithms, the creation of integrated hospital-based solutions (e.g., Long et al. 2017), or the monitoring of multiple sources of information using integrated database networks that include data collected inside and outside of the health care system (e.g., from smart phones and other sensors) (Mohr et al. 2017). For the clinician, this will mean that predictions may be available related to diagnostic, prognostic, and treatment decisions that could augment care.

Alongside the opportunities that exist for improvement of patient care, there are also serious procedural and ethical questions that need to be considered. As with the introduction of any new medical technology, there may be changes to practice and potentially to certain specialized medical roles (e.g., radiology) (Jha & Topol 2016), and therefore, ongoing communication among medical practitioners and consideration of counterarguments are required (Chen & Asch 2017, Rosenbaum 2015). Regulations from central agencies (e.g., the Food and Drug Administration or the European Medicines Agency) also need to be carefully considered to develop tests and methods that will be approved for widespread use. In particular, this involves consulting and implementing guidelines for the development of biomarkers that include interactions among academia, industry, government, and consortia to build strong evidence bases for qualification and assess risks through predefined regulatory pathways (Amur et al. 2015, Goodsaid & Mattes 2013, Woodcock et al. 2011).

The ethical implications of introducing prognostic tools also need to be seriously considered. For example, a prediction regarding whether an individual may have another episode of psychosis may affect expectations about their potential and their care, which could then contribute to an iatrogenic self-fulfilling prophecy where there is more potential for the prediction to be realized. Similar problems have been highlighted in examples of machine learning algorithms entrenching pre-existing discriminatory biases already present in society (O'Neil 2016). These problems require sustainable learning methods that recalibrate the models over time with changing contexts (Widmer & Kubat 1996), but they also necessitate close monitoring of robotic predictions and ethical oversight. If machine learning algorithms are integrated into future clinical care, as they have been integrated into our daily lives, then it will be important to monitor and regulate this process from moral and ethical perspectives. To this end, important partnerships have been established by AI-affiliated organizations (**https://www.partnershiponai.org/**), but it will also be necessary to have independent input, especially when there is the potential for commercial gain or vested interests.

CONCLUSIONS

The topic of machine learning encompasses an approach to problems as much as a set of specific methods. This approach fundamentally aims to learn information from multivariate data to fulfill the pragmatic goal of research translation by predicting outcomes for individuals rather than groups. The methods are beginning to be more widely used in clinical psychology and psychiatry and offer a promising future direction for translational research and, ultimately, clinical care. As with any emerging technology, caution needs to be used judiciously to overcome optimistic biases and, especially, to always serve the best interests of the people who the technology is designed to help.

SUMMARY POINTS

1. Diagnostic, prognostic, and treatment decisions could be improved by addressing the limitations of the dominant statistical methodology employed in clinical psychology and psychiatry.
2. Machine learning is a computational strategy that automatically determines (i.e., learns) methods and parameters to reach an optimal solution to a problem, rather than being programmed by a human to deliver a fixed solution a priori.
3. Generalizability is a central component of machine learning and needs to be assessed at multiple levels in different individuals and contexts to produce replicable research and, ultimately, practically useful models.

4. Statistical pipelines are embedded within cross-validated schemes consisting of the preprocessing of data operations and the creation of a statistical function. The parameters for the data operations are automatically determined (i.e., learned) with the aim of increasing predictive accuracy and generalizability.
5. Novel statistical techniques (e.g., SVMs) are available that can find multivariate and nonlinear functions (i.e., patterns) in high-dimensional data. Methods are available to combine these models to facilitate the highest predictive accuracy and generalizability.
6. Extant research demonstrates the success of machine learning techniques in stratifying individuals in reference to diagnostic, prognostic, and treatment decisions. These decisions may be aided by better biomarker detection and monitoring.
7. Future clinical psychology and psychiatry may combine the use of machine learning algorithms with existing experimental designs to enhance translational research and practically implement clinical tools.

FUTURE ISSUES

1. The validity of existing diagnostic and prognostic labels requires ongoing consideration.
2. Sample representativeness and generalizability related to machine learning models need to be increased.
3. Further research is required to develop mechanistic understandings of how the models work.
4. The translational feasibility of research designs and analyses needs to be more widely assessed and emphasized.
5. Accessible machine learning education and tool development is required to facilitate understanding and usage in the wider clinical research community.
6. Further consideration of moral and ethical questions is required as machine learning algorithms are more widely implemented in clinical practice.

DISCLOSURE STATEMENT

As member of the scientific advisory board of Spring Health, N.K. has provided consultancy to this company concerning the development of tools for the prediction of treatment outcomes in depression and psychosis. He does not hold equity and has not received any financial remuneration from this company. This affiliation did not influence the writing of this manuscript.

ACKNOWLEDGMENTS

This work was supported by the EU-FP7 project PRONIA (Personalised Prognostic Tools for Early Psychosis Management) under the grant 602152 to N.K.

LITERATURE CITED

Abi-Dargham A, Horga G. 2016. The search for imaging biomarkers in psychiatric disorders. *Nat. Med.* 22:1248–55

Alvarez-Jimenez M, Parker AG, Hetrick SE, McGorry PD, Gleeson JF. 2011. Preventing the second episode: a systematic review and meta-analysis of psychosocial and pharmacological trials in first-episode psychosis. *Schizophr. Bull.* 37:619–30

Amur S, LaVange L, Zineh I, Buckman-Garner S, Woodcock J. 2015. Biomarker qualification: toward a multiple stakeholder framework for biomarker development, regulatory acceptance, and utilization. *Clin. Pharmacol. Ther.* 98:34–46

Arbabshirani MR, Plis S, Sui J, Calhoun VD. 2017. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. *NeuroImage* 145:137–65

Arnedo J, Svrakic DM, del Val C, Romero-Zaliz R, Hernandez-Cuervo H, et al. 2015. Uncovering the hidden risk architecture of the schizophrenias: confirmation in three independent genome-wide association studies. *Am. J. Psychiatry* 172:139–53

Askland KD, Garnaat S, Sibrava NJ, Boisseau CL, Strong D, et al. 2015. Prediction of remission in obsessive compulsive disorder using a novel machine learning strategy. *Int. J. Methods Psychiatr. Res.* 24:156–69

Ball TM, Stein MB, Ramsawh HJ, Campbell-Sills L, Paulus MP. 2014. Single-subject anxiety treatment outcome prediction using functional neuroimaging. *Neuropsychopharmacology* 39:1254–61

Bastani O, Kim C, Bastani H. 2017. Interpretability via model extraction. arXiv:1706.09773 [cs.LG]

Begley CG, Ellis LM. 2012. Drug development: raise standards for preclinical cancer research. *Nature* 483:531–33

Bertocci MA, Bebko G, Versace A, Iyengar S, Bonar L, et al. 2017. Reward-related neural activity and structure predict future substance use in dysregulated youth. *Psychol. Med.* 47:1357–69

Bishop CM. 2006. *Pattern Recognition and Machine Learning*. New York: Springer

Bohannon J. 2015. The synthetic therapist. *Science* 349(6245):250−51

Borsboom D, Cramer AO. 2013. Network analysis: an integrative approach to the structure of psychopathology. *Annu. Rev. Clin. Psychol.* 9:91–121

Breiman L, Spector P. 1992. Submodel selection and evaluation in regression: the X-random case. *Int. Stat. Rev.* 60:291–319

Brown MR, Sidhu GS, Greiner R, Asgarian N, Bastani M, et al. 2012. ADHD-200 Global Competition: diagnosing ADHD using personal characteristic data can outperform resting state fMRI measurements. *Front. Syst. Neurosci.* 6:69

Bzdok D, Varoquaux G, Thirion B. 2016. Neuroimaging research: from null-hypothesis falsification to out-of-sample generalization. *Educ. Psychol. Meas.* 77:868–80

Bzdok D, Yeo BTT. 2017. Inference in the age of big data: future perspectives on neuroscience. *NeuroImage* 155:549–64

Prediction of psychosis using classical statistical methods highlighted as a counterpoint to using a machine learning approach for prediction in psychiatry.

Cannon TD, Yu C, Addington J, Bearden CE, Cadenhead KS, et al. 2016. An individualized risk calculator for research in prodromal psychosis. *Am. J. Psychiatry* 173:980–88

Carrion RE, Cornblatt BA, Burton CZ, Tso IF, Auther AM, et al. 2016. Personalized prediction of psychosis: external validation of the NAPLS-2 psychosis risk calculator with the EDIPPP project. *Am. J. Psychiatry* 173:989–96

Carter KW, Francis RW, Bresnahan M, Gissler M, Grønborg TK, et al. 2015. ViPAR: a software platform for the Virtual Pooling and Analysis of Research data. *Int. J. Epidemiol.* 45:408–16

Chekroud AM, Gueorguieva R, Krumholz HM, Trivedi MH, Krystal JH, McCarthy G. 2017. Reevaluating the efficacy and predictability of antidepressant treatments: a symptom clustering approach. *JAMA Psychiatry* 74(4):370−78

Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, et al. 2016. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry* 3:243–50

Chen JH, Asch SM. 2017. Machine learning and prediction in medicine: beyond the peak of inflated expectations. *N. Engl. J. Med.* 376:2507–9

Clementz BA, Sweeney JA, Hamm JP, Ivleva EI, Ethridge LE, et al. 2016. Identification of distinct psychosis biotypes using brain-based biomarkers. *Am. J. Psychiatry* 173:373–84

Cortes C, Vapnik V. 1995. Support-vector networks. *Mach. Learn.* 20:273–97

Csernansky JG, Schindler MK, Splinter NR, Wang L, Gado M, et al. 2004. Abnormalities of thalamic volume and shape in schizophrenia. *Am. J. Psychiatry* 161:896–902

Cumming G. 2014. The new statistics: why and how. *Psychol. Sci.* 25:7–29

Davatzikos C, Fan Y, Wu X, Shen D, Resnick SM. 2008. Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging. *Neurobiol. Aging* 29:514–23

de Wit S, Ziermans TB, Nieuwenhuis M, Schothorst PF, van Engeland H, et al. 2017. Individual prediction of long-term outcome in adolescents at ultra-high risk for psychosis: applying machine learning techniques to brain imaging data. *Hum. Brain Mapp.* 38:704–14

Deco G, Kringelbach ML. 2014. Great expectations: using whole-brain computational connectomics for understanding neuropsychiatric disorders. *Neuron* 84:892–905

Diniz BS, Lin CW, Sibille E, Tseng G, Lotrich F, et al. 2016. Circulating biosignatures of late-life depression (LLD): towards a comprehensive, data-driven approach to understanding LLD pathophysiology. *J. Psychiatr. Res.* 82:1–7

Doehrmann O, Ghosh SS, Polli FE, Reynolds GO, Horn F, et al. 2013. Predicting treatment response in social anxiety disorder from functional magnetic resonance imaging. *JAMA Psychiatry* 70:87–97

Drysdale AT, Grosenick L, Downar J, Dunlop K, Mansouri F, et al. 2017. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nat. Med.* 23:28–38

Du Y, Pearlson GD, Liu J, Sui J, Yu Q, et al. 2015. A group ICA based framework for evaluating resting fMRI markers when disease categories are unclear: application to schizophrenia, bipolar, and schizoaffective disorders. *NeuroImage* 122:272–80

Eberhart R, Kennedy J. 1995. *A new optimizer using particle swarm theory*. Presented at Int. Symp. Micro Mach. Hum. Sci., 6th, Nagoya, Jpn.

Eklund A, Andersson M, Josephson C, Johannesson M, Knutsson H. 2012. Does parametric fMRI analysis with SPM yield valid results? An empirical study of 1484 rest datasets. *NeuroImage* 61:565–78

Eklund A, Nichols TE, Knutsson H. 2016. Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *PNAS* 113:7900–5

Demonstrates the limitations of inferences made from magnetic resonance imaging studies using standard methods.

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, et al. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542:115–18

Etkin A, Patenaude B, Song YJ, Usherwood T, Rekshan W, et al. 2015. A cognitive-emotional biomarker for predicting remission with antidepressant medications: a report from the iSPOT-D trial. *Neuropsychopharmacology* 40:1332–42

Fair DA, Bathula D, Nikolas MA, Nigg JT. 2012. Distinct neuropsychological subgroups in typically developing youth inform heterogeneity in children with ADHD. *PNAS* 109:6769–74

Fajutrao L, Locklear J, Priaulx J, Heyes A. 2009. A systematic review of the evidence of the burden of bipolar disorder in Europe. *Clin. Pract. Epidemiol. Ment. Health* 5:3

Filzmoser P, Liebmann B, Varmuza K. 2009. Repeated double cross validation. *J. Chemometr.* 23:160–71

Fisher RA. 1938. The statistical utilization of multiple measurements. *Ann. Hum. Genet.* 8:376–86

Fong R, Vedaldi A. 2017. Interpretable explanations of black boxes by meaningful perturbation. arXiv:1704.03296 [cs.CV]

Fornito A, Zalesky A, Breakspear M. 2015. The connectomics of brain disorders. *Nat. Rev. Neurosci.* 16:159–72

Freedman R, Lewis DA, Michels R, Pine DS, Schultz SK, et al. 2013. The initial field trials of DSM-5: new blooms and old thorns. *Am. J. Psychiatry* 170:1–5

Fu CH, Mourao-Miranda J, Costafreda SG, Khanna A, Marquand AF, et al. 2008. Pattern classification of sad facial processing: toward the development of neurobiological markers in depression. *Biol. Psychiatry* 63:656–62

Fusar-Poli P, Borgwardt S, Bechdolf A, Addington J, Riecher-Rossler A, et al. 2013. The psychosis high-risk state: a comprehensive state-of-the-art review. *JAMA Psychiatry* 70:107–20

Gabrieli JD, Ghosh SS, Whitfield-Gabrieli S. 2015. Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. *Neuron* 85:11–26

Galton F. 1907. Vox populi (the wisdom of crowds). *Nature* 75:450–51

Goodman S. 1992. A comment on replication, p-values and evidence. *Stat. Med.* 11:875–79

Goodman SN, Fanelli D, Ioannidis JP. 2016. What does research reproducibility mean? *Sci. Transl. Med.* 8:341ps12

Goodsaid F, Mattes WB. 2013. *The Path from Biomarker Discovery to Regulatory Qualification*. Cambridge, MA: Academic

Guo W, Zhang K, Lin L, Huang S, Xing X. 2017. Towards interrogating discriminative machine learning models. arXiv:1705.08564 [cs.LG]

Hahn T, Kircher T, Straube B, Wittchen HU, Konrad C, et al. 2015. Predicting treatment response to cognitive behavioral therapy in panic disorder with agoraphobia by integrating local neural information. *JAMA Psychiatry* 72:68–74

Harrison G, Hopper K, Craig T, Laska E, Siegel C, et al. 2001. Recovery from psychotic illness: a 15- and 25-year international follow-up study. *Br. J. Psychiatry* 178:506–17

Hart A, Wyatt J. 1990. Evaluating black-boxes as medical decision aids: issues arising from a study of neural networks. *Med. Inform.* 15:229–36

Hasan A, Wobrock T, Guse B, Langguth B, Landgrebe M, et al. 2017. Structural brain changes are associated with response of negative symptoms to prefrontal repetitive transcranial magnetic stimulation in patients with schizophrenia. *Mol. Psychiatry* 22:857–64

Hastie T, Tibshirani R, Friedman J. 2009. *The Elements of Statistical Learning*. New York: Springer

Hegarty JD, Baldessarini RJ, Tohen M, Waternaux C, Oepen G. 1994. One hundred years of schizophrenia: a meta-analysis of the outcome literature. *Am. J. Psychiatry* 151:1409–16

Hirschfeld R, Lewis L, Vornik LA. 2003. Perceptions and impact of bipolar disorder: How far have we really come? Results of the National Depressive and Manic-Depressive Association 2000 survey of individuals with bipolar disorder. *J. Clin. Psychiatry* 64:161–74

Hoexter MQ, Miguel EC, Diniz JB, Shavitt RG, Busatto GF, Sato JR. 2013. Predicting obsessive-compulsive disorder severity combining neuroimaging and machine learning methods. *J. Affect. Disord.* 150:1213–16

Hofmann SG, Asnaani A, Vonk IJ, Sawyer AT, Fang A. 2012. The efficacy of cognitive behavioral therapy: a review of meta-analyses. *Cogn. Ther. Res.* 36:427–40

Hotelling H. 1933. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* 24:417–41

Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, et al. 2010. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *Am. J. Psychiatry* 167:748–51

Insel TR, Cuthbert BN. 2015. Brain disorders? Precisely. *Science* 348:499–500

Ioannidis JP. 2005. Why most published research findings are false. *PLOS Med.* 2:e124

Influential study questioning the dominant statistical framework and research design.

Ioannidis JP. 2016. Why most clinical research is not useful. *PLOS Med.* 13:e1002049

James G, Witten D, Hastie T, Tibshirani R. 2015. *An Introduction to Statistical Learning with Applications in R*. New York: Springer

Jha S, Topol EJ. 2016. Adapting to artificial intelligence: radiologists and pathologists as information specialists. *JAMA* 316:2353–54

Jordan MI, Mitchell TM. 2015. Machine learning: trends, perspectives, and prospects. *Science* 349:255–60

Kambeitz J, Cabral C, Sacchet MD, Gotlib IH, Zahn R, et al. 2016. Detecting neuroimaging biomarkers for depression: a meta-analysis of multivariate pattern recognition studies. *Biol. Psychiatry* 82:330–38

Kambeitz J, Kambeitz-Ilankovic L, Leucht S, Wood S, Davatzikos C, et al. 2015. Detecting neuroimaging biomarkers for schizophrenia: a meta-analysis of multivariate pattern recognition studies. *Neuropsychopharmacology* 40:1742–51

Meta-analysis of machine learning studies conducted in schizophrenia demonstrating their potential as biomarkers.

Kapur S, Phillips AG, Insel TR. 2012. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Mol. Psychiatry* 17:1174–79

Katuwal GJ, Chen R. 2016. Machine learning model interpretability for precision medicine. arXiv:1610.09045 [q-bio.QM]

Kessler RC, van Loo HM, Wardenaar KJ, Bossarte RM, Brenner LA, et al. 2016. Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. *Mol. Psychiatry* 21:1366–71

Khodayari-Rostamabad A, Hasey GM, Maccrimmon DJ, Reilly JP, de Bruin H. 2010. A pilot study to determine whether machine learning methodologies using pre-treatment electroencephalography can predict the symptomatic response to clozapine therapy. *Clin. Neurophysiol.* 121:1998–2006

Khodayari-Rostamabad A, Reilly JP, Hasey GM, de Bruin H, Maccrimmon DJ. 2013. A machine learning approach using EEG data to predict response to SSRI treatment for major depressive disorder. *Clin. Neurophysiol.* 124:1975–85

Kloppel S, Stonnington CM, Barnes J, Chen F, Chu C, et al. 2008. Accuracy of dementia diagnosis: a direct comparison between radiologists and a computerized method. *Brain* 131:2969–74

Koh PW, Liang P. 2017. Understanding black-box predictions via influence functions. arXiv:1703.04730 [stat.ML]

Konig IR, Malley JD, Weimar C, Diener HC, Ziegler A. 2007. Practical experiences on the necessity of external validation. *Stat. Med.* 26:5499–511

Koutsouleris N, Borgwardt S, Meisenzahl EM, Bottlender R, Moller HJ, Riecher-Rossler A. 2012. Disease prediction in the at-risk mental state for psychosis using neuroanatomical biomarkers: results from the FePsy study. *Schizophr. Bull.* 38:1234–46

Koutsouleris N, Davatzikos C, Borgwardt S, Gaser C, Bottlender R, et al. 2014. Accelerated brain aging in schizophrenia and beyond: a neuroanatomical marker of psychiatric disorders. *Schizophr. Bull.* 40:1140–53

Koutsouleris N, Kahn RS, Chekroud AM, Leucht S, Falkai P, et al. 2016. Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. *Lancet Psychiatry* 3:935–46

Demonstrates the assessment of multiple forms of generalizability using a leave-site-out methodology.

Koutsouleris N, Meisenzahl E, Borgwardt S, Riecher-Rossler A, Frodl T, et al. 2015. Individualized differential diagnosis of schizophrenia and mood disorders using neuroanatomical biomarkers. *Brain* 138:2059–73

Koutsouleris N, Meisenzahl EM, Davatzikos C, Bottlender R, Frodl T, et al. 2009. Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. *Arch. Gen. Psychiatry* 66:700–12

First study to show the potential of machine learning techniques in neuroimaging to predict the onset of psychosis.

Lavagnino L, Amianto F, Mwangi B, D'Agata F, Spalatro A, et al. 2015. Identifying neuroanatomical signatures of anorexia nervosa: a multivariate machine learning approach. *Psychol. Med.* 45:2805–12

LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* 521:436–44

Lessov-Schlaggar CN, Rubin JB, Schlaggar BL. 2016. The fallacy of univariate solutions to complex systems problems. *Front. Neurosci.* 10:267

Lisboa PJ. 2002. A review of evidence of health benefit from artificial neural networks in medical intervention. *Neural Netw*. 15:11–39

Long E, Lin H, Liu Z, Wu X, Wang L, et al. 2017. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. *Nat. Biomed. Eng.* 1:0024

Lueken U, Straube B, Yang Y, Hahn T, Beesdo-Baum K, et al. 2015. Separating depressive comorbidity from panic disorder: a combined functional magnetic resonance imaging and machine learning approach. *J. Affect. Disord.* 184:182–92

Marquand AF, Rezek I, Buitelaar J, Beckmann CF. 2016. Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. *Biol. Psychiatry* 80:552–61

Mechelli A, Lin A, Wood S, McGorry P, Amminger P, et al. 2017. Using clinical information to make individualized prognostic predictions in people at ultra high risk for psychosis. *Schizophr. Res.* 184:32–38

Miller KL, Alfaro-Almagro F, Bangerter NK, Thomas DL, Yacoub E, et al. 2016. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat. Neurosci.* 19:1523–36

Miotto R, Li L, Kidd BA, Dudley JT. 2016. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci. Rep.* 6:26094

Modai I, Israel A, Mendel S, Hines EL, Weizman R. 1996. Neural network based on adaptive resonance theory as compared to experts in suggesting treatment for schizophrenic and unipolar depressed in-patients. *J. Med. Syst.* 20:403–12

Modai I, Kurs R, Ritsner M, Oklander S, Silver H, et al. 2002. Neural network identification of high-risk suicide patients. *Med. Inform. Internet Med.* 27:39–47

Modai I, Stoler M, Inbarsaban N, Saban N. 1993. Clinical decisions for psychiatric inpatients and their evaluation by a trained neural-network. *Methods Inf. Med.* 32:396–99

Mohr DC, Zhang M, Schueller SM. 2017. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annu. Rev. Clin. Psychol.* 13:23–47

Molenaar PC, Campbell CG. 2009. The new person-specific paradigm in psychology. *Curr. Dir. Psychol. Sci.* 18:112–17

Questions the classical psychological paradigm of latent variables and highlights the importance of individual differences.

Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack CR, et al. 2005. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimers Dement*. 1:55–66

Nuzzo R. 2014. Scientific method: statistical errors. *Nature* 506:150–52

O'Neil C. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown

Opel N, Redlich R, Kaehler C, Grotegerd D, Dohm K, et al. 2017. Prefrontal gray matter volume mediates genetic risks for obesity. *Mol. Psychiatry* 22:703–10

Orru G, Pettersson-Yeo W, Marquand AF, Sartori G, Mechelli A. 2012. Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neurosci. Biobehav. Rev.* 36:1140–52

Perlis RH. 2013. A clinical risk stratification tool for predicting treatment resistance in major depressive disorder. *Biol. Psychiatry* 74:7–14

Perry PJ, Bever KA, Arndt S, Combs MD. 1998. Relationship between patient variables and plasma clozapine concentrations: a dosing nomogram. *Biol. Psychiatry* 44:733–38

Pettersson-Yeo W, Benetti S, Marquand A, Dell'Acqua F, Williams S, et al. 2013. Using genetic, cognitive and multi-modal neuroimaging data to identify ultra-high-risk and first-episode psychosis at the individual level. *Psychol. Med.* 43:2547–62

Polikar R. 2006. Ensemble based systems in decision making. *IEEE Circuits Syst. Mag.* 6:21–45

Ramyead A, Studerus E, Kometer M, Uttinger M, Gschwandtner U, et al. 2016. Prediction of psychosis using neural oscillations and machine learning in neuroleptic-naive at-risk patients. *World J. Biol. Psychiatry* 17:285–95

Redlich R, Almeida JR, Grotegerd D, Opel N, Kugel H, et al. 2014. Brain morphometric biomarkers distinguishing unipolar and bipolar depression: a voxel-based morphometry–pattern classification approach. *JAMA Psychiatry* 71:1222–30

Rosenbaum L. 2015. Transitional chaos or enduring harm? The EHR and the disruption of medicine. *New Engl. J. Med.* 373:1585–88

Rosenblatt F. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* 65:386–408

Rush AJ, Trivedi MH, Wisniewski SR, Nierenberg AA, Stewart JW, et al. 2006. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR* D report. *Am. J. Psychiatry* 163:1905–17

Sawilowsky SS. 2009. New effect size rules of thumb. *J. Mod. Appl. Stat. Methods* 8:597–99

Schizophr. Work. Group. 2014. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511:421–27

Schmaal L, Marquand AF, Rhebergen D, van Tol MJ, Ruhe HG, et al. 2015. Predicting the naturalistic course of major depressive disorder using clinical and multimodal neuroimaging information: a multivariate pattern recognition study. *Biol. Psychiatry* 78:278–86

Schnack HG, Kahn RS. 2016. Detecting neuroimaging biomarkers for psychiatric disorders: sample size matters. *Front. Psychiatry* 7:50

Schooler JW. 2014. Metascience could rescue the "replication crisis." *Nature* 515:9

Schumann G, Binder EB, Holte A, de Kloet ER, Oedegaard KJ, et al. 2014. Stratified medicine for mental disorders. *Eur. Neuropsychopharmacol.* 24:5–50

Setoyama D, Kato TA, Hashimoto R, Kunugi H, Hattori K, et al. 2016. Plasma metabolites predict severity of depression and suicidal ideation in psychiatric patients: a multicenter pilot analysis. *PLOS ONE* 11:e0165267

Silva RF, Castro E, Gupta CN, Cetin M, Arbabshirani M, et al. 2014. The tenth annual MLSP competition: schizophrenia classification challenge. In *2014 IEEE International Workshop on Machine Learning for Signal Processing*, ed. M Mboup, T Adali, E Moreau, J Larsen. New York: IEEE. **https://doi.org/10.1109/MLSP.2014.6958889**

Siskind D, McCartney L, Goldschlager R, Kisely S. 2016. Clozapine v. first- and second-generation antipsychotics in treatment-refractory schizophrenia: systematic review and meta-analysis. *Br. J. Psychiatry* 209:385–92

Snoek J, Larochelle H, Adams RP. 2012. Practical Bayesian optimization of machine learning algorithms. arXiv:1206.2944 [stat.ML]

Stone M. 1974. Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Stat. Methodol.* 36:111–47

Thompson PM, Stein JL, Medland SE, Hibar DP, Vasquez AA, et al. 2014. The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data. *Brain Imaging Behav*. 8:153–82

Tognin S, Pettersson-Yeo W, Valli I, Hutton C, Woolley J, et al. 2014. Using structural neuroimaging to make quantitative predictions of symptom progression in individuals at ultra-high risk for psychosis. *Front. Psychiatry* 4:187

Torgalsboen AK, Rund BR. 2002. Lessons learned from three studies of recovery from schizophrenia. *Int. Rev. Psychiatry* 14:312–17

Tran T, Luo W, Phung D, Harvey R, Berk M, et al. 2014. Risk stratification using data from electronic medical records better predicts suicide risks than clinician assessments. *BMC Psychiatry* 14:76

Tulio Ribeiro M, Singh S, Guestrin C. 2016. "Why should I trust you?" Explaining the predictions of any classifier. arXiv:1602.04938 [cs.LG]

Early study providing techniques to open the black box of machine learning.

van de Leemput IA, Wichers M, Cramer AO, Borsboom D, Tuerlinckx F, et al. 2014. Critical slowing down as early warning for the onset and termination of depression. *PNAS* 111:87–92

van der Gaag M, Smit F, Bechdolf A, French P, Linszen DH, et al. 2013. Preventing a first episode of psychosis: meta-analysis of randomized controlled prevention trials of 12 month and longer-term follow-ups. *Schizophr. Res.* 149:56–62

Van Der Maaten L, Postma E, Van den Herik J. 2009. Dimensionality reduction: a comparative review. *J. Mach. Learn. Res.* 10:66–71

Van Schependom J, Yu WP, Gielen J, Laton J, De Keyser J, et al. 2015. Do advanced statistical techniques really help in the diagnosis of the metabolic syndrome in patients treated with second-generation antipsychotics? *J. Clin. Psychiatry* 76:E1292–99

Varma S, Simon R. 2006. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformat*. 7:91

Varol E, Sotiras A, Davatzikos C. 2016. HYDRA: revealing heterogeneity of imaging and genetic patterns through a multiple max-margin discriminative analysis framework. *NeuroImage* 145:346–64

Varoquaux G, Raamana PR, Engemann DA, Hoyos-Idrobo A, Schwartz Y, Thirion B. 2017. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage* 145:166–79

Vickers AJ, Elkin EB. 2006. Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Making* 26:565–74

Visser RM, Haver P, Zwitser RJ, Scholte HS, Kindt M. 2016. First steps in using multi-voxel pattern analysis to disentangle neural processes underlying generalization of spider fear. *Front. Hum. Neurosci.* 10:222

Wachter R. 2015. *The Digital Doctor: Hope, Hype, and Harm at the Dawn of Medicine's Computer Age*. New York: McGraw-Hill

Wang L, Alpert KI, Calhoun VD, Cobia DJ, Keator DB, et al. 2016. SchizConnect: mediating neuroimaging databases on schizophrenia and related disorders for large-scale integration. *NeuroImage* 124:1155–67

West JB. 2005. The physiological challenges of the 1952 Copenhagen poliomyelitis epidemic and a renaissance in clinical respiratory physiology. *J. Appl. Physiol.* 99:424–32

Whelan R, Garavan H. 2014. When optimism hurts: inflated predictions in psychiatric neuroimaging. *Biol. Psychiatry* 75:746–48

Whelan R, Watts R, Orr CA, Althoff RR, Artiges E, et al. 2014. Neuropsychosocial profiles of current and future adolescent alcohol misusers. *Nature* 512:185–89

Whitfield-Gabrieli S, Ghosh SS, Nieto-Castanon A, Saygin Z, Doehrmann O, et al. 2016. Brain connectomics predict response to treatment in social anxiety disorder. *Mol. Psychiatry* 21:680–85

Widmer G, Kubat M. 1996. Learning in the presence of concept drift and hidden contexts. *Mach. Learn.* 23:69–101

Wolpert DH. 1992. Stacked generalization. *Neural Netw*. 5:241–59

Wong EHF, Yocca F, Smith MA, Lee CM. 2010. Challenges and opportunities for drug discovery in psychiatric disorders: the drug hunters' perspective. *Int. J. Neuropsychopharmacol.* 13:1269–84

Woo CW, Chang LJ, Lindquist MA, Wager TD. 2017. Building better biomarkers: brain models in translational neuroimaging. *Nat. Neurosci.* 20:365–77

Excellent review of the use of machine learning for translational neuroimaging.

Woodcock J, Buckman S, Goodsaid F, Walton MK, Zineh I. 2011. Qualifying biomarkers for use in drug development: a US Food and Drug Administration overview. *Expert Opin. Med. Diagn.* 5:369–74

Wu MJ, Mwangi B, Bauer IE, Passos IC, Sanches M, et al. 2016. Identification and individualized prediction of clinical phenotypes in bipolar disorders using neurocognitive data, neuroimaging scans and machine learning. *NeuroImage* 145:254–64

Wunderink L, Sytema S, Nienhuis FJ, Wiersma D. 2009. Clinical recovery in first-episode psychosis. *Schizophr. Bull.* 35:362–69

Yang C, Delcher C, Shenkman E, Ranka S. 2016. *Predicting 30-day all-cause readmissions from hospital inpatient discharge data*. Presented at IEEE Int. Conf. E-Health Netw. Appl. Serv., 18th, Sept. 14–17, Munich, Ger.

Yu KH, Zhang C, Berry GJ, Altman RB, Re C, et al. 2016. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat. Commun.* 7:12474

Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Stat. Methodol.* 67:301–20

RELATED RESOURCES

Appenzeller T. 2017. The scientists' apprentice. *Science* 357:16–17

Lewis-Kraus G. 2016. The great A.I. awakening: how Google used artificial intelligence to transform Google Translate, one of its more popular services—and how machine learning is poised to reinvent computing itself. *The New York Times*, Dec. 14. **https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html**

Ng A. *Machine learning*. Course, Stanford Univ., Coursera. **https://www.coursera.org/learn/machine-learning**

Annual Review of Clinical Psychology

Volume 14, 2018

Contents

Errata

An online log of corrections to *Annual Review of Clinical Psychology* articles may be found at http://www.annualreviews.org/errata/clinpsy