American Psychological Association

Psychological Methods

© 2020 American Psychological Association
ISSN: 1082-989X

2021, Vol. 26, No. 2, 236-254
http://dx.doi.org/10.1037/met0000317

Psychometric and Machine Learning Approaches for Diagnostic Assessment and Tests of Individual Classification

Oscar Gonzalez
University of North Carolina at Chapel Hill

Abstract

Assessments are commonly used to make a decision about an individual, such as grade placement, treatment assignment, job selection, or to inform a diagnosis. A psychometric approach to classify respondents based on the assessment would aggregate items into a score, and then each respondent's score is compared to a cut score. In contrast, a machine learning approach to classify respondents would build a model to predict the probability of belonging to a specific class from assessment items, and then respondents are classified based on their predicted probability of belonging to that class. It remains unclear whether psychometric and machine learning methods have comparable classification accuracy or if 1 method is preferable in all or some situations. In the context of diagnostic assessment, this study used Monte Carlo simulation methods to compare the classification accuracy of psychometric and machine learning methods as a function of the diagnosis-test correlation, prevalence, sample size, and the structure of the diagnostic assessment. Results suggest that machine learning models using logistic regression or random forest could have comparable classification accuracy to the psychometric methods using estimated item response theory scores. Therefore, machine learning models could provide a viable alternative for classification when psychometric methods are not feasible. Methods are illustrated with an empirical example predicting an oppositional defiant disorder diagnosis from a behavior disorders scale in children of age seven. Strengths and limitations for each of the methods are examined, and the overlap between the field of machine learning and psychometrics is discussed.

Translational Abstract

Assessments and tests are often used to make decisions about an individual, such as deciding who will graduate from high school, who gets hired for a job, and who is referred to more services or treatment. In these tests or assessments, individuals are asked to respond to a series of items, and depending on their item responses, a decision about the individual is made. Traditionally, researchers and assessment specialists have used a psychometric approach to score item responses and then check if the score is above a cut score. However, data-driven, exploratory methods from the area of machine learning could be used for similar purposes. For example, machine learning methods could be used to predict the assessment decision directly from the item responses, without the need to aggregate item responses beforehand. It remains unclear if the classification accuracy of machine learning methods is comparable with the psychometric methods typically used for classification. This study used Monte Carlo simulation methods to study classification accuracy of psychometric and machine learning methods across a variety of assessment conditions, among which are the number of items and response categories of the assessment, the strength of the relationship between the diagnosis and the assessment, sample size, and diagnosis prevalence. The methods were also illustrated in an empirical example predicting an oppositional defiant disorder diagnosis in children of age 7. Strengths and limitations for each of the methods are examined, and the overlap between the field of machine learning and psychometrics is discussed.

Keywords: psychometrics, machine learning, classification, diagnostic assessment, item response theory

Supplemental materials: http://dx.doi.org/10.1037/met0000317.supp

One of the goals of assessment is to make a decision about an individual. There are many different types of assessments that could be used to select candidates for a job, determine if a student passes or fails a course, or inform a patient's medical condition. This article largely focuses on diagnostic assessments, which play an important role in psychology and medicine as they can be used to screen individuals for disorders or to supplement a clinician's evaluation before giving a diagnosis. Examples of diagnostic assessments or screening measures include the Child Behavior Checklist (CBCL), a measure of emotional, behavioral, and social problems used to screen children for autism spectrum disorder (Achenbach & Rescorla, 2013); the Center for Epidemiologic Studies Depression (CES-D) Scale, a measure used to screen for depression in older adults (Lewinsohn, Seeley, Roberts, & Allen, 1997); and the Parent General Behavior Inventory (P-GBI), a

This article was published Online First July 2, 2020.

Data collection for the example was supported by NIMH 55625, NIMH 55584, and NIMH 58144.

Correspondence concerning this article should be addressed to Oscar Gonzalez, Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, 235 East Cameron Avenue, Chapel Hill, NC 27599. E-mail: ogonza13@unc.edu

measure used to assess children for pediatric bipolar disorder (Youngstrom, Frazier, Demeter, Calabrese, & Findling, 2008). In these situations, the diagnostic assessment comprises a set of discrete items that measure a latent variable related to a *gold standard* that dictates the true diagnosis (e.g., a structured clinical interview or a laboratory test). A gold standard tends to be costly, extensive, invasive, or all of these, so in a primary care setting it may be more convenient to administer an assessment than a gold standard (Gibbons, Weiss, Frank, & Kupfer, 2016). After observing item responses, assessment specialists often make a binary decision about the respondent, such as to follow-up or not follow-up with the respondent, or decide if the assessment supports or does not support the clinician's diagnostic evaluation (Liu, 2012). This article discusses two statistical approaches to binary classification of respondents as diagnosed or not based on their item responses. The approaches discussed could be easily extended to settings in which classification based on achievement, aptitude, or other cognitive abilities is of interest.

Typically, diagnostic classifications are decided using a psychometric approach. First, item responses are aggregated (e.g., as a summed score or an item response theory score), and then receiver operating characteristic (ROC) curves are used to determine a cut score (Youngstrom, 2014). Respondents are classified as diagnosed or not if their aggregate score is above or below the cut score (e.g., a respondent is likely to be diagnosed with depression if the CES-D summed score is ≥16). On the other hand, a similar approach could be carried out in a machine learning framework (Lu & Petkova, 2014). Machine learning methods could use item responses directly to predict the probability that the respondent has the diagnosis. The prediction model could be built using either a parametric approach, such as logistic regression, or a nonparametric approach, such as random forest. Respondents would be classified as diagnosed or not depending on the probability of diagnosis (e.g., a respondent is likely to be diagnosed with depression if the predicted probability of diagnosis is above 50%). Across both approaches, assessment specialists often aim to maximize the proportion of respondents correctly classified as having the diagnosis (known as sensitivity), and the proportion of respondents correctly classified as not having the diagnosis (known as specificity; Gibbons, Weiss, Frank, & Kupfer, 2016; James, Witten, Hastie, & Tibshirani, 2013; Liu, 2012; Youngstrom, 2014; Zweig & Campbell, 1993).

Although both the psychometric and machine learning approaches have the same end goal, it is not currently clear how machine learning approaches compare to the psychometric approaches commonly used. The goal of this study is to compare the classification accuracy of psychometric and machine learning approaches to predict a diagnosis from assessment items, and to discuss relative benefits of using each approach. The structure of this article is the following. First, the general goals of diagnostic assessment are introduced. Second, psychometric approaches to classification using summed scores and item response theory scores are described. Third, machine learning approaches to classification using regularized logistic regression and binary recursive partitioning are described. Fourth, the methods examined are evaluated using Monte Carlo simulations and demonstrated in an empirical example. Finally, the strengths and limitations of psychometric and machine learning methods for classification tests are discussed, and future directions are considered.

Diagnostic Assessment

Diagnostic accuracy studies investigate how well a diagnostic assessment discriminates between respondents with and without a condition (Zhou, McClish, Obuchowski, 2011; Zweig & Campbell, 1993). Diagnostic accuracy studies have three components: (a) a sample of respondents; (b) item responses to the diagnostic assessment; and (c) a gold standard, independent of the diagnostic assessment, that indicates the respondent's "true" condition (Zhou et al., 2011). For example, consider a gold standard to diagnose depression (e.g., a clinical interview), where θ_{Diag} is a respondent's standing on the latent variable for the gold standard. A threshold $T_{\theta_{Diag}}$ could be imposed on θ_{Diag} to determine the true diagnosis, D_{Diag}, such as,

$$D_{Diag} = 1 \Leftrightarrow \theta_{Diag} \geq T_{\theta_{Diag}}$$

$$D_{Diag} = 0 \Leftrightarrow \theta_{Diag} < T_{\theta_{Diag}}.$$

Suppose that we are interested in examining if the Revised Hamilton Rating Scale for Depression (HRSD-R), a cheaper and less extensive measure than the clinical interview, can discriminate between clinically depressed ($D_{Diag} = 1$) and nondepressed respondents ($D_{Diag} = 0$; McFall & Treat, 1999). The items in HRSD-R assess the unobserved latent variable θ_{Assmt}, which correlates highly, but not perfectly, with θ_{Diag} (see Figure 1). The task in the diagnostic accuracy study is to establish if the true D_{Diag} can be predicted from the observed $\hat{\theta}_{Assmt}$, which are estimated latent variable scores from the HRSD-R items. As such, researchers would estimate $T_{\hat{\theta}_{Assmt}}$, the optimal cut score on $\hat{\theta}_{Assmt}$, to obtain a predicted diagnosis from the assessment D_{Assmt} that classifies respondents as depressed ($D_{Assmt} = 1$) or not depressed ($D_{Assmt} = 0$),

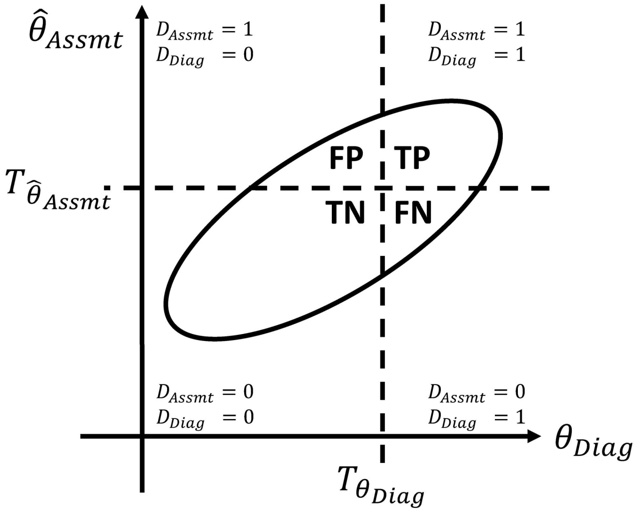

Figure 1. Relationship between the diagnosis latent variable and the assessment latent variable measured by assessment items. FP = false positive; TP = true positive; TN = true negative; FN = false negative, $\hat{\theta}_{Assmt}$ is the estimated latent variable score from the assessment, $T_{\hat{\theta}_{Assmt}}$ is the threshold that determines the predicted diagnosis D_{Assmt}, θ_{Diag} is the latent variable of the gold standard and $T_{\theta_{Diag}}$ is the threshold on θ_{Diag} that determines the true diagnosis D_{Diag}.

$$D_{Assmt} = 1 \Leftrightarrow \hat{\theta}_{Assmt} \geq T_{\hat{\theta}_{Assmt}}$$

$$D_{Assmt} = 0 \Leftrightarrow \hat{\theta}_{Assmt} < T_{\hat{\theta}_{Assmt}}.$$

The goal of the diagnostic accuracy study is to determine the similarity between the predicted diagnosis D_{Assmt} and the true diagnosis D_{Diag}. As mentioned above, the HRSD-R $\hat{\theta}_{Assmt}$ is not perfectly correlated with the diagnosis latent variable θ_{Diag}, so it will misclassify respondents. For example, lowering $T_{\hat{\theta}_{Assmt}}$ (sliding down the horizontal dashed line in Figure 1) would correctly classify most respondents who are depressed (high sensitivity), but would also misclassify many nonclinically depressed respondents as depressed (low specificity). On the other hand, increasing $T_{\hat{\theta}_{Assmt}}$ (i.e., sliding up the horizontal dashed line in Figure 1) would correctly classify most nondepressed respondents (high specificity), but would misclassify many respondents who are depressed (low sensitivity). A strategy to determine $T_{\hat{\theta}_{Assmt}}$ is to study the trade-off between sensitivity and specificity using ROC curves (further discussed in the Method section); however, choosing $T_{\hat{\theta}_{Assmt}}$ depends on the goal of the application—maximizing true positives or true negatives. It is important to highlight that the goal of the HRSD-R, and of any other diagnostic or screening measure, is not to replace the gold standard, but to inform assessment specialists about who they should follow up in a primary care setting or to supplement a clinician's diagnostic evaluation.

From a psychometric perspective, diagnostic assessment focuses on aggregating item responses to obtain a precise estimate of $\hat{\theta}_{Assmt}$. If $\hat{\theta}_{Assmt}$ is precise, then one might be able to determine a D_{Assmt} that closely resembles D_{Diag}. In this case, there is an indirect prediction from the item responses to D_{Diag} through $\hat{\theta}_{Assmt}$ because, typically, information about D_{Diag} is not considered in the model that estimates $\hat{\theta}_{Assmt}$. From a machine learning perspective, diagnostic assessment focuses on building a model that predicts D_{Diag} directly from the item responses. In the sections below, psychometric methods to estimate $\hat{\theta}_{Assmt}$ using item response theory and machine learning methods to predict D_{Diag} from item responses are discussed.

Psychometric Methods

As mentioned before, psychometric methods for diagnostic assessment involve the aggregation of the observed item responses, and then respondents are classified as diagnosed or not diagnosed if the score is above or below a cut score, respectively. One psychometric approach for diagnostic assessment uses classical test theory, where a total score of the items is used for classification (Gibbons et al., 2013). A total score is usually estimated by summing item responses, where each item response is unit-weighted. Another psychometric approach for diagnostic assessment uses item response theory (IRT) to obtain scores for classification. IRT refers to a family of mathematical models used for two main purposes: item analysis and scoring item responses (Thissen & Steinberg, 2009). IRT has a long history and underlies many high-stakes and large-scale assessments (Thissen & Wainer, 2001). An appropriate IRT model for Likert-type items is the graded response model (GRM; Samejima, 1969), expressed as,

$$P(u_i = j \mid \theta, b_{ij}, a_i) = \frac{1}{1 + \exp[-a_i(\theta - b_{ij})]} - \frac{1}{1 + \exp[-a_i(\theta - b_{i(j+1)})]}. \quad (1)$$

The GRM indicates that the probability of endorsing a response category depends on properties of the respondent and properties of the items. Specifically, θ refers to the respondent's standing on the latent variable assessed by the items (singular in this case as this is a unidimensional version of the GRM), and the *a*- and *b*-parameters describe the item. The *a*-parameter indicates the strength of the relation between the item and the latent variable (analogous to a factor loading), and the *b*-parameters indicate the θ value where the respondent has a 50% probability of endorsing category *j* or a higher category. If the item consists of *J* response categories, then there would be *J*-1 *b*-parameters. By definition, the probability of responding to the lowest category or higher is 1, and the probability of responding higher than the highest category $\{J\}$ is 0 (Edelen & Reeve, 2007). For items with two response categories, the GRM reduces to the two-parameter logistic model (2PLM; Birnbaum, 1968)

$$P(u_i = 1 \mid \theta, a_i, b_i) = \frac{1}{1 + \exp[-a_i(\theta - b_i)]}, \quad (2)$$

a popular IRT model for measures comprised of binary items (with symbols as defined above).

The estimation of item parameters and θ values is commonly carried out in two steps—the *calibration* step and the *scoring* step. For calibration, a common approach to estimate item parameters is marginal maximum likelihood estimation using the expectation-maximization (EM) algorithm (MML-EM; Bock & Aitkin, 1981; de Ayala, 2009). MML-EM assumes that respondents are randomly sampled from a population, so their θ values are marginalized over during the calibration step and only item parameters are obtained. In the scoring step, the estimated item parameters are treated as fixed and θ values are estimated per response pattern (van der Linden & Pashley, 2010). A common scoring method to estimate θ values is expected a posteriori (EAP) scoring (Thissen & Steinberg, 2009). The EAP$[\hat{\theta}]$ is the mean of the distribution defined by the product of the likelihood of each of the response patterns and the prior distribution of the θ values (typically assumed to follow a normal distribution). The variability of the EAP$[\hat{\theta}]$ estimate can also be estimated by the standard deviation of the distribution just described. EAP$[\hat{\theta}]$ estimates are considered shrunken estimates because they pull the EAP$[\hat{\theta}]$ toward the mean of the prior and have a smaller variance than the likelihood. Previous research suggests that EAP$[\hat{\theta}]$ scores could prevent poor score estimation when the likelihood is bimodal and are more robust to IRT model misspecifications than other scoring methods (Thissen & Wainer, 2001).

IRT models require meeting certain assumptions so the estimates can be meaningfully interpreted. IRT models assume that the correct number of latent variables that influence item responses have been specified. The 2PLM and the GRM described above assume only one latent variable, but multidimensional versions of the 2PLM and GRM are also available (Gibbons & Hedeker, 1992; Reckase, 2009). Closely related, IRT models also assume local independence, where the joint probability of endorsing two items is the product of the probabilities of endorsing each of the items (i.e., that items are not related to each other conditional on θ). Two sources of local dependence are unmodeled multidimensionality, known as underlying local dependence, or similar item content or item location in the assessment, known as surface local dependence (Edwards, Houts, & Cai, 2018). Finally, a distribution (typically standard normal) must be chosen for the latent variable. The present study assumes that the assessment is unidimensional and, in most conditions, that the items are locally independent.

Overall, the psychometric approach to diagnostic assessment estimates a score from the assessment items, and then makes a decision about the diagnosis by comparing the score to a *cut score*. A score could be estimated by summing all item responses or by fitting an IRT model to estimate an EAP[$\hat{\theta}$] score. A strength of the psychometric methods is that they focus on improving the measurement precision of an assessment, so that the scores produced are reliable and interpretable. Potential problems with the psychometric methods are that the prediction of the diagnosis is a by-product measurement process; and that classification accuracy of the diagnosis might depend on how closely the item responses meet the assumptions of the scoring model (Gibbons et al., 2013). Also, as the complexity of the psychometric model increases, a larger sample size would be needed to provide stable parameter estimates (Reise & Yu, 1990).

Machine Learning Methods

Machine learning is a subfield of artificial intelligence that deals with data-driven, statistical algorithms that learn patterns from data and that improve through experience (Jordan & Mitchell, 2015; Witten, Frank, Hall, & Pal, 2016). This article largely discusses supervised machine learning models, which are models that learn by example. The learning process involves a dataset with an observed outcome and predictors, and an algorithm that determines the relations between the predictors and the outcome. A goal of machine learning is to examine the likelihood that the relations found would generalize to other cases (Yarkoni & Westfall, 2017). Therefore, the general machine learning framework is to learn the patterns/relations and build a model using one part of the data (referred to as a training dataset), and then evaluate model performance on a different part of the data (referred to as a testing dataset). It is important to highlight that the predictors and the outcome are observed in the testing dataset, so model evaluation involves comparing the observed outcome in the testing dataset to the predicted outcome by the model given the predictors in the testing dataset.

For diagnostic assessment, machine learning approaches could be used to build a model to predict the probability of diagnosis from individual item responses, and then classify respondents depending whether the probability of diagnosis is above or below a threshold. The models could be built using a parametric or a nonparametric approach. Here, a parametric approach prespecifies the functional form between the predictors and the outcome, such as using regularized logistic regression to predict the diagnosis (i.e., logistic regression assumes that the predictors combine linearly to predict the outcome). In contrast, a nonparametric model uses a data-driven approach to determine the functional form between the predictors and the outcome, such as using binary recursive partitioning to predict the diagnosis. Below, regularized logistic regression and binary recursive partitioning are discussed.

Logistic Regression and Regularization

For our case, consider predicting the probability of a binary diagnosis using a linear model with two predictors,

$$P(Y = 1 \mid \boldsymbol{X}) = b_o + b_1 X_1 + b_2 X_2. \tag{3}$$

Probability is bounded by 0 and 1, so it might not be appropriate to use OLS regression—a linear function might make predictions outside of that bound. One can turn to an S-shaped, logistic function bounded by 0 and 1. The logistic function is,

$$\textit{Logisticfunction}: P(Y = 1 \mid \boldsymbol{X}) = \frac{e^{b_o + b_1 X_1 + b_2 X_2}}{1 + e^{b_o + b_1 X_1 + b_2 X_2}}. \tag{4}$$

After manipulation, one can determine that there is a linear relationship between the natural logarithm of the odds and the predictors,

$$\ln\left(\frac{P(Y = 1 \mid X)}{1 - P(Y = 1 \mid X)}\right) = b_o + b_1 X_1 + b_2 X_2. \tag{5}$$

All else constant, the regression coefficients indicate the change in the logit (log of the odds unit) of the outcome for a one-unit change in the predictor. The regression coefficients are typically estimated using maximum likelihood. Below is the log-likelihood function for logistic regression,

$$l(\beta) = \sum_{i=1}^{N} \left\{ y_i \boldsymbol{\beta}^T x_i - \log\left(1 - e^{\boldsymbol{\beta}^T x_i}\right) \right\}, \tag{6}$$

where $\boldsymbol{\beta}^T$ is a $q + 1$ vector of Q regression coefficients and the intercept. The coefficients $\boldsymbol{\beta}^T$ can then be used to estimate predicted probabilities (from Equation 4), and in turn predicted classes (such as a predicted diagnosis) conditional on a probability threshold.

Moreover, regularization could be used to improve prediction and interpretability of the logistic regression model (Hastie, Tibshirani, & Friedman, 2009; Hastie, Tibshirani, & Wainwright, 2015; Tibshirani, 1996). Logistic regression coefficient estimates are unbiased for the data where the model is developed, but if the estimated coefficients are used to predict outcomes in new data, these predictions may be inaccurate. Regularization involves fitting a logistic regression model to the data, but adds a penalty to the logistic regression log-likelihood function to control coefficient size and shrink the coefficients toward zero. Two penalties for regularization discussed in this article are the ridge penalty and the Lasso penalty. The loss function for ridge logistic regression augments the logistic regression log-likelihood function from Equation 6 by including a penalty to control for the coefficient size,

$$l(\beta)^{L2} = \sum_{i=1}^{N} \left\{ y_i \boldsymbol{\beta}^T x_i - \log\left(1 - e^{\boldsymbol{\beta}^T x_i}\right) \right\} - \lambda \sum_{q=1}^{Q} \beta_Q^2. \tag{7}$$

In ridge logistic regression, the squared value of the size of the coefficient is penalized and coefficients are shrunk toward zero. The impact of the penalty is controlled by the nonnegative tuning parameter λ, determined by k-fold cross-validation to reduce prediction error. On the other hand, the Lasso augments the logistic regression log-likelihood function to control for the coefficient size as shown below,

$$l(\beta)^{L1} = \sum_{i=1}^{N}\left\{y_i\boldsymbol{\beta}^{\boldsymbol{T}}x_i - \log(1 - e^{\boldsymbol{\beta}^{\boldsymbol{T}}x_i})\right\} - \lambda\sum_{q=1}^{Q}|\beta_q|. \quad (8)$$

In Lasso logistic regression, the absolute value of the size of the coefficient is penalized and it could shrink some regression coefficients to exactly zero, acting as a proxy for variable selection and helping on model interpretability. Similarly, the weight of the penalty is also controlled by λ.

Regularization would induce some bias to the regression coefficients (i.e., parameter estimates and prediction may not be optimal for our specific data), but prediction would be less variable when those coefficients are used to predict outcomes in new data. If one wanted to estimate unbiased coefficients for ridge logistic regression, one would fit the full logistic regression model. If one wanted to estimate unbiased coefficients for Lasso, but still focus on variable selection, an unrestricted logistic regression model could be fitted again only with the variables selected by the Lasso. This approach is known as the relaxed Lasso (Jacobucci, Grimm, & McArdle, 2016).

Binary Recursive Partitioning

Instead of assuming a linear functional form between the predictors and the outcome, binary recursive partitioning could be used to determine the functional form between the predictors and the outcome. Binary recursive partitioning is the algorithm behind classification and regression trees (CART), an umbrella term that refers to using decision trees to predict continuous (regression trees) or categorical (classification trees) outcomes (Breiman, Friedman, Stone, & Olshen, 1984). In sum, trees are useful, easy to interpret, and they do not depend on a functional form (Strobl, Malley, & Tutz, 2009).

In general, CART selects the best binary split among all unique predictor values to partition the data into two regions where outcome values are most similar. Then, the procedure is carried out recursively in each derived region until the algorithm meets a stopping rule. In this case, splits depend on the predictor type. For a continuous predictor, splits are considered along all unique values of the predictor. For example, suppose that an item response is a predictor of the diagnosis; a proposed partition could be defined by respondents who scored less than 3 on the item and respondents who scored 3 or higher. There are *m*-1 candidate splits when a predictor has *m* unique values. On the other hand, splits on unordered categorical predictors are defined in the one (or some)-versus-rest form for all combination of predictors. For example, suppose that the clinician is a predictor of diagnosis. A proposed partition could be defined by respondents who have Johnny as their doctor and those who do not have Johnny as their doctor (such as those who have Kelly or Rick). There are $2^{m\text{-}1}$-1 candidate splits when a predictor has *m* unordered categories. For classification trees (which would be used to predict a binary diagnosis), the best split could be defined as the split that minimizes the misclassification rate of the outcome in each region. The misclassification rate is the proportion of cases in which the observed outcome and predicted outcome differ. The observed outcome comes from the data, and the predicted outcome is the mode of the outcome in each of the derived regions (note that all cases in the region have the same predicted value). Once the best split has been decided, the procedure is carried out again in the defined regions until the algorithm meets a stopping rule, often based on prediction improvement or a minimum number of cases in each region. The regions derived by CART could be summarized in a decision tree, and the branches of the trees are interpreted as interactive effects of predictors on the outcome.

Binary recursive partitioning is considered a greedy algorithm because once it decides on a predictor to split on, that decision is not reconsidered even though splitting on a different predictor could have been better for prediction accuracy across the whole model. As a result, CART could learn the idiosyncrasies of the data (also referred to as overfitting the data) and yield inaccurate predictions when the tree predicts outcomes in new data. Cost-complexity pruning, a CART extension, could help overcome some of these limitations (James et al., 2013). Cost-complexity pruning consists in overgrowing a tree in a training dataset (e.g., growing a tree with many splits that learns all of the idiosyncrasies of the data), and then pruning back branches of the tree with the goal of reducing prediction error in the testing dataset. Cost-complexity pruning introduces a penalty parameter that controls the trade-off between tree size and prediction accuracy (James et al., 2013). The final tree size is determined by *k*-fold cross-validation, where the number of regions remaining in the final tree is the size that minimizes the prediction error in the left-out fold. Predictions based on the pruned tree in new data are likely to be more accurate than predictions based on the overgrown tree.

The random forest algorithm is another CART extension that can improve the tree stability and prediction in new data (Breiman, 2001). Random forest grows an ensemble of decision trees in bootstrapped datasets and averages predictions across trees. However, the random forest algorithm grows the trees slightly differently than the trees grown for CART—every time that one of the trees makes a split, a random subset of predictors is chosen as candidates to split on. Splitting on a random subset of predictors helps to increase tree diversity in bootstrapped datasets, and it is expected to increase both tree stability and prediction accuracy. An important consideration of using random forest is that the interpretability of the model is lost in exchange for prediction—the model is the whole ensemble of many trees, not just one tree. In this case, variable importance measures could be used to identify the predictors that affect prediction accuracy the most.

Overall, the machine learning methods for diagnostic assessments build a model to predict the probability of diagnosis using assessment items, and then determine a diagnosis if the probability is above or below a threshold. The model could be built using a parametric approach, such as logistic regression with regularization (Lu & Petkova, 2014), or a nonparametric approach, such as random forest or classification trees with binary recursive partitioning (Gibbons et al., 2016; McArdle, 2013; Yan, Lewis, & Stocking, 2004). A strength of the machine learning methods is that the prediction of the diagnosis is the main product of the analysis. A potential problem with the machine learning methods is that items are assumed to be perfect predictors, when in fact they suffer from measurement error.

Expected Differences Between Psychometric and Machine Learning Models

There are several strengths and weaknesses in psychometric and machine learning models that could be expected given the properties of each approach. First, suppose that there was a true linear relation between the items and a function of the diagnosis. Approaches like the psychometric models, logistic regression, ridge logistic regression, Lasso, and relaxed Lasso would be expected to perform well because they are based on a linear model. On the other hand, the performance of classification trees and random forest would depend on how well they are able to approximate the linear relation between the items and diagnosis with step functions via binary recursive partitioning (James et al., 2013). However, if the relation between the items and some function of the diagnosis is nonlinear, the opposite would be expected—approaches that use binary recursive partitioning would be able to approximate the nonlinear relation, while methods restricted to a linear model would miss it. Furthermore, suppose that there are item interactions in the dataset, where the relation between a reported symptom and the diagnosis depends on the level of a different reported symptom. Methods that rely on binary recursive partitioning could naturally find and accommodate interactions among the items, while psychometric and machine learning models that assume linear relations would have some difficulty handling nonhypothesized interactions—psychometric models typically do not accommodate interactions among items, and (regularized) logistic regression requires the prespecification of the interactions. Finally, psychometric models might have estimation problems in the presence of both underlying and surface local dependence. In this case, items are excessively correlated, which might lead to instability in the item parameter estimates. On the other hand, machine learning methods that rely on regularization could treat locally dependent items as redundant and might set the coefficient of one of those items to zero or close to zero. Lastly, for classification trees and random forest, the effect of local dependence on classification might depend on the type of local dependence. In the case of underlying local dependence, classification trees and random forest might accommodate locally dependent items in one branch of the tree, and in the case of surface local dependence, only one of the items would be selected to split on. Also, note that highly correlated items could bias the variable importance measures from random forest (Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008), which could be addressed with a data preprocessing step as recommended in many machine learning applications (Kuhn & Johnson, 2013). The present study includes some conditions to study how surface local dependence affects classification accuracy, but nonlinear relations between the items and the diagnosis and interactions between the items are not studied.

Present Study

Researchers could use a psychometric or a machine learning approach to diagnostic assessment, but it is less clear which approach to use. In general, psychometric models have worked well in previous large-scale applications (Reeve et al., 2007), so the relations suggested by psychometric models might represent a decent approximation to the relation between assessment items and diagnoses. Consequently, this study examines if the machine learning methods approximate the performance of the commonly used psychometric methods when the relations between the assessment items and the diagnosis are simulated from a psychometric model. Specifically, the Monte Carlo simulation examines how classification accuracy across the methods is influenced by sample size, prevalence of the diagnosis, the diagnosis-test correlation, and assessment structure. Also, the methods are demonstrated by predicting an oppositional defiant disorder (ODD) diagnosis from the ODD subscale in the Disruptive Behavior Disorders Rating Scale (Pelham, Gnagy, Greenslade, & Milich, 1992) in a sample of children of age 7.

For the simulation, it is hypothesized that classification accuracy across models would increase as sample size, number of item categories, number of items, prevalence of the diagnosis, and the diagnosis-test correlation increase. Given that there is a linear relation between the items and a function of the diagnosis, it is expected that psychometric models and machine learning models that assume linear relations would perform better than classification trees and random forest, although it would depend on how close the trees are able to approximate the linear relation. Also, it is expected that psychometric models would yield less accurate item parameter estimates in conditions in which items exhibit local dependence, while the estimation of the machine learning models would not be affected. Finally, it is hypothesized that psychometric methods would outperform machine learning methods when the prevalence and diagnosis-test correlation are low because machine learning models might struggle to find relations affected by measurement error, although this might depend on the number of predictor items in the model.

Method

Data-Generation

Data-generating theta and true diagnosis. For each respondent, data were generated by drawing two continuous, positively correlated variables, θ_{Assmt} and θ_{Diag}, from a bivariate normal distribution. As mentioned above, the variable θ_{Diag} was the score on a latent variable that determined the respondent's true diagnosis (Loeffelman et al., 2020), such as,

$$D_{Diag} = 1 \Leftrightarrow \theta_{Diag} \geq T_{\theta_{Diag}}$$

$$D_{Diag} = 0 \Leftrightarrow \theta_{Diag} < T_{\theta_{Diag}},$$

where D_{Diag} is the true diagnosis and $T_{\theta_{Diag}}$ is a threshold derived by the quantiles of the normal distribution to yield a specific diagnosis prevalence. If $D_{Diag} = 1$, then respondents had the diagnosis, and if $D_{Diag} = 0$, then respondents did not have the diagnosis. The variable θ_{Assmt} represents the respondent's data-generating IRT score on the assessment, which in turn was used to generate the item responses.

Item responses. IRT item parameters were dependent on the number of item categories and were sampled at each replication. The distribution of item parameters likely mimics the parameters found in traditional assessments (Hill, 2004). Binary items only have one threshold, so the *b*-parameters were simulated from the normal distribution *N*(0, 1), and the *a*-parameters were simulated from the normal distribution *N*(1.7, .3). Polytomous items have *J*-1 thresholds for items with *J* categories, so the first *b*-parameter was simulated from the normal distribution *N*(−0.6, 1), and the remaining *b*-parameters were sequentially higher than the previous one by adding a random number from the uniform distribution *U*(.5, .9). Datasets included in this study had each item response endorsed at least five times across all items. The *a*-parameters for polytomous items were also simulated from the normal distribution $N(1.7, .3)$. With a simulated θ_{Assmt} and item parameters, the IRT models in either Equation 1 or Equation 2 were used to obtain the expected probability of endorsing each item category per item. Finally, the item response was determined by comparing if the expected probability of the model was higher than a random probability from the uniform distribution $U(0, 1)$. Generally, the psychometric and machine learning approaches would use the item responses simulated from θ_{Assmt} to predict the binary diagnosis determined from θ_{Diag}.

Simulation factors. There were six factors varied in this simulation: training sample size ($N = 250, 500, 1{,}000$), number of items (10, 30), number of item categories (two, five), correlation between θ_{Assmt} and θ_{Diag}, referred to as the diagnosis-test correlation, (.30, .50, .70), prevalence of the diagnosis in the training sample (.05, .10, .20), and violation of the local independence assumption of IRT models (no violation or for five items in 30% of the sample). Specifically, this simulation examined surface local dependence, where participants might respond to a set of items identically because of similar item content or item location (Edwards et al., 2018). In other words, the IRT model does not determine the responses to the item, but the item response is conditional on the response to another item. In the condition of $LD = .30$, data were generated so that 30% of the participants had identical responses to the first five items, regardless of the expected response to the item given by the θ_{Assmt} value. Overall, there were 216 conditions, with 500 replications per condition.

Determining Cut Scores and Probability Thresholds

The psychometric and machine learning models depend on a threshold to classify respondents, such as a cut score or probability threshold. Traditionally, machine learning methods focused on reducing prediction error and assigned respondents to the diagnosis class that is most likely. This practice would translate to using a fixed probability threshold at .50, and cases with a predicted probability of diagnosis above 50% would be classified as diagnosed (James et al., 2013). When machine learning approaches use a fixed .50 probability threshold for classification, we say that the approach uses a Bayes classifier.

On the other hand, a valuable tool to determine thresholds in the psychometric and machine learning frameworks is a receiver operating characteristic (ROC) curve (Egan, 1975; McFall & Treat, 1999; Pepe, 2003; Zou, O'Malley, & Mauri, 2007; Zweig & Campbell, 1993). Empirical ROC curves describe how sensitivity and specificity for classification change as a function of the cut scores or probability thresholds. In this case, sensitivity and specificity are estimated using a 2 × 2 confusion table that cross-tabulates the observed diagnosis in the data and the predicted diagnosis by the model for a specific threshold (see footnote on Table 1 for equations to estimate classification accuracy statistics). In essence, a 2 × 2 confusion table is computed for each unique cut score (for the psychometric approach) or unique predicted probability (for the machine learning approach), and the sensitivity and specificity to identify respondents with or without the diagnosis are estimated and plotted. The shape and height of the ROC curve indicate how well the diagnostic assessment score or the model discriminates, and allows for a graphical comparison across the different approaches. A 45-degree line (the chance diagonal) can be used as a reference because it represents an assessment that makes diagnoses at random—the true positive rate and the false positive rate are the same. Also, the area under the ROC curve (AUC) could be used as a summary statistic to compare different models. The AUC indicates the likelihood of making correct classifications when two cases (one per group) are chosen at random, which is also interpreted as the average sensitivity (or specificity) across all values of specificity (or sensitivity). The AUC of the chance diagonal is .50, and an AUC close to 1.0 indicates that the assessment classifies respondents almost perfectly. The ROC curve can be used to determine a threshold with a specific property, such as balancing sensitivity and specificity. As such, researchers could choose the threshold that maximizes the Youden index (Liu, 2012),

Table 1
Sample 2 × 2 Table to Estimate Classification Accuracy

	Gold standard	
Score or pred. probability	Not diagnosed	Diagnosed
Below threshold	TN	FN
Above threshold	FP	TP

Note. TP = true positive; FP = false positive; FN = false negative; TN = true negative; sensitivity = TP/(TP + FN); specificity = TN/(TN + FP); classification rate = (TP + TN)/(TP + TN + FP + FN).

$$\textit{Youden index}: Max\{Se + Sp - 1\}. \tag{9}$$

Graphically, the Youden index is the point in the ROC curve that is farthest from the chance diagonal. Depending on the assessment application, researchers might give more value to finding more true positives or more true negatives, so ROC curves allow researchers to operate along the curve and investigate how changing the threshold affects sensitivity or specificity. When psychometric or machine learning approaches use an ROC curve to determine the cut score or probability threshold for classification, we say that the approach uses a ROC classifier.

Once a threshold is determined, then the threshold is imposed on either the scores from the psychometric approach or the predicted probabilities from the model to get a predicted diagnosis, D_{Assmt}. For the psychometric approach, if T_{EAP} is the EAP$[\hat{\theta}]$ cut score that maximizes the Youden index using a ROC classifier, then,

$$D_{Assmt} = 1 \Leftrightarrow \mathrm{EAP}[\hat{\theta}] \geq T_{EAP}$$
$$D_{Assmt} = 0 \Leftrightarrow \mathrm{EAP}[\hat{\theta}] < T_{EAP}.$$

In this case, if $D_{Assmt} = 1$, then the predicted class is that the respondent is diagnosed, and if $D_{Assmt} = 0$, then the predicted class is that the respondent is not diagnosed. Note that similar steps would be used to determine D_{Assmt} from the summed score. For machine learning approaches, if threshold T_{RF} is the predicted probability by a random forest model, $\mathrm{RF}_{\mathrm{pred}}$, that maximizes the Youden index using a ROC classifier, then,

$$D_{Assmt} = 1 \Leftrightarrow \mathrm{RF}_{\mathrm{pred}} \geq T_{RF}$$
$$D_{Assmt} = 0 \Leftrightarrow \mathrm{RF}_{\mathrm{pred}} < T_{RF}.$$

Similar steps would be used to determine D_{Assmt} from the predicted probabilities of other machine learning models, or when the Bayes classifier threshold ($T_{RF} = .50$) is used.

Data Analysis

The simulated datasets were analyzed with three psychometric models: the data-generating θ_{Assmt}, the raw summed score, and the estimated $\hat{\theta}_{Assmt}$, which is EAP[$\hat{\theta}$]. Also, the simulated datasets were analyzed with six machine learning models: logistic regression, ridge logistic regression,[1] Lasso logistic regression, relaxed Lasso logistic regression, CART, and random forest. Therefore, there are nine different statistical approaches to diagnostic assessment examined in this study. Five out of the six machine learning approaches predicted the diagnosis with two different classifiers: the Bayes classifier with a fixed threshold at .50 and the ROC classifier with a data-specific threshold determined by maximizing the Youden index (note that CART with a ROC classifier was not used given the limited variability in the predicted probabilities). Overall, there were 14 models of analysis.

For the psychometric models, a training dataset was used for item calibration (for IRT models) and to determine the cut score using ROC curves. For the machine learning models, a training dataset was used to build models to predict the probability of diagnosis and also to determine a probability threshold using ROC curves which was then used by the ROC classifier to assign a diagnosis. Note that the machine learning algorithms did not determine class assignment directly; instead, the predicted probability of diagnosis by the machine learning model was compared to a probability threshold. Finally, classification accuracy for both psychometric and machine learning models were evaluated in a testing dataset of $N = 5{,}000$. For the machine learning models, most tuning parameters were dataset-specific and determined using 10-fold cross-validation, except in conditions with either a small sample size ($N = 250$) or low prevalence (5%) where both five-fold cross-validation and stratified sampling were used to guarantee that every fold had both diagnosed and nondiagnosed cases. More information about the specification and metaparameter tuning of each model can be found in the Appendix. Finally, the main simulation outcomes were the classification rate (proportion of respondents correctly classified), sensitivity (proportion of respondents correctly identified as having the diagnosis), and specificity (proportion of respondents correctly identified as not having the diagnosis). Simulation outcomes were analyzed using both linear regression and random forest. For linear regression, the simulation outcomes were predicted by the simulation factors, including all possible interactions. For random forest, the simulation outcomes were predicted by the simulation factors in half of the Monte Carlo simulated datasets (training datasets) and variable importance measures were obtained, and the model was evaluated in the remaining half of the Monte Carlo simulated datasets (testing datasets; Gonzalez, O'Rourke, Wurpts, & Grimm, 2018). Linear regression effect sizes and random forest variable importance measures were used to identify which simulation factors are important predictors of classification rate, sensitivity, and specificity.

When models are estimated in the simulation, there may be situations in which IRT models do not converge or machine learning models assign respondents only to the most prevalent class in order to maximize prediction accuracy. Therefore, simulation conditions discussed in this study had at least 50% of the models across replications converge, or at least 50% of the models across replications assign any respondents to the diagnosis (least prevalent) class. Furthermore, the median of classification accuracy indices per model (see Table 2) and effect sizes, such as R^2 and partial-η^2, were used to guide the presentation of the results. Regression models with R^2 greater than .01 were further examined, and predictors with partial-η^2 greater than .01 and high variable importance measures were interpreted. Detailed tables for all relevant analyses are presented in the online supplemental materials.

Results

In this section, the term classification accuracy is used to refer to classification rate, sensitivity, and specificity jointly. Preliminary analyses suggest that .02% of the IRT models did not converge, but that the θ_{Assmt} and item parameters were appropriately recovered in the vast majority of cases. Violations of local dependence slightly influenced item parameter recovery, but they did not impact classification accuracy, so results presented are averaged across local dependence conditions. Machine learning models with a Bayes classifier had very low sensitivity and inflated specificity compared to the sensitivity and specificity of the data-generating model (see Table 2). Therefore, machine learning methods using a Bayes classifier are not discussed.

Classification Accuracy of the Psychometric Models

Linear models predicting classification accuracy differences suggested that there were no practical differences in classification accuracy between using data-generating θ_{Assmt} and EAP[$\hat{\theta}$] (classification rate $R^2 = .003$; sensitivity $R^2 = .001$; and specificity $R^2 = .002$); between using data-generating θ_{Assmt} and the raw summed scores (classification rate $R^2 = .004$; sensitivity $R^2 = .001$; and specificity $R^2 = .003$); and between using the EAP[$\hat{\theta}$] and the raw summed scores (classification rate $R^2 = .001$; sensitivity $R^2 = .001$; and specificity $R^2 = .001$). It is likely that these models performed similarly because the raw summed score and EAP[$\hat{\theta}$] were highly correlated. Therefore, results are presented only for classification using EAP[$\hat{\theta}$].

The classification rate using EAP[$\hat{\theta}$] ranged from .58 to .80; sensitivity ranged from .58 to .82; and specificity ranged from .59 to .79. In a regression predicting classification rate of EAP[$\hat{\theta}$] from the simulation factors, the variance explained by the predictors was $R^2 = .337$. The classification rate increased as the diagnosis-test correlation increased ($b = .146$, $SE = .005$, $t = 26.964$, $p < .001$, partial-$\eta^2 = .321$) and as the prevalence decreased ($b = -.032$, $SE = .005$, $t = -5.887$, $p < .001$, partial-$\eta^2 = .021$). In a regression predicting sensitivity from the simulation factors, the variance explained by the predictors was $R^2 = .280$. Sensitivity increased as the diagnosis-test correlation increased ($b = .175$, $SE = .007$, $t = 24.010$, $p < .001$, partial-$\eta^2 = .272$). In a regression predicting specificity from the simulation factors, the variance explained by the predictors was $R^2 = .236$. Specificity increased as the diagnosis-test correlation increased ($b = .144$, $SE = .007$, $t = 21.177$, $p < .001$, partial-$\eta^2 = .220$) and as the prevalence decreased ($b = -.039$, $SE = .007$, $t = -5.793$, $p < .001$, partial-$\eta^2 = .017$). Overall, the best predictors of classifica-

[1] Note that analyses and estimates for ridge logistic regression were carried out in additional simulated datasets. Item parameter recovery for the psychometric models and estimates from the remaining machine learning models were comparable in the additional stimulated datasets.

Table 2
Median of Classification Accuracy Indices by Model

	Classification rate			Sensitivity			Specificity		
Prevalence	.05	.10	.20	.05	.10	.20	.05	.10	.20
Data-generating θ	.73	.71	.68	.74	.72	.71	.73	.71	.69
EAP[$\hat{\theta}$]	.72	.69	.67	.73	.72	.70	.72	.69	.67
Raw summed score	.71	.69	.67	.74	.72	.70	.71	.68	.67
CART$^{\wedge}$	—	—	.74	—	—	.26	—	—	.86
RF Bayes classifier	—	—	.80	—	—	.14	—	—	.96
Logistic reg Bayes classifier$^{@}$	.94	.89	.79	.05	.08	.19	.98	.98	.94
Ridge logistic reg Bayes classifier$^{\$}$	—	—	.82	—	—	.15	—	—	.99
Lasso logistic reg Bayes classifier$^{\%}$	—	—	.82	—	—	.16	—	—	.99
Relaxed lasso logistic reg Bayes classifier$^{\#}$	.95	.90	.82	.13	.21	.37	.99	.98	.94
RF ROC classifier	.67	.65	.64	.71	.71	.70	.67	.64	.64
Logistic reg ROC classifier	.67	.66	.65	.72	.71	.69	.67	.66	.65
Ridge logistic reg ROC classifier***	.76	.75	.73	.84	.81	.78	.76	.74	.73
Lasso logistic reg ROC classifier***	.74	.73	.72	.81	.79	.76	.74	.72	.71
Relaxed lasso reg ROC classifier***	.75	.73	.73	.81	.79	.76	.74	.73	.72

Note. Reg = regression; ROC = receiver operating characteristic; RF = random forest.
$^{\wedge}$ Only for conditions greater than $N = 250$. $^{\$}$ Only for conditions with five-category items, diagnosis-test correlation of .70 and sample size of $N =$ 1,000. $^{\%}$ For conditions with five-category items, diagnosis-test correlation of .70, and sample size greater than $N = 250$. $^{\#}$ For conditions with five-category items and diagnosis-test correlation of .70. $^{@}$ For conditions with 30 items. — is for conditions not analyzed because less than 50% of replications per condition assigned cases to the least prevalent class.
*** Only for conditions with diagnosis-test correlation of .70. For data-generating θ, classification rates were [.80, .77, .75], sensitivity were [.83, .81, .78], and specificity were [.80, .77, .75] for prevalence [.05, .10, .20].

tion accuracy in psychometric models were the diagnosis-test correlation and prevalence.

Classification Accuracy of Machine Learning Models With ROC Classifiers

Table 3 shows the R^2 and the partial-η^2 effect sizes for the predictors of classification accuracy, and Figure 2 shows the variable importance measures for predictors across machine learning methods. Linear models predicting classification accuracy differences suggested that there were no practical differences in classification accuracy between Lasso logistic regression and relaxed Lasso logistic regression as a function of simulation factors (classification rate $R^2 = .004$; sensitivity $R^2 = .004$; specificity $R^2 = .004$). Therefore, only relaxed Lasso estimates are discussed in the rest of the document.

Logistic regression and random forest. For logistic regression and random forest, the diagnosis-test correlation was the most important predictor of classification accuracy according to the variable importance measures. For random forest, the number of items, prevalence, and number of item categories also influenced sensitivity. On average, sensitivity increased as the diagnosis-test correlation, number of item categories, prevalence, and number of items increased. However, as number of items and item categories increased, differences in sensitivity across prevalence decreased.

Ridge logistic regression and relaxed lasso. For ridge logistic regression and the relaxed Lasso, only the conditions with diagnosis-test correlation of .70 were likely to assign cases to the diagnosis class. In conditions with a diagnosis-test correlation of .70, predictors of classification accuracy for both models were the main effects of item categories, sample size, number of items, and prevalence. On average, classification accuracy increased as sample size, number of items, and number of item categories increased, and classification accuracy decreased as prevalence increased. Given that ridge logistic regression and relaxed Lasso required a substantially large correlation between the assessment and the diagnosis to detect the least prevalent class, it is unlikely that they would be the preferred tools to predict a diagnosis from item responses. However, it might still be interesting to compare ridge logistic regression and the relaxed Lasso to random forest and logistic regression in conditions where the diagnosis-test correlation is .70 to understand more about their performance. Linear models predicting classification accuracy differences suggested that there were differences in classification accuracy between ridge logistic regression and logistic regression (classification rate $R^2 =$ .074; sensitivity $R^2 = .306$; and specificity $R^2 = .056$) and between ridge logistic regression and random forest (classification rate $R^2 = .141$; sensitivity $R^2 = .420$; and specificity $R^2 = .138$). For classification accuracy, most conditions favored ridge logistic regression over logistic regression and random forest. Differences tended to decrease as the number of items and prevalence increased. Also, linear models predicting classification accuracy differences suggested that there were differences in classification accuracy between relaxed Lasso and logistic regression (classification rate $R^2 = .120$; sensitivity $R^2 = .393$; and specificity $R^2 =$ .111) and between relaxed Lasso and random forest (classification rate $R^2 = .120$; sensitivity $R^2 = .393$; and specificity $R^2 = .111$). Conditions with 30 items favored relaxed Lasso, and conditions with 10 items favored logistic regression, although these differences decreased as prevalence and sample size increased. On the other hand, the relaxed Lasso had higher classification accuracy than random forest in most conditions. It is important to note that random forest has higher classification rate, higher specificity, and lower sensitivity in conditions with 10 binary items than the relaxed Lasso. It is likely that random forest was assigning most cases to the nondiagnosed class in conditions with limited information (i.e., conditions with the fewest items and item categories).

Table 3
Variance Explained and Partial-η^2 *Effect Sizes for the Predictors of Classification Accuracy for Machine Learning Models Using ROC Classifiers*

	RF ROC			Logistic reg ROC			Ridge ROC			Lasso ROC			Relaxed Lasso ROC		
Factor	cr	se	sp	cr	se	sp	cr	se	sp	cr	se	sp	cr	se	sp
Overall R^2	.506	.456	.433	.506	.408	.404	.240	.327	.212	.096	.140	.081	.112	.139	.094
Nitem	.000	.085	.000	.003	**.014**	.002	**.090**	**.042**	**.062**	**.028**	**.013**	**.020**	**.031**	**.012**	**.022**
Ncat	.000	.093	.001	.009	.003	.008	**.045**	**.006**	**.034**	**.016**	**.012**	**.011**	**.024**	**.015**	**.016**
Cor	.424	**.219**	**.334**	**.492**	**.362**	**.387**	-	-	-	-	-	-	-	-	-
Ss	.010	.010	.010	**.015**	**.030**	.009	**.000**	**.000**	**.000**	**.035**	**.028**	**.023**	**.030**	**.034**	**.018**
Prev	.016	.012	.023	.006	.005	.008	**.133**	**.302**	**.133**	**.019**	**.090**	**.027**	**.030**	**.081**	**.037**
nitem:ncat	.063	.092	.061	.000	.004	.000	.010	.000	.008	.001	.001	.001	.003	.000	.002
nitem:cor	.008	.001	.004	.003	.001	.002	-	-	-	-	-	-	-	-	-
nitem:ss	.009	.015	.009	.000	.008	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
nitem:prev	.029	.048	.027	.000	.006	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
ncat:cor	.007	.001	.005	.000	.001	.000	-	-	-	-	-	-	-	-	-
ncat:ss	.008	.019	.008	.000	.003	.001	.000	.000	.000	.000	.000	.000	.000	.001	.000
ncat:prev	.016	.053	.015	.001	.002	.000	.000	.000	.000	.000	.000	.000	.000	.001	.000
cor:ss	.001	.000	.001	.001	.004	.001	-	-	-	-	-	-	-	-	-
cor:prev	.001	.004	.002	.007	.001	.007	-	-	-	-	-	-	-	-	-
ss:prev	.004	.005	.002	.001	.008	.001	.000	.000	.000	.000	.001	.000	.001	.000	.001
nitem:ncat:cor	**.011**	.001	.009	.001	.000	.001	-	-	-	-	-	-	-	-	-
nitem:ncat:ss	**.013**	**.018**	**.013**	.001	.002	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000
nitem:ncat:prev	**.043**	**.053**	**.034**	.000	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
nitem:cor:ss	.001	.000	.000	.000	.004	.000	-	-	-	-	-	-	-	-	-
nitem:cor:prev	.008	.001	.007	.001	.005	.001	-	-	-	-	-	-	-	-	-
nitem:ss:prev	.004	.005	.003	.002	.006	.002	.000	.000	.000	.000	.000	.000	.000	.000	.000
ncat:cor:ss	.001	.000	.001	.000	.000	.000	-	-	-	-	-	-	-	-	-
ncat:cor:prev	.007	.001	.006	.000	.000	.000	-	-	-	-	-	-	-	-	-
ncat:ss:prev	.004	.007	.003	.000	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
cor:ss:prev	.000	.000	.000	.001	.005	.001	-	-	-	-	-	-	-	-	-

Note. Neither local dependence or four/five-way interactions had partial-$\eta^2 > .010$ model. Colon (:) is for interactions. In bold are the highest-order terms in which each simulation factor is involved with partial-$\eta^2 > .010$. Cells with dashes are for predictors not included in the model. reg = regression; ROC = receiver operating characteristic; RF = random forest; cr = classification rate; se = sensitivity; sp = specificity; rf = random forest; reg = regression; nitem = number of items; ncat = number of item categories; cor = diagnosis-test correlation; ss = sample size; prev = prevalence.

Overall, results suggest that ridge logistic regression and the relaxed Lasso work best when there is a high correlation between the item responses and the diagnosis and when there are more predictor items to regularize.

Comparing logistic regression to random forest. Linear models predicting classification accuracy differences suggested that there were differences in classification accuracy between logistic regression and random forest (classification rate R^2 = .184; sensitivity R^2 = .351; and specificity R^2 = .170). Similar to the comparison with the relaxed Lasso, random forest had higher classification rate and specificity and lower sensitivity than logistic regression in conditions with 10 binary items and low prevalence. Largely, in the rest of the conditions, logistic regression had higher classification rate and specificity than random forest when there was a diagnosis-test correlation of .30, and random forest had higher classification rate than logistic regression when the diagnosis test-correlation was .70. On the other hand, logistic regression and random forest had similar sensitivity, except in the case of 10 binary items and low prevalence, discussed above.

Comparing Classification Accuracy Across Psychometric and Machine Learning Models

So far, the simulation results suggest that: (a) there are no significant differences between classification of data-generating θ_{Assmt} and EAP[$\hat{\theta}$]; (b) machine learning algorithms with a Bayes classifier were not likely to assign cases to the least prevalent class; and (c) ridge logistic regression, Lasso, and relaxed Lasso with a ROC classifier assign respondents to the diagnosis class in conditions in which there is a diagnosis-test correlation of .70. Therefore, this section only compares the classification accuracy between EAP[$\hat{\theta}$] and logistic regression with a ROC classifier, and between EAP[$\hat{\theta}$] and random forest with a ROC classifier.

Comparing EAP[$\hat{\theta}$] to logistic regression. EAP[$\hat{\theta}$] had higher classification accuracy than logistic regression across the vast majority of conditions (see Figure 3). Differences in classification rate between logistic regression and EAP[$\hat{\theta}$] ranged between −.10 to .02; differences in sensitivity ranged from −.28 to .03; and differences in specificity ranged from −.10 to .03. Although there were differences in classification rate (R^2 = .040) and specificity (R^2 = .028) across simulation factors, none of the predictors had a partial-$\eta^2 > .01$. Sensitivity (R^2 = .064) estimates were similar between logistic regression and EAP[$\hat{\theta}$], except in conditions with 30 items, a small sample size, and .05 prevalence, where EAP[$\hat{\theta}$] had higher sensitivity than logistic regression. It is likely that the combination of a small sample size, small prevalence, and many items could lead to estimation problems in logistic regression.

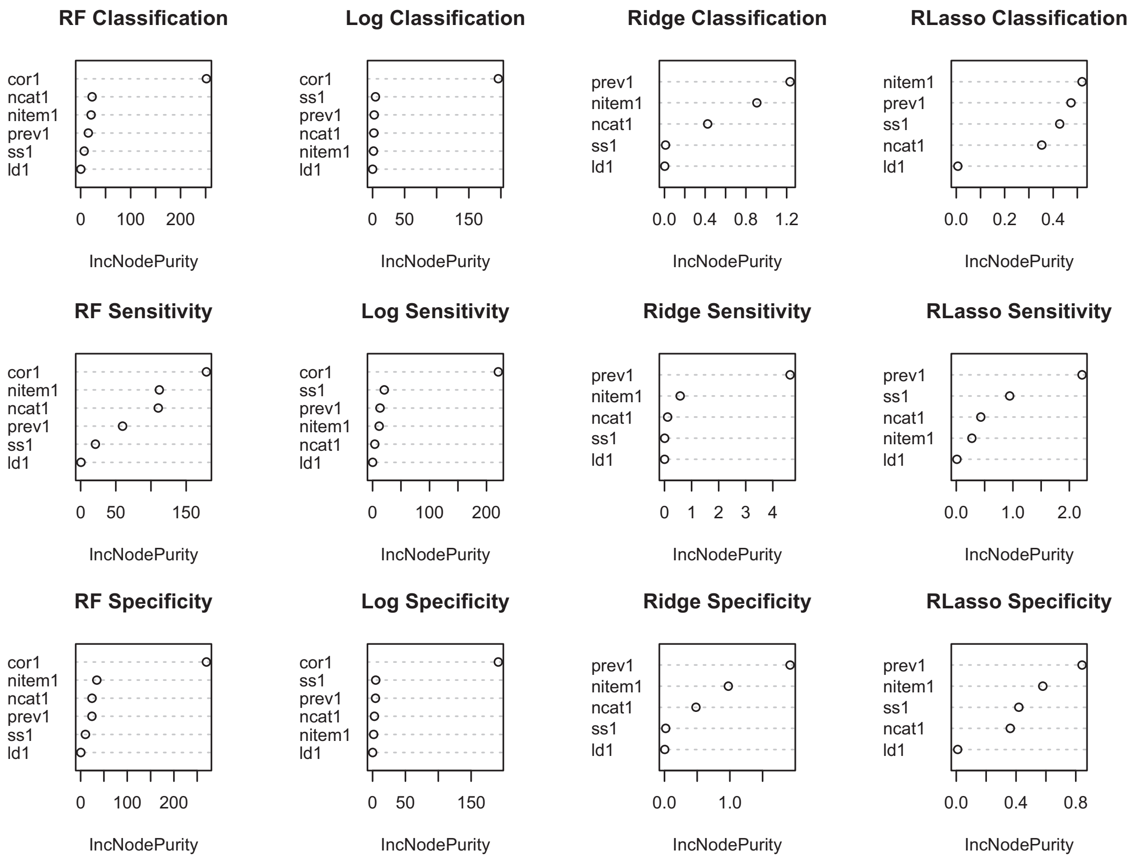

Figure 2. Random forest variable importance measures for machine learning algorithms with ROC classifiers. RF = random forest; Log = logistic regression; Ridge = ridge logistic regression; RLasso = relaxed Lasso logistic regression; nitem1 = number of items; ncat1 = number of item categories; cor1 = diagnosis-test correlation; ss1 = sample size; prev1 = prevalence. Note that variable importance measures for ridge logistic regression and relaxed Lasso logistic regression are only for conditions with diagnosis-test correlation of .70. Also, note that all of the plots have different x-axes.

Comparing EAP[$\hat{\theta}$] to random forest. EAP[$\hat{\theta}$] had a higher classification accuracy than random forest across the vast majority of conditions (see Figure 4). Differences in classification rate between the random forest model and EAP[$\hat{\theta}$] ranged between −.12 to .20; differences in sensitivity ranged from −.42 to .05; and differences in specificity ranged from −.13 to .24. Similar to previous findings, random forest appears to be assigning most respondents to the most prevalent category in conditions with 10 binary items and a prevalence of .05 or .10. As a result, random forest had higher classification rate and specificity, and lower sensitivity than the data generating model in those conditions. On average, differences in classification rate (R^2 = .167) and specificity (R^2 = .151) decreased as the diagnosis-test correlation, number of items, prevalence, and number of item categories increased. Also, differences in sensitivity (R^2 = .256) decreased as prevalence, sample size, number of items, and number of item categories increased.

Empirical Illustration

The methods from this article are illustrated using data from the RIGHT Track longitudinal study, an ongoing study which investigates how health, social, and emotional development in children predict future outcomes (Wideman et al., 2016). There were N = 447 children across three cohorts who were originally recruited for RIGHT Track and who have been followed since they were age 2. For this application, the subset of the data analyzed was collected when the children were age 7. Children were evaluated for ODD by trained interviewers using the Diagnostic Interview Schedule for Children (DISC), and then children received a diagnosis. Also, one of the child's parents completed the eight-item ODD subscale in the Disruptive Behavior Disorders Rating Scale (DBDRS; Pelham et al., 1992). Researchers could use psychometric or machine learning approaches to investigate if the parent-reported ODD DBDRS subscale discriminates children who meet or do not meet the ODD diagnostic criteria by the DISC.

In this case, children who were given diagnostic interviews and whose parents completed the ODD DBDRS subscale were included in the analysis. Therefore, the sample size for these analyses was N = 278, with a 12.6% of respondents meeting the ODD diagnostic criteria by the DISC. The eight items in the ODD subscale reflect different symptoms of ODD (anger, arguing, etc.) rated across four response categories (*not at all, just a little, pretty*

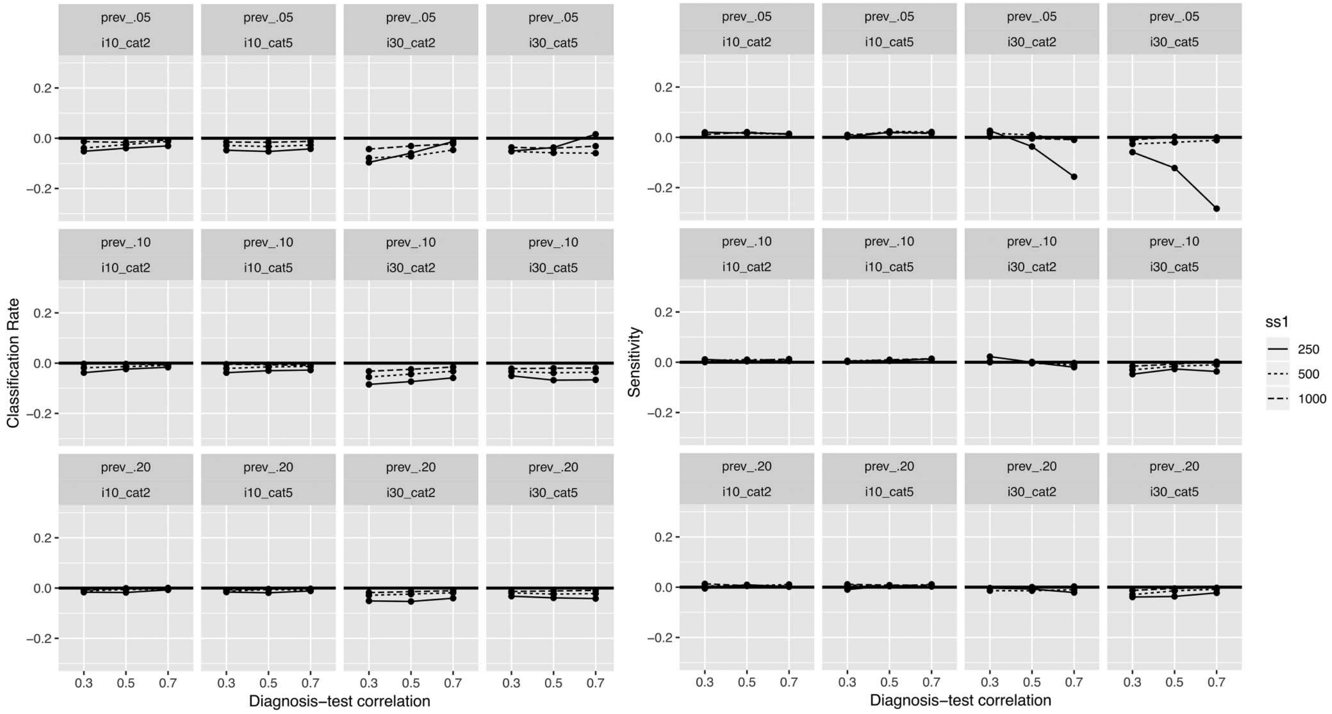

Figure 3. Difference in classification accuracy between logistic regression (above zero line) and EAP[$\hat{\theta}$] (below zero line) across sample size, diagnosis-test correlation, prevalence, number of items, and number of item categories.

much, and *very much*). Higher categories indicated higher symptom severity. In this example, the psychometric approaches illustrated are based on summed scores and EAP[$\hat{\theta}$] scores. The machine learning approaches illustrated predicted the probability of the ODD diagnosis from the ODD DBDRS item responses using random forest, logistic regression, ridge logistic regression, Lasso logistic regression, and relaxed Lasso logistic regression.

Parallel analysis suggests that the ODD subscale is largely unidimensional (first eigenvalue was 4.553, second eigenvalue was .854). Descriptive statistics suggest that less than 5% of the parents endorsed the last item category, so the third and the fourth categories were collapsed to prevent estimation problems in the IRT model. The full dataset was split into a training (N = 139) and a testing dataset (N = 139). The estimation steps and tuning of the metaparameters for both psychometric and machine learning methods largely follow those outlined in the Method section for the simulation and described in the Appendix. For the psychometric approaches, the training dataset was used to determine the cut score for both the summed score and the EAP[$\hat{\theta}$] score using ROC curves, along with estimating item parameters from the graded response model. For the machine learning approaches, the training dataset was used to develop the models, obtain predicted probabilities, and determine the probability threshold (via ROC curves) to assign a respondent to the diagnosis class. The testing dataset was used to evaluate the sensitivity, specificity, and classification rate of the models developed where the ODD DBDRS subscale predicts the ODD diagnosis by the DISC.

Figure 5 shows the ROC curves for the seven methods evaluated. Table 4 contains the area under the ROC curve (AUC), which is an indicator of classification accuracy by the models developed in the training dataset. The lowest AUC value across all of the models is .858 (classifying respondents based on EAP[$\hat{\theta}$] scores), and the highest was .970 (classifying respondents based on a random forest model). Thus, the AUC values suggested that the models developed in the training dataset are likely to classify participants well in the testing dataset.

Table 4 also contains sensitivity, specificity, and classification rates for the testing dataset for each of the methods illustrated. In this case, high sensitivity would suggest that the ODD DBDRS subscale is able to correctly identify children who meet the ODD diagnostic criteria by the DISC, and high specificity would suggest that the ODD DBDRS subscale is able to correctly rule out children who do not meet the ODD diagnostic criteria by the DISC. Results suggest that classification based on the predicted probabilities of logistic regression and Lasso logistic regression maximized both sensitivity and specificity of the ODD DBDRS subscale in the testing dataset. Random forest, whose AUC value suggested that the ODD diagnosis was classified almost perfectly in the training dataset, had lower sensitivity and specificity in the testing dataset than both logistic regression and Lasso logistic regression. On the other hand, the psychometric approaches using summed scores and EAP[$\hat{\theta}$] had lower sensitivity and higher specificity than some of the machine learning approaches.

Discussion

Diagnostic assessments are important because they are cheaper and less invasive than a gold standard that dictates diagnoses. Diagnostic assessments allow assessment specialists to decide if they should follow-up with a respondent or not, or determine if the

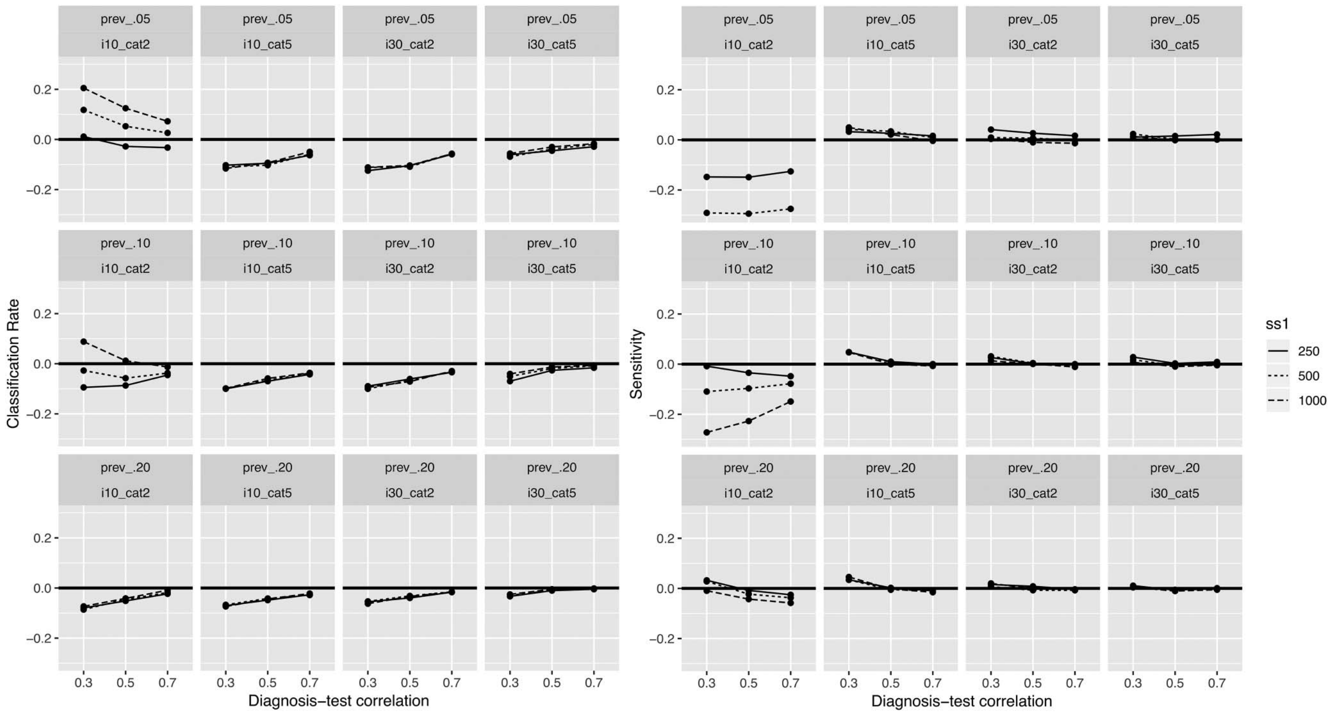

Figure 4. Difference in classification accuracy between random forest (above zero line) and EAP[$\hat{\theta}$] (below zero line) across sample size, diagnosis-test correlation, prevalence, number of items, and number of item categories.

assessment supports a clinician's diagnostic evaluation or not. After an assessment has been administered, it is important to decide if the respondent should be classified as diagnosed or not.

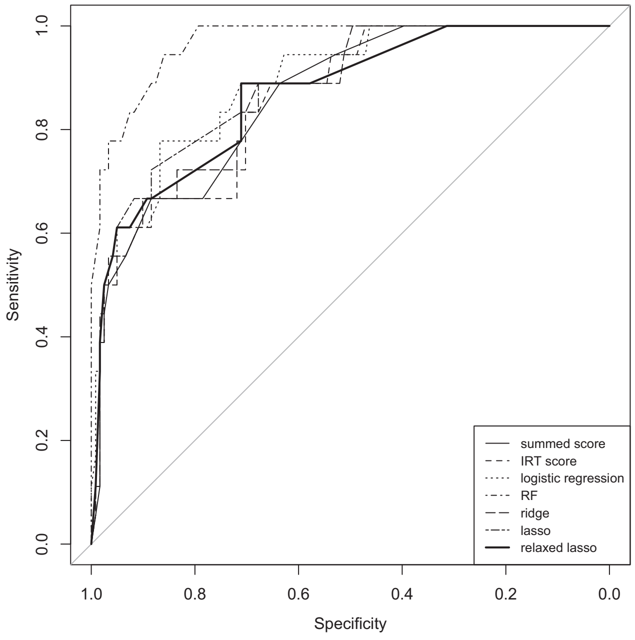

Figure 5. ROC curves for the psychometric and machine learning methods used to evaluate if the ODD DBDRS subscale can discriminate respondents who meet the ODD diagnostic criteria by the DISC.

This article investigated psychometric and machine learning approaches to make classification decisions. Psychometric methods determine the diagnosis by aggregating item responses, and then classify respondents if the assessment score is above or below a diagnosis cut score. Summed scores or IRT scores could be used to aggregate item responses. On the other hand, machine learning methods first predict the probability of diagnosis using the item responses, and then classify respondents according to the class that is most probable (Bayes classifier), or the class corresponding to their location above or below a probability threshold (ROC classifier). The machine learning methods discussed were classification trees, random forest, logistic regression, ridge logistic regression, Lasso logistic regression, and relaxed Lasso logistic regression.

There are several important findings from the simulation study. First, the most important factors that affected classification accuracy across psychometric and machine learning models are the diagnosis-test correlation and prevalence of the diagnosis. In other words, classification accuracy improves when assessment items are highly related to the diagnosis. It is presumed that higher prevalence is associated with lower classification accuracy because there is a greater chance that the model might confound diagnosed and nondiagnosed respondents. Second, machine learning models that focused on reducing prediction error by using a Bayes classifier mostly assigned respondents to the most prevalent class, leading to high specificity and low sensitivity across conditions. This finding is not surprising given that there is class imbalance in the simulation datasets (i.e., the 5% prevalence condition translates to only one out of 20 respondents having the diagnosis). Perhaps the main point of this finding is that machine learning methods should not be used blindly because the algo-

Table 4

Classification Performance of the Psychometric and Machine Learning Methods Used to Evaluate if the ODD DBDRS Subscale can Discriminate Respondents Who Meet the ODD Diagnosis by the DISC

Model	Training AUC	Cut score/ threshold	Testing sensitivity	Testing specificity	Testing classification rate
Summed score	.861	7.500	.765	.893	.878
EAP[$\hat{\theta}$] score	.858	.684	.705	.893	.871
Logistic regression	.890	.153	.823	.869	.863
Random forest	.970	.071	.765	.795	.791
Ridge logistic reg	.867	.166	.647	.926	.892
Lasso	.877	.145	.823	.869	.863
Relaxed Lasso	.868	.078	.941	.631	.669

Note. $N = 139$, out of which 17 respondents met the ODD diagnosis by the DISC. Note that thresholds are on different metrics. ODD = oppositional defiant disorder; DBDRS = Disruptive Behaviors Disorder Rating Scale; DISC = Diagnostic Interview Schedule for Children; AUC = area under the ROC curve; EAP[$\hat{\theta}$] = expected a posteriori IRT score; Reg = regression.

rithms might misclassify respondents in the least prevalent class when the sole focus is reducing prediction error. This finding is not novel, but it bears repeating. Finally, there were conditions in which logistic regression with a ROC classifier and random forest with a ROC classifier had comparable classification accuracy to the psychometric models, suggesting that a machine learning approach could be a viable alternative to a psychometric approach to diagnostic assessment.

Specifically, logistic regression with a ROC classifier had similar sensitivity and slightly lower classification rate (and specificity) than EAP[$\hat{\theta}$] across conditions (see Figure 3). On the other hand, the estimates of classification accuracy of random forest with a ROC classifier and EAP[$\hat{\theta}$] were comparable as sample size, diagnosis-test correlation, prevalence, number of items, and number of item categories increased (see Figure 4). In other words, classification accuracy of the random forest improved when the algorithm was given more information, as in more respondents, items, and item categories (thus, more candidate splits). Also, ridge logistic regression, Lasso logistic regression, and relaxed Lasso logistic regression with ROC classifiers were likely to assign respondents to the least prevalent class only in conditions with a diagnosis-test correlation of .70. Perhaps the Lasso also needed a strong relation between the items and the diagnosis to be able to select an appropriate number of items, and in turn predict the diagnosis accurately. Generally, the implication of these findings is that logistic regression or random forest could be used to make classification decisions when a psychometric approach is not feasible. These findings are not limited just to the area of diagnostic assessment, but could extend to any setting in which classification decisions are based on a unidimensional assessment. Some of these settings include situations where aptitude is assessed for job placement, achievement is assessed to determine grade placement, or simply to any test that assesses a construct and the process reduces to a binary decision. However, researchers should consider the application and assessment consequences before moving from a psychometric to a machine learning approach, further discussed below.

Psychometrics and Machine Learning

One of the overarching themes of this article is how psychometrics and machine learning could be used to address the same problem from different angles. Many of the attractive features that machine learning brings into psychology and medicine are not particularly new for the field of psychometrics. Both machine learning and psychometrics have traditionally addressed classification problems. Whereas machine learning applications have, for example, been used for identifying customers that display a certain behavior or predicting the likelihood of certain events, psychometrics has focused on finding the best candidates for a job, students who are college-ready, or physicians qualified to be licensed. The learning process of training the model in one part of the data and evaluating the model in a testing dataset from machine learning also is largely similar to the item calibration and scoring steps in large-scale testing, or a routine exploratory-confirmatory factor analysis. Likewise, the exploratory nature of finding patterns in machine learning methods could be similar to early stages of assessment development, where exploratory factor analysis is carried out to understand the structure of the construct that was measured. Closely related, both areas have focused on dimension reduction. Whereas machine learning has preferred atheoretical principal components, psychometrics has favored factor analysis due to its interpretability and replicability (Widaman, 2018). Finally, some machine learning methods have approaches to select the most important variables with a specific property, which could be similar to items selected in automated test assembly or computerized adaptive testing in psychometrics. Perhaps the point of this juxtaposition is to highlight that psychometrics and machine learning might be more similar than they are different, and that each of the areas could learn from each other.

In this instance, despite the similarity in classification accuracy across models, it is important to examine the process that psychometrics and machine learning took to make a decision about the diagnosis. The psychometric approach focused on finding the best way to assess the construct by aggregating item responses into a score, and then deciding on the diagnosis based on a cut score. A model to derive the score is likely to provide a precise estimate of the score and an interpretation of how the score is related to the diagnosis. The machine learning approach bypassed the estimation of a score and examined how item responses predicted the diagnosis. In essence, the main difference between the two approaches could be described in terms of needing to aggregate and interpret item responses. Depending on the application and consequences of assessment, the interpretation of the scores might or might not be the primary focus. On the other hand, machine learning models were more likely to predict the diagnosis when the diagnosis-test correlation was high. If the diagnosis-test correlation is not high, then aggregating item responses might make it easier to detect the relation between item responses and the diagnosis. Finally, a hybrid approach using machine learning and psychometrics could be considered. From a pool of calibrated items, one could use a machine learning algorithm to select the items that predict the diagnosis (primary product of the analysis), and then one could score the item responses using an IRT model (secondary product of the analysis), such that the score is reliable and could be ascribed valid interpretation.

Limitations and Future Directions

There were several limitations to the simulation that could be addressed in future studies. One of the goals of the article was to study how violating assumptions of the psychometric model affected classification accuracy. The violations of local independence simulated were shown to slightly affect item parameter recovery, but they did not predict classification accuracy in any of the models. A future direction would be to include a stronger manipulation of local dependence or other IRT model misspecifications, such as unmodeled multidimensionality or model error at the population level (MacCallum, Widaman, Preacher, & Hong, 2001). In the same vein, another goal of the article was to examine if classification accuracy of machine learning approaches would be favored over psychometric approaches in conditions with a small sample size. The vast majority of IRT models converged and item parameters were accurately recovered, so the estimation of the EAP[$\hat{\theta}$] was not adversely affected by small sample sizes. However, there were several small sample conditions in which machine learning models were likely to assign cases only to the most prevalent class. As mentioned in the Method section, in order to train machine learning models in conditions with a small sample size, five-fold cross-validation and stratified sampling were used to guarantee that each fold had both diagnosed and nondiagnosed respondents. The remaining conditions used 10-fold cross-validation. There is a possibility that differences in model training could have affected the comparability of results across conditions. A future direction would be to examine approaches that overcome class imbalance problems in machine learning, such as oversampling the minority class, undersampling the majority class, or creating synthetic cases of the least prevalent class using nearest-neighbor algorithms (SMOTE; Kuhn & Johnson, 2013). Other approaches to overcome class imbalance would be to introduce weights to the classes or to use the prevalence (if known) as a probability threshold for classification, similar to the Bayes classifier. Overall, it would be important to continue to study the small sample properties of some of these methods, which would be useful in settings where diagnostic assessments are developed for populations who are difficult to access.

As mentioned before, diagnostic assessment is usually carried out in a psychometric framework (Gibbons et al., 2013), and item responses were simulated to largely match what is found in commonly used assessments (Hill, 2004). The simulation setup of this article allows us to examine if the machine learning approaches could approximate the classification accuracy of the psychometric methods that are typically used. A future direction would be to simulate diagnostic assessment data that favors machine learning approaches (Lu & Petkova, 2014). For example, the main predictors of the diagnosis could have been some, but not all, of the assessment items, the items could have had a nonlinear relation with the diagnosis, or there could have been complex interactions among the items. Also, the classification accuracy between EAP[$\hat{\theta}$] and the summed scores was similar. A future direction would be to simulate items that yield a lower summed score reliability and examine how classification accuracy is affected. Finally, alternative models could have been used to predict the diagnosis. Boosting, a machine learning algorithm that uses binary recursive partitioning to fit trees to residual structures, or support vector machines, a machine learning algorithm that attempts to find separating planes for classification, could have been used to predict the diagnosis from item responses (Hastie et al., 2009). Other machine learning methods with varying levels of complexity and tuning parameters could have been used in this study, but given enough data, most machine learning algorithms might be expected to perform pretty similarly (Domingos, 2012). It is important to highlight that the machine learning methods used in this study were chosen because applied researchers are likely to have encountered these methods in the literature (e.g., logistic regression or decision trees), which in turn helps to introduce direct extensions that are unfamiliar to them (e.g., regularized logistic regression and random forest).

Conclusion

The results of this study suggest that machine learning algorithms could be used for diagnostic assessment and classification tests without sacrificing much classification accuracy, and that they could provide a viable alternative to the psychometric models. However, it is important to consider the implications for assessment when one takes either a psychometric or a machine learning approach to diagnostic assessment. The psychometric approach focuses on improving the measurement of the construct in the assessment by introducing a latent variable. There is an indirect prediction of the diagnosis because the diagnosis is not part of the measurement process. On the other hand, machine learning builds a model to predict the probability of diagnosis. Therefore, there is a direct prediction of the diagnosis. Keeping the goals of diagnostic assessment in mind could help researchers decide what approach to take, and the results of this simulation could suggest under what circumstances each model could perform well. Finally, it is important to reflect on the overlap between psychometrics and machine learning, and understanding advantages and disadvantages from each of the models could help each of these fields learn from the other.

References

Achenbach, T., & Rescorla, L. (2013). Achenbach system of empirically based assessment. In F. Volkman (Ed.), *Encyclopedia of autism spectrum disorders* (pp. 31–39). New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4419-1698-3_219

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–472). Reading, MA: Addison Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459. http://dx.doi.org/10.1007/BF02293801

Breiman, L. (2001). Random forests. *Machine Learning, 45,* 5–32. http://dx.doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth International Group.

Chalmers, R. P. (2012). A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48,* 1–29. http://dx.doi.org/10.18637/jss.v048.i06

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55,* 78–87. http://dx.doi.org/10.1145/2347736.2347755

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16,* 5–18. http://dx.doi.org/10.1007/s11136-007-9198-0

Edwards, M. C., Houts, C. R., & Cai, L. (2018). A diagnostic procedure to detect departures from local independence in item response theory models. *Psychological Methods, 23,* 138–149. http://dx.doi.org/10.1037/met0000121

Egan, J. P. (1975). *Signal detection theory and ROC analysis*. New York, NY: Academic Press.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33,* 1–22. http://dx.doi.org/10.18637/jss.v033.i01

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57,* 423–436. http://dx.doi.org/10.1007/BF02295430

Gibbons, R. D., Hooker, G., Finkelman, M. D., Weiss, D. J., Pilkonis, P. A., Frank, E., . . . Kupfer, D. J. (2013). The CAD-MDD: A computerized adaptive diagnostic screening tool for depression. *The Journal of Clinical Psychiatry, 74,* 669–674. http://dx.doi.org/10.4088/JCP.12m08338

Gibbons, R. D., Weiss, D. J., Frank, E., & Kupfer, D. (2016). Computerized adaptive diagnosis and testing of mental health disorders. *Annual Review of Clinical Psychology, 12,* 83–104. http://dx.doi.org/10.1146/annurev-clinpsy-021815-093634

Gonzalez, O., O'Rourke, H. P., Wurpts, I. C., & Grimm, K. J. (2018). Analyzing Monte Carlo simulation studies with classification and regression trees. *Structural Equation Modeling, 25,* 403–413. http://dx.doi.org/10.1080/10705511.2017.1369353

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. New York, NY: Springer. http://dx.doi.org/10.1007/978-0-387-84858-7

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The LASSO and generalizations*. Boca Raton, FL: CRC Press. http://dx.doi.org/10.1201/b18401

Hill, C. D. (2004). *Precision of parameter estimates for the graded item response model*. Unpublished manuscript, The University of North Carolina at Chapel Hill, Chapel Hill, NC.

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling, 23,* 555–566. http://dx.doi.org/10.1080/10705511.2016.1154793

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4614-7138-7

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349,* 255–260. http://dx.doi.org/10.1126/science.aaa8415

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4614-6849-3

Lewinsohn, P. M., Seeley, J. R., Roberts, R. E., & Allen, N. B. (1997). Center for Epidemiologic Studies Depression Scale (CES-D) as a screening instrument for depression among community-residing older adults. *Psychology and Aging, 12,* 277–287. http://dx.doi.org/10.1037/0882-7974.12.2.277

Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. *R News, 2,* 18–22.

Liu, X. (2012). Classification accuracy and cut point selection. *Statistics in Medicine, 31,* 2676–2686. http://dx.doi.org/10.1002/sim.4509

Loeffelman, J. E., Steinley, D., Boness, C. L., Trull, T. J., Wood, P. K., Brusco, M. J., & Sher, K. J. (2020). Combinatorial optimization of clustering decisions: An approach to refine psychiatric diagnoses. *Multivariate Behavioral Research*. Advance online publication. http://dx.doi.org/10.1080/00273171.2020.1717921

Lu, F., & Petkova, E. (2014). A comparative study of variable selection methods in the context of developing psychiatric screening instruments. *Statistics in Medicine, 33,* 401–421. http://dx.doi.org/10.1002/sim.5937

MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research, 36,* 611–637. http://dx.doi.org/10.1207/S15327906MBR3604_06

McArdle, J. (2013). Adaptive testing of the number series test using standard approaches and a new decision tree analysis approach. In J. McArdle & G. Ritschard (Eds.), *Contemporary issues in exploratory data mining in the behavioral sciences* (pp. 312–344). New York, NY: Routledge. http://dx.doi.org/10.4324/9780203403020

McFall, R. M., & Treat, T. A. (1999). Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology, 50,* 215–241. http://dx.doi.org/10.1146/annurev.psych.50.1.215

Pelham, W. E., Jr., Gnagy, E. M., Greenslade, K. E., & Milich, R. (1992). Teacher ratings of *DSM–III–R* symptoms for the disruptive behavior disorders. *Journal of the American Academy of Child & Adolescent Psychiatry, 31,* 210–218. http://dx.doi.org/10.1097/00004583-199203000-00006

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford, UK: Oxford University Press.

Reckase, M. (2009). *Multidimensional item response theory*. New York, NY: Springer. http://dx.doi.org/10.1007/978-0-387-89976-3

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., . . . Cella, D. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care, 45,* S22–S31. http://dx.doi.org/10.1097/01.mlr.0000250483.85507.04

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27,* 133–144. http://dx.doi.org/10.1111/j.1745-3984.1990.tb00738.x

Ripley, B. (2016). tree: Classification and regression trees (R package version 1.0–37) [Computer software]. Retrieved from https://CRAN.R-project.org/package=tree

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12,* 77. http://dx.doi.org/10.1186/1471-2105-12-77

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. Psychometrika Monograph No. 17 (4, Pt. 2). http://dx.doi.org/10.1007/BF03372160

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics, 9,* 307. http://dx.doi.org/10.1186/1471-2105-9-307

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods, 14,* 323–348. http://dx.doi.org/10.1037/a0016973

Thissen, D., & Steinberg, L. (2009). Item response theory. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The Sage handbook of quantitative methods in psychology* (pp. 148–177). London, UK: Sage Publications. http://dx.doi.org/10.4135/9780857020994.n7

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring*. Mahwah, NJ: Erlbaum Publishers. http://dx.doi.org/10.4324/9781410604729

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B. Methodological, 58,* 267–288. http://dx.doi.org/10.1111/j.2517-6161.1996.tb02080.x

van der Linden, W. J., & Pashley, P. J. (2010). Item selection and ability estimation adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 3–30). New York, NY: Springer. http://dx.doi.org/10.1007/978-0-387-85461-8

Widaman, K. F. (2018). On common factor and principal component representations of data: Implications for theory and for confirmatory replications. *Structural Equation Modeling, 25,* 829–847. http://dx.doi.org/10.1080/10705511.2018.1478730

Wideman, L., Calkins, S. D., Janssen, J. A., Lovelady, C. A., Dollar, J. M., Keane, S. P., . . . Shanahan, L. (2016). Rationale, design and methods for the RIGHT Track Health Study: Pathways from childhood self-regulation to cardiovascular risk in adolescence. *BMC Public Health, 16,* 459. http://dx.doi.org/10.1186/s12889-016-3133-7

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques*. Cambridge, MA: Morgan Kaufmann.

Yan, D., Lewis, C., & Stocking, M. (2004). Adaptive testing with regression trees in the presence of multidimensionality. *Journal of Educational and Behavioral Statistics, 29,* 293–316. http://dx.doi.org/10.3102/10769986029003293

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12,* 1100–1122. http://dx.doi.org/10.1177/1745691617693393

Youngstrom, E. A. (2014). A primer on receiver operating characteristic analysis and diagnostic efficiency statistics for pediatric psychology: We are ready to ROC. *Journal of Pediatric Psychology, 39,* 204–221. http://dx.doi.org/10.1093/jpepsy/jst062

Youngstrom, E. A., Frazier, T. W., Demeter, C., Calabrese, J. R., & Findling, R. L. (2008). Developing a 10-item mania scale from the Parent General Behavior Inventory for children and adolescents. *The Journal of Clinical Psychiatry, 69,* 831–839. http://dx.doi.org/10.4088/JCP.v69n0517

Zhou, X. H., McClish, D. K., & Obuchowski, N. A. (2011). *Statistical methods in diagnostic medicine*. New York, NY: Wiley. http://dx.doi.org/10.1002/9780470906514

Zou, K. H., O'Malley, A. J., & Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation, 115,* 654–657. http://dx.doi.org/10.1161/CIRCULATIONAHA.105.594929

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry, 39,* 561–577. http://dx.doi.org/10.1093/clinchem/39.4.561

(*Appendix follows*)

Appendix

Information on Parameter Tuning

Below is a more detailed description about the specification and estimation of the models described in the Method section, tuning metaparameters, and how the training and testing datasets were used to estimate each model.

1. **Data-generating θ**. In the training dataset, use the data-generating θ in a ROC analysis (using the pROC R package; Robin et al., 2011) to choose a cut score to determine the diagnosis.

   a. Determine the predicted diagnosis per respondent in the testing dataset using the cut score from the data-generating θ from the training dataset.

2. **Raw summed score.** In the training dataset, estimate a summed score by summing all item responses. Use the summed score in a ROC analysis to choose a cut score to determine the diagnosis.

   a. Estimate the summed score for respondents in the testing dataset.

   b. Determine the predicted diagnosis per respondent using the summed cut score from the training dataset.

3. **Estimated EAP[$\hat{\theta}$].** In the training dataset, fit a unidimensional IRT model to calibrate the items (using MML-EM in the mirt R package; Chalmers, 2012) and estimate the EAP[$\hat{\theta}$] score. Use the EAP[$\hat{\theta}$] in a ROC analysis to choose a cut score to determine the diagnosis.

   a. Estimate EAP[$\hat{\theta}$] scores for respondents in the testing dataset using the item parameters from the training dataset.

   b. Determine the predicted diagnosis per respondent using the EAP[$\hat{\theta}$] cut score from the training dataset.

4. **Logistic regression.** In the training dataset, predict the probability of diagnosis from item responses using logistic regression (Equation 5, using the glm R function). There are no tuning parameters for the logistic regression model used in this study. Use the predicted probabilities in a ROC analysis to choose a probability threshold to determine the diagnosis.

   a. Predict the probability of diagnosis in the testing dataset using the logistic regression model from the training dataset.

   b. Determine a predicted diagnosis per respondent if the respondent has greater than 50% probability of having the diagnosis (Bayes classifier).

   c. Determine a predicted diagnosis per respondent using the probability threshold from the training dataset (ROC classifier).

5. **Ridge logistic regression.** In the training dataset, predict the probability of diagnosis from item responses using regularized logistic regression with a ridge penalty (Equation 7, using the glmnet R package; Friedman, Hastie, & Tibshirani, 2010). The tuning parameter for ridge logistic regression is the weight given to the ridge penalty in the ridge logistic regression log-likelihood function. Use *k*-fold cross-validation to determine the penalty tuning parameter that minimizes prediction error in the left-out fold. The value of the tuning parameters tried were the default grid values of the glmnet package. Choose a penalty parameter *one-standard-error* away from the one that minimized the prediction error to regularize the parameters (James et al., 2013). Note that the metaparameter was tuned for each simulated dataset. Use the predicted probabilities in a ROC analysis to choose a probability threshold to determine the diagnosis.

   a. Predict the probability of diagnosis in the testing dataset using the ridge logistic regression model from the training dataset.

   b. Determine a predicted diagnosis per respondent if the respondent has greater than 50% probability of having the diagnosis (Bayes classifier).

   c. Determine a predicted diagnosis per respondent using the probability threshold from the training dataset (ROC classifier).

(*Appendix continues*)

6. **Lasso logistic regression.** In the training dataset, predict the probability of diagnosis from item responses using regularized logistic regression with a lasso penalty (Equation 8, using the glmnet R package; Friedman et al., 2010). The tuning parameter for lasso logistic regression is the weight given to the lasso penalty in the lasso logistic regression log-likelihood function. Use *k*-fold cross-validation to determine the penalty parameter that minimizes prediction error in the left-out fold. The value of the tuning parameters tried were the default grid values of the glmnet package Choose a penalty parameter *one-standard-error* away from the one that minimized the prediction error to regularize the parameters (James et al., 2013). Note that the metaparameter was tuned for each simulated dataset. Use the predicted probabilities in a ROC analysis to choose a probability threshold to determine the diagnosis.

   a. Predict the probability of diagnosis in the testing dataset using the lasso logistic regression model from the training dataset.

   b. Determine a predicted diagnosis per respondent if the respondent has greater than 50% probability of having the diagnosis (Bayes classifier).

   c. Determine a predicted diagnosis per respondent using the probability threshold from the training dataset (ROC classifier).

7. **Relaxed Lasso logistic regression**. In the training dataset, predict the probability of diagnosis from item responses using regularized logistic regression with a Lasso penalty (similar to Model 6, with tuning parameters described above). Save the predictors with the nonzero regression coefficients remaining in the model. Predict the probability of diagnosis from the remaining predictors using an unstructured logistic regression model (as in Model 4). Use the predicted probabilities in a ROC analysis to choose a probability threshold to determine the diagnosis.

   a. Predict the probability of diagnosis in the testing dataset using the relaxed Lasso logistic regression model from the training dataset.

   b. Determine a predicted diagnosis per respondent if the respondent has greater than 50% probability of having the diagnosis (Bayes classifier).

   c. Determine a predicted diagnosis per respondent using the probability threshold from the training dataset (ROC classifier).

8. **Classification and regression trees**. In the training dataset, predict the probability of diagnosis from the item responses using a classification tree (using the tree R package; Ripley, 2016). There are several tuning parameters for trees, among which is the size of the tree. Overgrow the tree in a training dataset (deviance of .0001) and then prune back using cost-complexity pruning. Determine the size of the tree by *k*-fold cross-validation to find the tree size that minimizes misclassification in the left-out fold. Note that the metaparameter was tuned for each simulated dataset.

   a. Predict the probability of diagnosis in a testing dataset using the pruned tree from the training dataset.

   b. Determine a predicted diagnosis per respondent if the respondent has greater than a 50% probability of having the diagnosis (Bayes classifier).

   c. Given the limited variability in the predicted probabilities using CART, the ROC classifier to determine predicted diagnoses was not carried out.

9. **Random forest.** In the training dataset, predict the probability of diagnosis from the item responses using an ensemble deep classification trees grown in bootstrapped datasets (using the randomForest R package; Liaw & Wiener, 2002). There are tuning parameters for random forest, among which are the number of trees grown and the number of candidate predictors to split on. However, these tuning parameters were fixed to growing 500 trees in bootstrapped datasets and selecting $\sqrt{p}$ candidate *p* predictors to split on. Note that $\sqrt{p}$ candidate predictors was used as a heuristic (James et al., 2013) and preliminary analyses suggested that less than 500 trees are needed to stabilize prediction error, but we decided to use 500 because of the low burden in computation time. Use the predicted probabilities in a ROC analysis to choose a probability threshold to determine the diagnosis.

   a. Predict the probability of diagnosis in a testing dataset using the random forest model from the training dataset.

   b. Determine a predicted diagnosis per respondent if the respondent has greater than a 50% probability of having the diagnosis (Bayes classifier).

   c. Determine a predicted diagnosis per respondent using the probability threshold from the training dataset (ROC classifier).

Received June 23, 2019
Revision received May 7, 2020
Accepted May 11, 2020 ■