American Psychological Association

Psychological Assessment

ISSN: 1040-3590

2020, Vol. 32, No. 5, 473–492
http://dx.doi.org/10.1037/pas0000808

Generalizability of Statistical Prediction From Psychological Assessment Data: An Investigation With the MMPI-2-RF

William H. Menton
Kent State University

In the present study, the author employed tools and principles from the domain of machine learning to investigate four questions related to the generalizability of statistical prediction in psychological assessment. First, to what extent do predictive methods common to psychology research and machine learning actually tend to predict new data points in new settings? Second, of what practical value is parsimony in applied prediction? Third, what is the most effective way to select model predictors when attempting to maximize generalizability? Fourth, how well do the methods considered compare with one another with respect to prediction generalizability? To address these questions, the author developed various types of predictive models on the basis of Minnesota Multiphasic Personality Inventory (MMPI)-2-RF scales, using multiple prediction criteria, in a calibration inpatient sample, then externally validated those models by applying them to one or two clinical samples from other settings. Model generalizability was then evaluated based on prediction accuracy in the external validation samples. Noteworthy findings from the present study include (a) statistical models generally demonstrated observable performance shrinkage across settings regardless of modeling approach, though they nevertheless tended to retain non-negligible predictive power in new settings; (b) of the modeling approaches considered, regularized (penalized) regression methods appeared to produce the most consistently robust predictions across settings; (c) parsimony appeared more likely to reduce than to enhance model generalizability; and (d) multivariate models whose predictors were selected automatically tended to perform relatively well, often producing substantially more generalizable predictions than models whose predictors were selected based on theory.

Public Significance Statement
This study evaluated how well prediction models developed using a variety of approaches accurately generate predictions in new clinical samples, a question rarely considered in psychological assessment research. Among other findings, results suggest that researchers may be able to improve their predictive accuracy across samples by using a more data-driven approach to model construction.

Keywords: generalizability, machine learning, MMPI-2-RF, statistical prediction, validation

Supplemental materials: http://dx.doi.org/10.1037/pas0000808.supp

Although psychological assessment data can be gathered from many sources (e.g., interview, psychological test, behavioral observation), the evaluator eventually must combine these data to generate inferences about the individual being assessed. Psychologists involved in assessment tend to base their decision-making (e.g., diagnosis, recommendations) upon informal integration of the assessment data; however, a substantial body of evidence suggests that the decision process is most effective when driven by application of formal statistical prediction models or other mechanical procedures (Ægisdóttir et al., 2006; Grove, Zald, Lebow, Snitz, & Nelson, 2000; Meehl, 1954). On its face, psychological assessment appears a ripe opportunity for applied statistical prediction; many of the data gathered during a psychological assessment (psychological test results, e.g.) are quantitative in nature, or are at least quantifiable. Such data lend themselves particularly well to integration in predictive statistical models, which provide an empirical basis for generating predictions and guiding other decisions and recommendations.

However, the *generalizability* of a statistical prediction model limits its practical utility. That is, such a model is useful only insofar as it allows us to meaningfully predict *new* data, not just the data used to develop the model (Busemeyer & Wang, 2000). The generalizability problem is of great practical importance. Despite the apparent superiority of statistical prediction over human judgment, traditional statistical prediction models, such as linear and logistic regression, have met only limited success when validated

Editor's Note. Cecil Reynolds, Guest Editor, served as the sole action editor for this submission.—YSB-P

This article was published Online First February 6, 2020.

Correspondence concerning this article should be addressed to William H. Menton, Department of Psychological Sciences, Kent State University, Kent Hall, 600 Hilltop Drive, Kent, OH 44240. E-mail: wmenton@kent.edu

on external data, particularly when the prediction models include multiple predictor variables (Browne, 2000; Mosier, 1951). Although human behavior is inherently complex and multifactorial, and is therefore difficult to predict, the problem of limited prediction generalizability also owes substantially to limitations of the statistical methods employed (Babyak, 2004; Copas, 1983; Darlington, 1968).

The Case for Statistical Prediction

Meehl (1954) argued that when an appropriate statistical prediction model is available, clinical psychologists should base their decision-making on the output of the model, rather than on their own expert judgment. To support this argument, Meehl (1954) presented a survey of approximately 20 studies in which clinical prediction and statistical prediction were directly compared; in each case, statistical prediction performed as well as, or better than, expert clinical prediction. This finding was counterintuitive to many clinicians, and a lively debate on the relative merits of clinical and statistical prediction ensued. Nevertheless, subsequent empirical investigations, including two large-scale meta-analyses (Ægisdóttir et al., 2006; Grove et al., 2000), have vindicated Meehl's (1954) position. These studies found that across applied scientific fields (not just within the domain of clinical psychology), experts are rarely able to outperform statistical models when making predictions, even when the experts have access to more information. Rather, in most cases, statistical prediction performs as well as, or better than expert judgment. No situations in which a clinician might be expected to systematically outperform a statistical prediction model were identified.[1] Human cognitive limitations appear to be at fault for the underperformance of expert judgment relative to statistical prediction. For example, human beings are prone to cognitive biases, and as such, they have a difficult time objectively weighting various predictor variables when attempting to arrive at an optimal probabilistic judgment.

These findings, though influential, have not led to the proliferation of applied statistical prediction within the field of psychology that Meehl had originally envisioned (Grove & Meehl, 1996; see also Meehl, 1956, 1986). Nevertheless, the demonstrable general superiority of a statistically oriented approach to prediction and decision-making suggests that development of useful statistical prediction models is a worthwhile pursuit for assessment-oriented psychology researchers. In fact, given that the process of psychological assessment potentially provides a great deal of information that can be used to drive predictions and decision-making with significant real-life consequences, development of decision-guiding statistical models is arguably a central task for the science of psychological assessment (Meehl, 1956). Moreover, because statistical prediction models offer a general advantage over expert human judgment, development and successful application of such models is particularly consonant with the goals and ethics of clinical psychologists and other applied psychologists who seek to use psychological assessment to improve the well-being of others, as improved prediction could lead to more optimal decision-making and, therefore, to improved outcomes.

Similarly, for both practical and ethical reasons, psychologists are increasingly expected to engage in evidence-based practice (EBP) while operating clinically, using sound empirical research to guide intervention (Anderson, 2006). Accordingly, psychologists conducting assessments in such settings should engage in evidence-based assessment (EBA), using empirically supported methods to optimally meet the needs of evaluees, service professionals, third-party payers and referral sources, and other affected parties (Anderson, 2006; Hunsley & Mash, 2007). Thus, if formal, applied statistical models offer a potential advantage over the subjective judgment-based prediction methods (e.g., impressionistic combination of assessment data) that currently predominate many areas of applied psychology, then a focus of assessment psychology research should indeed be on the development of statistical procedures that can be applied in practice to aid real-world decision-making. Moreover, in areas in which some limited form of statistical prediction is already in place (e.g., in using single-scale cut scores to determine the applicability of empirically validated interpretive statements), the potential for improvement of prediction through more advanced statistical models merits investigation.

Generalizability of Traditional Statistical Prediction Methods

Statistical prediction in psychological assessment research most commonly takes the form of linear or logistic regression, likely owing at least in part to the ease with which such methods can be applied and interpreted. For a number of methodological reasons, the models produced by these techniques are potentially limited with respect to their generalizability. For example, psychology researchers generally favor simple predictive models, perhaps including only a single predictive variable, yet criteria of interest in psychology are generally complex and may share nonredundant explanatory variance with any number of additional potential predictors. Limiting the information used to generate predictions restricts the amount of predictable information, which in turn limits the validity of predictions generated by simple models. Thus, more complex predictive models appear to offer a significant potential advantage over simple predictive models because inclusion of additional predictor variables hypothetically permits the model to account for a greater amount of criterion variance.

However, this gives rise to a problem often underappreciated in the psychological sciences. Namely, as the number of predictors in a model increases, conventional techniques for multiple regression tend to produce increasingly unstable predictor weights of limited generalizability (Babyak, 2004; Browne, 2000; Copas, 1983; Darlington, 1968; Mosier, 1951). This is at least partly attributable to the fact that these models are typically optimized using ordinary least squares (OLS) or maximum likelihood estimation (MLE), which algorithmically optimize the regression model by minimizing squared errors of prediction across the range of data used to develop the model.

Model fit with respect to the calibration data used to develop the model will nevertheless tend to improve as the model becomes

[1] Meehl (1957) did argue that clinicians have an advantage in identifying "broken leg cases": cases in which special and unusual circumstances (such as breaking one's leg, which could lead to a deviation from normally predictable behavior) justify ignoring or overruling a statistical prediction model. However, he also argued that clear-cut broken leg cases occur very seldomly in applied psychology (perhaps much more seldomly than clinicians believe), and as such, clinicians are better off relying upon statistical prediction the vast majority of the time.

more complex. However, when a complex model is used to predict new data points for cases not included in the calibration data, model performance tends to be much poorer than one would expect, given its good fit to the calibration data. In such a case, the model is said to be *overfit* to the calibration data, meaning that it is unlikely to perform comparably well in any new dataset. The attenuation of goodness-of-fit indices observed when using an overfit model to predict new data has been termed *shrinkage* (see Copas, 1983). Generalizability of statistical prediction from psychological assessment data is further potentially limited by discrepancies between the criterion distributions in the model development sample and external applied samples, such as base rate shifts (see Hunsley & Meyer, 2003; Meehl & Rosen, 1955; Wiggins, 1973) or range restriction (Linn, 1968), yet assessment psychologists rarely formally account for these phenomena in their predictions (Linn, Harnisch, & Dunbar, 1981; Meehl & Rosen, 1955).

A Way Forward: Machine Learning

The problems just described are not limited to the assessment area of psychology. Indeed, some have argued that failure to produce generalizable predictive models is a significant shortcoming for psychology research broadly and that research psychologists in general should focus on producing predictive models that generalize across samples, rather than on evaluating explanatory models based on their fit to calibration data (see Yarkoni & Westfall, 2017).[2] This raises the question of whether traditional statistical methods in the study of psychology can be improved upon or replaced to advance generalizable applied prediction. As Yarkoni and Westfall (2017) have suggested, one promising avenue for such improvement is through adopting the methods of a successful field in which generalizable statistical prediction is a central focus: machine learning.

Machine learning, also sometimes called statistical learning, is often associated with computer science and complex computer learning problems such as image recognition and artificial intelligence. However, virtually any statistical prediction method in modern use potentially falls within the domain of machine learning; indeed, linear and logistic regression are among the most commonly used machine learning techniques (James, Witten, Hastie, & Tibshirani, 2014). Machine learning also includes many statistical methods not commonly employed by psychologists but potentially applicable to prediction problems in psychology. Furthermore, machine learning's approach to statistics notably differs from the approaches traditionally used by psychology researchers in that it places a premium on the generalizability of findings. Indeed, in machine learning, the gold standard for model performance *is* the model's goodness of fit to validation data (i.e., data not used to develop the model; Efron & Hastie, 2016; Hastie, Tibshirani, & Friedman, 2009; James et al., 2014; Yarkoni & Westfall, 2017).

Machine Learning: A Brief Introduction

In this section, I provide a conceptual overview of key machine learning concepts relevant to applied prediction in psychology.[3]

Machine Learning Terminology

Some of the terms described below are also used in assessment psychology; however, common terms may differ in definition and usage between the fields of machine learning and psychological assessment. Additionally, these two areas share several concepts that are given different labels depending on the field. After these terms are introduced, the author will therefore adhere to the (generally more precise) machine learning language conventions for the remainder of this paper.

Training and test data. Simply put, *training data* are any data used to calibrate a statistical model, whereas *test data* are any data used to evaluate a statistical model. In the machine learning approach, no overlap should exist, under most circumstances, between training and test data, though both training and test data may represent subsets of a larger sample of data.

Validation and cross-validation. In assessment psychology, the term *validation* has many meanings. For example, although validation can refer to evaluation of a statistical model, as is relevant in the present discussion, it can also refer to establishment of various forms of theoretical or psychometric validity (e.g., construct, content, criterion). In machine learning, *validation* refers specifically to evaluation of a statistical model using test data. Similarly, in psychology research, *cross-validation* generally refers to any evaluation of a model or measure outside the training data (often using an entirely different sample). However, in machine learning, *cross-validation* usually refers specifically to a family of resampling procedures used in model selection and evaluation (Kohavi, 1995), described below (see the Method section). To minimize ambiguity in the remainder of the paper, in addition to using the machine learning definitions for validation and cross-validation, the author will use the term *external validation* to describe the specific case in which a model is validated on a test dataset entirely distinct from the original data (i.e., not merely a holdout set of the original sample).

Tuning variables. A *tuning variable*, *tuning parameter*, or *hyperparameter* is a variable used in the model training process to calibrate the model; however, tuning variables are not generally used to generate predictions directly. The values of tuning variables are often prespecified by the researcher or are selected by comparing cross-validated performances of competing models along various values of the tuning variable.

Considerations in Prediction Method Selection and Model Specification

In many cases, the predictive modeling approach that a researcher selects represents a compromise between competing goals of research and application (e.g., model interpretability vs. generalizable predictive accuracy), and an informed researcher must ensure that the model selected is appropriate to the data.

[2] Yarkoni and Westfall (2017) further argue that theoretical models in psychology should be operationalized in a predictive framework so that their generalizability can be directly tested, as many models commonly presented in psychology research do not readily lend themselves to clear testing.

[3] See the Method section for a technical discussion of specific techniques used in the present study.

Regression versus classification. In machine learning, the regression-classification distinction refers to the nature of the criterion variable involved in the prediction task. *Regression* in this sense refers to prediction of a continuous or quasi-continuous criterion, whereas *classification* refers to prediction of discrete group membership (Hastie et al., 2009; James et al., 2014; Kuhn & Johnson, 2013).

Flexibility, interpretability, and the bias-variance trade-off. *Flexibility* is the extent to which a predictive model can account for variations in distributions of data. *Interpretability* is the extent to which an individual examining the trained model can understand how the model produces predictions as a function of its predictors (see Hastie et al., 2009). Flexibility and interpretability are not inherently incompatible; however, highly flexible modeling approaches (e.g., polynomial regression with many higher-order terms) tend to be more difficult to interpret than less flexible approaches (James et al., 2014). When accuracy of prediction of external data is the primary research goal, interpretability may not be a major consideration in selecting a modeling approach; however, researchers interested in forming explanatory theoretical inferences about the predictors may wish to select a more interpretable model over one that has good predictive power but limited interpretability.[4] In addition, the standards of a given domain of applied prediction may weigh against use of an uninterpretable model.

Understanding model bias, model variance, and the trade-off between them is a fundamental requirement for the knowledgeable development and evaluation of generalizable predictive models. Virtually all prediction models applied to test data produce some errors of prediction. Error attributable to *model bias* is the amount that the model errs, on average, relative to its expected (i.e., predicted) values. Model bias error tends to result from reducing a complex process to a simplified statistical approximation. Error attributable to *model variance*, on the other hand, is the error in prediction that occurs due to the influence of variability in the training data (i.e., sampling error) on the model's calibrated values. Unfortunately, bias and variance are often at odds with one another. A model with low bias must be flexible enough to meaningfully approximate the form of the true underlying function; however, as model flexibility increases, so does the risk of overfitting the data. The key to producing a generalizable model is therefore to strike an optimal balance in the *bias-variance trade-off*, which requires an informed approach to model evaluation and selection (Efron & Hastie, 2016; Friedman, 1997; Hastie et al., 2009; James et al., 2014).

Complexity versus parsimony. A complex model is one that includes many predictors, or *features*; in contrast, a parsimonious model includes relatively few. A parsimonious model with few predictors tends toward high model bias but low model variance. As complexity increases (i.e., as more features are added to the model), variance tends to increase while bias tends to decrease. However, addition of a feature only decreases overall prediction error if the additional feature accounts for a sufficient amount of additional variance in the criterion to offset the additional error also included in the new feature (Miller, 2002).

Parsimonious models tend to be more interpretable than more complex ones. For these reasons, statisticians tend to prefer parsimonious models over more complex alternatives, even when model performance criteria suggest they have equivalent predictive power (see, e.g., Hastie et al., 2009; James et al., 2014; Miller, 2002). However, the most appropriate level of complexity for a model is not always clear; in fact, identifying the optimal number of features to include in a model is a common task in statistical learning (see the Feature Selection section below). Prediction methods that tend to produce effectively uninterpretable models are often referred to as *black box* methods.

Psychology researchers tend to place a premium on model parsimony in predictive modeling, strongly preferring simple to more complex models. This is especially the case when the more complex models under consideration include predictors that cannot be justified a priori by a strong theoretical association between the predictor and criterion. However, the practical advantages and disadvantages of parsimony in applied prediction in psychology are not well-understood, as models of varying complexity are rarely empirically compared with one another, and even when they are compared empirically, such models are rarely evaluated fairly with regard to their generalizable predictive power.

Model Evaluation and Selection

The distinction between training and test data is key to understanding the evaluation of a predictive model's generalizability. As discussed above, in psychological research, training data performance is often misleadingly put forth as the expected future performance of a predictive model (Babyak, 2004; Yarkoni & Westfall, 2017). However, because a predictive model's parameters are optimized to maximize the model's fit to the training data, such a model will virtually always fit its training data better than it will any test data, including data produced in the real-world settings in which professionals may wish to apply the model. Fairly evaluating the potential generalizability of a candidate predictive model therefore requires some criterion of model performance that accounts for expected discrepancies between training and test data. In machine learning, the gold standard of model performance is external validation, ideally using multiple external samples, though resampling techniques (e.g., *k*-fold cross-validation) and penalized training data performance criteria (e.g., adjusted R^2) are sometimes used when external data are not available (Hastie et al., 2009; James et al., 2014; Kuhn & Johnson, 2013).

Feature Selection

As noted above, feature selection is the process of selecting, from the pool of all possible candidate predictors, those that will be included in the final predictive model. This becomes particularly important when the number of potential predictors is very large; in fact, many conventional modeling techniques (e.g., OLS regression) cannot estimate model parameters if the number of predictors exceeds the number of observations. Thus, in such situations, feature selection (or regularization, as discussed below) is not only advisable, it is necessary. Although psychology researchers generally advocate for theory-driven feature selection, many mechanical alternatives are available and are commonly used in machine learning.

[4] As an analogy, consider orthogonal versus oblique factor rotation; the former is simple to interpret, but the latter tends to fit the data better, at the cost of interpretability.

Although psychology researchers are generally familiar with basic stepwise feature selection methods, these and other feature selection techniques have fallen out of favor owing to concerns regarding their limited generalizability. Fears regarding the generalizability of stepwise methods are not entirely unwarranted; indeed, empirical research has shown that such methods tend to overfit their training data and to include spurious features in their final model (Derksen & Keselman, 1992; Steyerberg, 2009). Notably, machine learning's toolbox includes a robust suite of techniques potentially useful for both automated feature selection and evaluation of automatically selected models, and many of those techniques include elements intended to mitigate concerns such as overfitting.

The Present Study

Thus far, I have endeavored to demonstrate three primary points. First, assessment psychologists have compelling reasons to develop and apply statistical prediction models to aid decision-making in real-world contexts. Second, for a variety of methodological reasons, traditional approaches to statistical prediction in psychology research tend to produce models of limited or questionable generalizability, and, moreover, the generalizable predictive power of these models is rarely evaluated properly. Third and finally, the field of machine learning, a branch of applied statistics, employs an array of principles and methods that could be incorporated into psychological assessment research to improve both statistical prediction and model evaluation with respect to generalizability.

In the present study, the author aims to evaluate the extent to which statistical prediction and model evaluation in psychological assessment may be improved by application of machine learning concepts and techniques. The author compares traditional prediction methods currently in wide use in psychology research, such as ordinary least squares (OLS) linear and logistic regression, to alternative modern prediction techniques commonly used in the field of machine learning. In accordance with best-practice machine learning principles, models are evaluated with an emphasis on test data performance, rather than on training data performance. In all analyses, test score data obtained from a broadband measure of personality and psychopathology, the MMPI-2 Restructured Form (MMPI-2-RF; Ben-Porath & Tellegen, 2008), serve as predictors, whereas various clinician-rated clinical constructs of interest (mental status exam variables, psychiatric diagnoses) serve as prediction criteria. Results of the present study will inform responses to four research questions related to the central issue of the generalizability of statistical predictions in psychological assessment. First, to what extent do the methods under consideration tend to produce generalizable predictions? Second, to what extent does a model benefit from having more or less predictive information (i.e., how should we value parsimony in applied prediction)? Third, what is the most effective way to approach feature selection when attempting to maximize generalizable predictive power in psychological assessment? Fourth and finally, how well do the methods under consideration compare with one another with respect to prediction generalizability? Responses to these questions, informed by the results of this empirical investigation, can provide important guidance for psychologists who wish to develop useful, generalizable predictive models.

Psychology and machine learning are vast fields, and testing more than a small subset of their respective techniques would be unfeasible. For the purposes of the present study, traditional techniques are drawn from the pool of statistical prediction methods judged by the author to be most commonly used in psychological assessment research (e.g., conventional OLS linear regression). Machine learning techniques in the present study are selected based on their apparent rates of usage and acceptance in machine learning and other areas of applied statistics, as well as their demonstrated off-the-shelf success.[5] This selection is guided both by leading texts in the area (Hastie et al., 2009; James et al., 2014; Kuhn & Johnson, 2013) and by past empirical comparative studies outside the field of psychology (particularly Caruana & Niculescu-Mizil, 2006; and Fernández-Delgado, Cernadas, Barro, & Amorim, 2014). Specific selected techniques are described in detail in the Data Analysis Strategy section below.

Method

Participants

The present study employs archival data originally collected at three separate clinical sites.[6] Two of these samples represent psychiatric inpatient populations: one from a large metropolitan county hospital (N = 1,524) in the American Midwest, the other from a VA hospital (N = 1,401) in the same geographic region. The third sample was drawn from a Midwestern outpatient community mental health center (N = 1,020). Participants producing invalid MMPI-2-RF protocols according to standard MMPI-2-RF criteria (any of: CNS $\geq$18, VRIN-$r \geq$ 80, TRIN-$r \geq$ 80, F-$r \geq$ 120, Fp-$r \geq$ 100; Ben-Porath & Tellegen, 2008) were excluded from analysis. Table 1 displays exclusion rates and descriptive demographic data for the participants retained from these three samples.

Measures

MMPI-2 Restructured Form (MMPI-2-RF). All retained participants produced valid protocols for the MMPI-2-RF, a 338-item, 51-scale broadband measure of personality and psychopa-

[5] That is, the machine learning techniques used in the present study are selected based on their ability to perform competitively in terms of predictive performance (relative to other candidate techniques) without placing a great burden of technical knowledge or excessive preparatory work upon the researcher developing the model. The requirement that these modeling techniques have a history of performing well off-the-shelf is stipulated for two reasons. First, this property mitigates the probability that apparent model performance in the present study reflects a misuse of the modeling technique, to the extent that the simplicity of these methods' applications precludes user error. Second, demonstration of generalizable prediction using a new technique that is easy to apply increases the likelihood that the technique will be adopted by others (relative, say, to a demonstration using a model producing similar performance but requiring greater technical knowledge on the part of the researcher).

[6] Data were collected in the 1990s under the approval of prior institutional review boards (IRBs), and no new data or identifiable private information were collected for the present study. As such, the present study was exempt from further IRB review.

Table 1
Sample Exclusion Rates and Demographic Characteristics

Characteristic	Psychiatric inpatients, county	Psychiatric inpatients, VA	Community outpatients
Percent excluded[a]	31%	31%	12%
Final *N*	1,050	972	895
Gender	57.0% male	93.3% male	61.2% female
Mean age	34.0	48.1	33.0
Mean years of education	9.3	9.7	12.5

Note. VA = Veterans Affairs.
[a] Percent excluded because of MMPI-2-RF protocol invalidity.

thology.[7] In addition to appraising threats to protocol validity (e.g., inconsistent responding, overreporting, underreporting), the MMPI-2-RF's scales measure substantive content spanning multiple domains of psychopathology, including internalizing dysfunction, externalizing dysfunction, thought dysfunction, interpersonal dysfunction, and somatization. The MMPI-2-RF has been the subject of numerous studies evaluating its psychometric properties (see Ben-Porath, 2012a, 2012b; Tellegen & Ben-Porath, 2008). Its scales have been found to be sufficiently reliable and to demonstrate good convergent and discriminant validity with respect to empirical associations with extratest criteria (Tellegen & Ben-Porath, 2008). Extensive MMPI-2-RF reliability and validity coefficients for the samples used in the present study are reported in Tellegen and Ben-Porath (2008). Internal consistency (Cronbach's alpha) values for the MMPI-2-RF substantive scales range from .60–.95 (M = .77) for the county hospital sample, .50–.92 (M = .75) for the VA hospital sample, and .57–.93 (M = .75) for the community outpatient sample.

Record review form scales. Research assistants completed identical record review forms (RRFs) for participants in the county hospital inpatient and VA inpatient samples (but not in the community outpatient sample). These RRFs included dichotomous mental status exam items indicating the presence or absence of specific patient symptoms or characteristics at the time of intake. In previous studies (see, e.g., the MMPI-2-RF Technical Manual; Tellegen & Ben-Porath, 2008), these mental status exam items were aggregated into scales measuring clinical constructs of interest. Four such RRF scales (Depression, Substance Abuse, Delusions, and Pain) are used in the present study as prediction criteria for regression analyses; any other previously created RRF scales are not used. Intraclass correlation (ICC) estimates of reliability for these scales range from .69 (Pain) to .91 (Delusions), with a mean ICC of .85 (Tellegen et al., 2003). Table 2 lists the dichotomous items comprising each of the RRF scales used in the present study. Items on these scales are scored 1 if the clinical feature is present or 0 if the feature is not present. Means and standard deviations for the RRF scales, separated by sample, are presented in Table 3.[8]

Clinical diagnoses. Clinical psychiatric diagnoses were assigned by all participants' treating clinicians according to the diagnostic criteria contained in the *Diagnostic and Statistical Manual of Mental Disorders*, third edition, revised (*DSM–III–R*; American Psychiatric Association [APA], 1987) or *Diagnostic and Statistical Manual of Mental Disorders*, fourth edition (*DSM–IV*; APA, 1994). Although operationalized in the *DSM* as discrete categorical diagnoses, a large body of research indicates that many psychological disorders are associated with one another through higher-order dimensional factors under which similar disorders are grouped (see Kotov et al., 2017). Thus, in the present study, four diagnostic groups are created to represent various major areas of psychopathology: Depressive Disorder, Substance Use Disorder, Thought Disorder, and Somatic Disorder. A participant is considered a member of one of these groups if the participant's chart notes a formal diagnosis of any of the disorders comprising the group. Membership statuses in these diagnostic groups are used as prediction criteria in the present study's classification method analyses. The specific *DSM* diagnoses defining each group are listed in Table 2. The base rates for each diagnostic group, separated by sample, are displayed in Table 3.

Data Analysis Strategy

General approach. The availability of predictive information in these models is systematically varied to explore the impact of different approaches to model parsimony and feature selection on model generalizability. Thus, in addition to being separated into regression and classification analysis branches, analyses are also organized into three groups representing different approaches to predictor selection. In the first group of analyses, the nine Restructured Clinical (RC) Scales are used as predictors. This scale set, arguably the most essential to interpretation of the MMPI-2-RF, covers a broad range of clinically important constructs.[9] In the second set of analyses, the full pool of all 42 substantive scales are used as predictors for each model. In the third and final analysis group, all 42 substantive scales are used to form a pool of *potential*

[7] Participants were originally administered the Minnesota Multiphasic Personality Inventory-2 (MMPI-2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989), a 567-item measure, rather than the MMPI-2-RF. All items from the MMPI-2-RF are found in the MMPI-2 item pool. As such, all MMPI-2-RF protocols used in the present study were rescored from MMPI-2 protocols. Tellegen and Ben-Porath (2008) present evidence of the equivalence of MMPI-2-RF scores generated from MMPI-2 and MMPI-2-RF administrations. In the two inpatient samples, but not in the outpatient sample, the psychiatrists who rendered diagnostic opinions had access to their patients' MMPI-2 (not MMPI-2-RF) results, though the extent to which this information was considered in making diagnoses is unknown.

[8] Visualizations of the distributions of the RRF scales in the county hospital and VA medical center samples are presented in the online supplemental materials as Figures S1 and S2, respectively.

[9] The RC Scales include RCd (Demoralization), RC1 (Somatic Complaints), RC2 (Low Positive Emotions), RC3 (Cynicism), RC4 (Antisocial Behavior), RC6 (Ideas of Persecution), RC7 (Dysfunctional Negative Emotions), RC8 (Aberrant Experiences), and RC9 (Hypomanic Activation).

Table 2
Prediction Criteria and Their Compositions

Regression criteria		Classification criteria	
RRF scale	Scale items	Diagnostic group	Included diagnoses
Depression	Worthlessness Guilt Helpless or hopeless Tearfulness Anhedonia	Depressive disorder	Major depressive disorder Dysthymia Depressive disorder NOS
Substance abuse	History of substance abuse Alcohol Polysubstance Hallucinogens Opioids Benzodiazepines Cocaine Marijuana	Substance use disorder	Substance abuse Substance dependence
Pain	Chronic pain Backache Headache Muscle ache	Somatic disorder	Hypochondriasis Somatization disorder Somatoform disorder Conversion disorder
Delusions	Delusions Ideas of reference Delusions of reference Persecutory delusions Paranoid/Suspicious	Thought disorder	Schizophrenia Schizoaffective disorder Schizophreniform disorder Psychotic disorder NOS

Note. RRF = record review form.

predictors; however, the predictors for the final models in this group are selected using various feature selection approaches. The specific modeling approaches used in each group of analyses are listed in Table 4. Brief descriptions of these methods are included in Table 5, along with rationales for their inclusion in the present study. More detailed descriptions of each modeling approach are included in the online supplemental materials.

Within each set of analyses, multiple modeling strategies are employed using the same starting set of predictors and the same prediction criteria. Use of multiple modeling strategies with the same training data and external validation data provides a basis for fair comparison of model performance for each model type. Multiple prediction criteria are employed for each modeling strategy to provide additional points of comparison. For all analyses, the county hospital inpatient sample is used for model training. For regression analyses, the VA hospital inpatient sample is used for external validation (with the RRF scales serving as prediction criteria), and for classification analyses, both the VA hospital sample and the community outpatient sample are used for external validation (with diagnostic groups as prediction criteria). Feature selection analyses employ both continuous and categorical prediction criteria (i.e., both the RRF scales and diagnostic groups), with external validation conducted on any applicable data. The use of training data from one clinical setting and test data from other clinical settings (as opposed to, say, generating training and test sets from data combined across settings) is intended to reflect a common reality in psychological research, as researchers often develop and present predictive models based on data from a single

Table 3
Distributional Characteristics of Prediction Criteria

Criterion	County hospital	VA medical center	Community outpatient
Record review form scales: *M (SD)*			
Depressed	1.90 (1.67)	0.93 (1.09)	
Substance abuse	2.15 (2.03)	1.85 (1.82)	
Delusions	0.71 (1.14)	0.43 (0.89)	
Pain	0.25 (0.70)	0.34 (0.77)	
Diagnostic groups: Base rates			
Depressive disorder	0.46	0.39	0.24
Substance use disorder	0.44	0.45	0.14
Thought disorder	0.17	0.14	0.02
Somatic disorder	0.01	0.02	<.01

Table 4
Modeling Techniques by Predictor Set

Predictors	Regression methods	Classification methods
RC Scales	OLS linear regression Negative binomial regression Ridge linear regression Random forests Support vector regression	MLE logistic regression Ridge logistic regression Random forests Support vector machines
All substantive scales	OLS linear regression Ridge linear regression Random forests Support vector regression	MLE logistic regression Ridge logistic regression Random forests Support vector machines
Feature selection	Single-predictor selection Manual selection Forward selection Backward deletion Forward-and-backward selection Lasso	Single-predictor selection Manual selection Forward selection Backward deletion Forward-and-backward selection Lasso

Note. RC = Restructured Clinical; OLS = ordinary least squares; MLE = maximum likelihood estimation.

setting with either an implicit or explicit assumption that the model will meaningfully generalize. Thus, this arrangement in the present study provides an opportunity to evaluate the ecological validity of applied prediction techniques.

Training and test data performance metrics are provided for all predictive models, including model performance on both the full training set and on external validation data. Additionally, whenever possible, *k*-fold cross-validation (Kohavi, 1995) performance is provided as well. *K*-fold cross-validation, a resampling procedure, is commonly used in machine learning as part of model development and evaluation. In *k*-fold cross-validation, the training data are randomly divided into *k* partitions (or *folds*) of approximately equal size. One data fold is designated the test set, and the remaining folds are combined to form the training set. The model is then trained using the training fold set, and its performance on the holdout test fold is recorded. Then, another fold is designated the test set, and the previous test fold is added back to the training data for the next round of model training. This process repeats until the model has been evaluated on each test fold, then predictive performance across all test folds is aggregated. In this way, *k*-fold cross-validation provides a relatively unbiased estimate of model performance, as test data are always excluded from model training. The present study employs 10-fold cross-validation (i.e., *k*-fold cross-validation using 10 folds) unless otherwise specified.

Machine learning routinely employs a number of model performance criteria, some of which can be used to evaluate models of many different forms. In the present study, the root mean squared error (RMSE), a measure of absolute model fit, is used to evaluate all regression models,[10] though R^2 values are also provided where appropriate. Classification models are evaluated using both overall classification accuracy, which evaluates performance of the model as a dichotomous classifier, and area under the curve (AUC), which here evaluates the discriminative performance of the model on the basis of its outputted probabilities.[11] All analyses are carried out in the R software environment (R Core Team, 2015).[12] R code for these analyses is presented as a supplementary file in the online supplemental materials.

Results

As a reminder, the data and validation procedures used in the present study are described using machine learning terminology. *Training data* refers to the dataset used to calibrate models: in this case, the county hospital inpatient mental health dataset. *Cross-validation* here refers specifically to *k*-fold cross-validation. *External validation* refers to validation on data drawn from settings other than that of the training data. Zero-order effect sizes between the predictors and prediction criteria (Pearson's *r* for regression criteria, AUC for classification criteria) are presented for all applicable samples in Supplemental Tables S1 and S2 in the online supplemental materials.

[10] RMSE is calculated by taking the root of the mean square error (MSE; the mean of all residuals for the model outcome variable). As a measure of error, RMSE is a badness-of-fit metric, with higher values indicating greater error and therefore poorer model performance. RMSE can be interpreted as the approximate average error of prediction for the model, in unstandardized units of the outcome variable. R^2 is not provided for all regression models because its use with nonlinear models is potentially inappropriate.

[11] Classification accuracy and AUC do not always lead to identical conclusions about comparative model performance. Generally speaking, classification accuracy is preferable when class balance is approximately equal (i.e., the base rate is around .5), and AUC is preferable when classes are highly imbalanced (i.e., the base rate is extremely high or extremely low). In addition, some modeling approaches (most relevantly including random forests, which is used in the present study) are known to perform well as classifiers and yet to produce poorly calibrated probabilities.

[12] Data organization and visualization is augmented through use of the tidyverse package (Wickham, 2017). Regularized regression is performed using the glmnet package (Friedman, Hastie, & Tibshirani, 2009). The MASS package (Ripley et al., 2013) is used to create negative binomial regression models and to construct stepwise-type models (through the stepAIC function). Random forests are generated using the randomForest package (Liaw & Wiener, 2002). Support vector regression and support vector machine models are courtesy of the kernlab package (Karatzoglou, Smola, Hornik, & Zeileis, 2004). Wherever possible, predictive models are passed through the caret package (Kuhn, 2008), which provides a generalized framework for training and testing predictive models. The pROC package (Robin et al., 2011) is used for some ROC analyses.

Table 5
Descriptions of Modeling Techniques

Method	Description	Rationale for inclusion
Data combination methods		
OLS linear regression[a]	A method that generates predictions using a linear combination of predictor weights calibrated to minimize the residual sum of squares (RSS) in the training data.	Commonly used in both psychology research and machine learning.
MLE logistic regression[b]	A generalization of OLS linear regression to dichotomous classification problems.	Commonly used in both psychology research and machine learning.
Negative binomial regression[a]	A count regression approach designed to model non-negative integers corresponding to the negative binomial distribution.	Recommended for wider use in assessment psychology research by some scholars (see Wright, Pincus, & Lenzenweger, 2012).
Ridge regression[a,b]	A regularized or penalized form of regression that produces more conservative (smaller) beta weight estimates than conventional OLS/MLE regression.	Regularization is thought to enhance generalizability and is common in machine learning.
Random forests[a,b]	An ensemble method that generates predictions by averaging them across many decision trees, each of which was trained on a bootstrapped sample of the training data.	Commonly used in machine learning.
Support vector regression/ machines[a,b]	A machine learning method that efficiently models nonlinear data and minimizes the influence of outlying data points in the training data.	Commonly used in machine learning.
Feature selection methods		
Single-variable selection[a,b]	Empirical selection of single predictors for bivariate models based on the strength of their zero-order associations to the prediction criteria in the training data. See Supplemental Table S3 for the single-variable models selected in the present study.	Intended to replicate common practice in assessment psychology of focusing on identification and interpretation of "best" single predictors.
Theory-driven (manual) selection[a,b]	Selection of predictors based on theory and past research, rather than purely empirical considerations. See Supplemental Table S4 for the predictors manually selected in the present study.	Generally recommended as "best practice" in psychology research.
Stepwise procedures[a,b]	A family of procedures (forward selection, backward deletion, and forward-and-backward stepwise selection) that incrementally add and/or subtract features until some performance criterion is optimized.	Automated feature selection methods that are well-known to, but generally maligned by, psychology researchers.
Lasso[a,b]	A regularization regression approach (very similar to ridge regression) that performs feature selection in addition to producing conservative predictor weights.	Commonly used in machine learning.

Note. OLS = ordinary least squares; MLE = maximum likelihood estimation.
[a] Used in regression analyses. [b] Used in classification analyses.

Regression Models

Regression models predicting record review form (RRF) criterion scales are presented in Supplemental Tables S5–S8 in the online supplemental materials, separated by prediction criterion. Two types of model performance statistics are presented for regression models: R^2 and root mean square error (RMSE), though as previously noted, the present study focuses on RMSE as the primary index of model fit. Detailed descriptions of the variable importance and model evaluation criteria presented in the presented study are provided in the online supplemental materials. Tables S5–S8 in the online supplemental materials also include model performance statistics for three sets of data: training data, *k*-fold cross-validation data, and external validation data. The external validation data performance values were calculated by applying the models calibrated in the training dataset to test data from the inpatient VA medical center sample. External validation performance statistics for all regression models are consolidated in Table 6.

A comparison of regression model performances in the training data with cross-validated and externally validated performances reveals a clear and expectable trend. That is, models generally fit training data the best, fit *k*-fold cross-validation data somewhat less well, and fit external validation data least well. However, this general trend does not hold true for all regression prediction criteria. Perhaps most notably, models predicting the Substance Abuse record review form scale appear to fit the external validation data approximately as well as, or sometimes even better than, the training data, as reflected in both R^2 and RMSE values. The cause of this unexpected finding is not obvious, though possible contributing factors include sampling variability, similar rates of substance abuse across settings, and unusually strong zero-order associations between predictors and the model criterion. The general magnitude of shrinkage in model performance between training and test data was substantial, in many cases corresponding to a decrement in R^2 from approximately .20 to approximately .10. In general, for models for which significant discrepancies between training and external validation data performance are observed, cross-validation performance appears to fall much closer to training data performance than to external validation performance. As a notable exception to this rule, the random forests and support

Table 6
External Validation (VA Medical Center) Performance of Regression Models

Model	R^2				RMSE			
	Depressed	Sub. abuse	Delusions	Pain	Depressed	Sub. abuse	Delusions	Pain
RC Scales								
Ordinary least squares	.10	.32	.13	.04	1.53	1.51	.87	.75
Negative binomial					1.52	1.58	.92	.76
Ridge	.10	.32	.13	.04	1.52	1.50	.87	.76
Random forests					1.54	1.55	.90	.76
Support vector regression					1.38	1.70	.92	.79
All substantive scales								
Ordinary least squares	.11	.38	.14	.06	1.54	1.46	.87	.75
Ridge	.12	.39	.15	.07	1.49	1.43	.85	.74
Random forests					1.53	1.45	.88	.75
Support vector regression					1.45	1.47	.84	.78
Feature selection								
Single predictor	.07	.37	.03	.06	1.56	1.45	.92	.75
Manual selection	.08	.39	.08	.07	1.59	1.43	.89	.74
Forward selection	.11	.38	.14	.06	1.49	1.45	.88	.75
Backward deletion	.11	.38	.14	.06	1.48	1.46	.87	.75
Stepwise selection	.11	.38	.14	.06	1.48	1.45	.87	.75
Lasso	.12	.40	.15	.07	1.49	1.42	.86	.74

Note. RC = Restructured Clinical; RMSE = root mean squared error; VA = Veterans Affairs.

vector regression models appear to overfit the training data to a much greater extent than any other modeling approach considered, though they nevertheless appear to produce external validation performance roughly comparable to other approaches. Models predicting the Pain RRF scale generally evidenced substantially poorer fit to the external validation data than models predicting other record review form scales. This likely reflects the overall low rate and variability of pain in both the training and external validation data.

A full discussion of each of the vast number of possible comparisons that can be made between the different regression models created as part of the present study is beyond the scope of this paper. Briefly, models predicting each of the regression criteria did not perform identically when applied to the external validation data, though models predicting the same criterion did appear to produce external validation performances in the same general range. More specifically, external validation RMSEs ranged from 1.38–1.59 for Depression, 1.42–1.70 for Substance Abuse, .86–.92 for Delusions, and .74–.79 for Pain.

The specific modeling approach producing the lowest external validation RMSE varied between regression criteria. Support Vector Regression (SVR) using the RC Scales as predictors produced the lowest RMSE (1.38) for the Depression criteria, relative to other models predicting the same criterion. However, curiously, SVR models produced external validation error rates among the *highest* of all competing methods when predicting Substance Abuse, Delusions, and Pain regression criteria (RMSE = 1.70, .92, and .79, respectively). The lasso produced the lowest RMSE among all models predicting Substance Abuse (1.42). The SVR model predicting Delusions using all Substantive Scales produced the lowest external validation RMSE (.84) among all competing models, though the other Delusions SVR model (i.e., the one using only the RC Scales as predictors) produced the highest external validation RMSE, as already noted. Several models were tied for the overall lowest external validation RMSE (.74) in predicting Pain (perhaps unsurprising given the narrow range of external validation performances for this particular criterion): ridge regression using all Substantive Scales, the manually selected model, and the lasso model.

To summarize the relative performances of each modeling approach both within and across regression criteria, a simple rank sum score was calculated for each approach. This score was computed by ranking each model within each regression criterion by external validation RMSE (lowest to highest, using the rounded values displayed in Table 6), then summing each model's ranks across all four regression criteria. Thus, the lower the score, the generally better the performance of the modeling approach's performance across regression criteria, relative to other modeling approaches, as measured by external validation RMSE. The results of this coarse post hoc analysis are displayed in Table 7. Rank sum scores for regression modeling approaches range from 10 (two regularized approaches: lasso regression and ridge regression using the full Substantive Scale set as predictors) to 48 (random forests regression using only the RC Scales as predictors). Four of the five top-performing methods are automated feature selection methods (the lasso, forward-and-backward stepwise regression, backward deletion, and forward selection, in that order). Models using all Substantive Scales as predictors generally performed better than models using only the RC Scales as predictors. Models constructed using manual (theory-based) feature selection or using a single, automatically selected predictor were systematically outperformed by automated feature selection methods permitted to consider and include multiple predictors.

Classification Models

Classification models predicting diagnostic group membership are presented in Supplemental Tables S9–S12 in the online supplemental materials. As with the regression models described above, model performance metrics for training data, *k*-fold cross-

Table 7
Rank Sum Scores for External Validation Performance of Regression Models Based on RMSE

Modeling approach	Feature set	Rank sum score
Lasso	Feature selection	10
Ridge	All substantive scales	10
Stepwise selection	Feature selection	15
Backward deletion	Feature selection	19
Forward selection	Feature selection	22
Random forests	All substantive scales	27
Support vector regression	All substantive scales	27
Ordinary least squares	All substantive scales	28
Manual selection	Feature selection	29
Ordinary least squares	RC Scales	30
Ridge	RC Scales	34
Single predictor	Feature selection	35
Support vector regression	RC Scales	44
Negative binomial regression	RC Scales	46
Random forests	RC Scales	48

Note. RC = Restructured Clinical; RMSE = root mean squared error.

validated test data, and external validation data are presented for the classification models generated as part of the present study. Two external validation samples were available for the present study, and external validation model performance statistics are therefore presented for both samples. External validation performance across all classification models is summarized in Table 8 for classification accuracy and in Table 9 for AUC.

Similar to the regression models already described, performance for the classification models used in the present study almost always appears strongest in the training data, next strongest in the *k*-fold cross-validation data, and least strong in the external validation data, regardless of the model performance criterion under consideration. An exception to this is observed when using classification accuracy to evaluate the performance of models with low base rates in the external validation sample (e.g., thought disorder diagnosis in the community mental health sample). In these cases, classification accuracy tends to appear much higher in the external validation sample than in the training sample, likely owing to the increasing influence of base rates on classification accuracy as base rates approach their extremes. As an additional exception to the general performance rule (training performance > cross-validated performance > externally validated performance), some models predicting substance use disorders produce higher performance in the external validation data than in the training data, mirroring similar findings from the regression analyses. Additionally, again echoing regression analysis results, many models demonstrated substantial performance shrinkage across settings. As a representative example, a standard logistic regression model predicting depressive disorder diagnosis classified such diagnoses with 72% accuracy in the county hospital training sample but with only 63% accuracy in the VA medical center test sample.

External validation performance of classification models appears to vary considerably between some prediction criteria, though various models predicting the same criterion variables tend to produce performance values in the same range within each setting, similar to findings from the regression model analyses. Some, but not all, models produce very different external validation performances between settings, particularly when base rates differ markedly. The highest-performing classification methods for each diagnostic variable are not listed here, as such a determination varies depending on both the performance metric used and the external validation setting; instead, the reader is referred to Tables 8 and 9.

Prediction of somatic disorder diagnosis in the present data appears particularly problematic, relative to prediction of other classification criteria. The somatic disorder diagnostic variable differs from other diagnostic variables in the present study in its low base rate across all samples (see Table 3),[13] an issue that typically limits predictive power at both training and testing stages. Indeed, many of the models predicting somatic disorder classification appear to predict no better than chance. Furthermore, overall predictive performance of somatic disorder classifiers appears relatively low compared with classifiers for other diagnostic groups. In addition, the predictor weights calibrated in some approaches to somatic disorder prediction (namely, forward selection, backward deletion, and forward-and-backward stepwise selection) are wildly out of bounds compared with the weights calibrated via other logistic regression approaches, indicating extreme overfitting to the training data. Nevertheless, several modeling approaches appear to substantially improve upon chance-level prediction of somatic disorder classification (e.g., lasso AUC = .70 [VA], .64 [community]).

As a post hoc analysis, rank sum scores were calculated for the classification models produced in the present study. This process was similar to the one used to calculate rank sum scores for the regression models, with some notable exceptions. First, only AUC performance was considered, as classification accuracy proved a problematic metric for evaluating performance in low base rate settings, as previously described. Second, model performance in predicting somatic disorder diagnosis was not considered because of the already-discussed problems in training and testing those models resulting from low statistical power. Third, performance was considered across both external validation settings. Thus, six ranks were calculated for each modeling approach: one for each of the three remaining classification criteria (depressive disorder diagnosis, substance use disorder diagnosis, and thought disorder diagnosis) in both external validation settings (VAMC and community mental health center). The results of this analysis are displayed in Table 10. Overall, the findings of this rank sum score analysis appear similar to the results of the corresponding analysis conducted for regression modeling approaches. Two of the top five modeling approaches, of the 14 total classification approaches considered, are regularization methods (the lasso and ridge regression using all Substantive Scales as predictors). Four of the top five approaches use automated feature selection of potentially multiple predictors (forward selection, the lasso, forward-and-backward stepwise selection, and backward deletion, in that order). The least generally effective classification methods in the present study, with respect to generalizability, appear to be support vector machines using all Substantive Scales, tied with single-predictor selection. In general, classification models incorporating predic-

[13] The low base rate of clinically diagnosable somatization in these samples is not surprising, as individuals with such problems tend to present at medical settings, rather than mental health care settings, for treatment.

Table 8
External Validation Performance of Classification Models (Overall Classification Accuracy)

Model	VA medical center				Community outpatient			
	Depressive	Substance	Thought	Somatic	Depressive	Substance	Thought	Somatic
RC Scales								
Maximum likelihood	.63	.71	.85	.98	.59	.73	.98	1.00
Ridge	.64	.71	.86	.98	.60	.74	.98	1.00
Random forests	.63	.70	.85	.98	.59	.69	.98	1.00
Support vector machine	.57	.65	.86	.98	.52	.74	.98	.99
All substantive scales								
Maximum likelihood	.64	.75	.86	.97	.64	.79	.97	.98
Ridge	.65	.75	.86	.98	.64	.79	.98	1.00
Random forests	.63	.76	.86	.98	.62	.73	.98	1.00
Support vector machine	.65	.75	.84	.98	.64	.77	.98	1.00
Feature selection								
Single predictor	.59	.71	.86	.98	.59	.70	.98	1.00
Manual selection	.62	.74	.86	.98	.61	.78	.99	1.00
Forward selection	.65	.75	.86	.97	.66	.79	.97	.98
Backward deletion	.65	.75	.86	.98	.65	.79	.97	.98
Stepwise selection	.65	.75	.86	.97	.66	.79	.97	.98
Lasso	.64	.76	.86	.98	.66	.81	.98	1.00

Note. Depressive = depressive disorders; Substance = substance use disorders; Thought = thought disorders; Somatic = somatic disorders; RC = Restructured Clinical; VA = Veterans Affairs.

tive information from all Substantive Scales perform better than models using only the RC Scales as predictors.

Discussion

Results of the present study provide insight into several issues related to the generalizability of statistical prediction in the domain of psychological assessment. In the following discussion, the author first considers the extent to which statistical prediction broadly appears to generalize across settings, as well as the extent to which the loss of generalizability appears, itself, to be predictable. This is followed by commentary on the apparent association between model parsimony and model generalizability, including a discussion of the (perhaps counterintuitive) finding that emphasizing parsimony may actually diminish predictive generalizability. Next, the author discusses the impact of various approaches to predictor selection on generalizability, particularly with respect to theory-driven predictor selection, automated multivariate approaches (e.g., stepwise-type methods), and empirically driven approaches that focus on identifying and applying single predictors in isolation. This is then followed by a discussion of the relative performances of the various modeling approaches considered in the present study. Finally, the author discusses limitations of the present study and directions for future research, then summarizes the major findings and implications of the study. As before, the training data referenced in the paper are the data collected from the inpatient county hospital setting, which were used to calibrate all models; cross-validation refers to use of the *k*-fold cross-validation resampling procedure on the training data; and the external validation data are the data sets collected from the VAMC inpatient and community outpatient settings, which were used to evaluate the generalizable predictive validity of the models developed in the training setting.

Overall Generalizability of Statistical Prediction in Psychological Assessment

The results of the present study support the longstanding concern in the psychology literature that models based on psychometric data often do not perform as expected in applied settings when that expectation is based upon a model's performance in its training setting. In most cases, the models developed as part of the present study fit the external validation data noticeably more poorly than they fit the training data. For example, many of the linear regression models evaluated tended to show R^2 values approximately half as large in the validation sample as they did in the training sample (e.g., R^2 dropping from approximately .20 to approximately .10, corresponding to a loss of predictive power of approximately 29%[14]). None of the considered modeling approaches appeared immune to this shrinkage effect, although some modeling approaches appeared to consistently produce more generalizable predictions than others. Nevertheless, many models demonstrated non-negligible predictive power in the external validation data despite significant performance shrinkage.

The extent to which models generalized in the present study varied significantly based not only on the modeling approach considered but also on the criterion variable. Somewhat unexpectedly, models predicting criterion variables related to substance abuse (i.e., the substance abuse record review form scale, substance abuse-related clinical diagnosis) were particularly robust against performance shrinkage, often performing approximately as

[14] This value is 29%, rather than 50%, because effect size is more accurately represented by the multiple correlation than the squared multiple correlation (i.e., R instead of R^2). Thus, the proportion of the effect in the test data to the effect in the training data is $\frac{\sqrt{.10}}{\sqrt{.20}} = .71$, and the effect size loss is $1 - .71 = .29$ (Darlington, 1990; Funder & Ozer, 2019; Ozer, 1985).

Table 9
External Validation Performance of Classification Models (Area Under the Curve)

Model	VA medical center				Community outpatient			
	Depressive	Substance	Thought	Somatic	Depressive	Substance	Thought	Somatic
RC Scales								
Maximum likelihood	.70	.77	.74	.61	.69	.77	.75	.62
Ridge	.70	.77	.75	.62	.69	.77	.75	.72
Random forests	.69	.74	.72	.56	.66	.75	.75	.61
Support vector machine	.67	.69	.67	.64	.66	.73	.76	.56
All substantive scales								
Maximum likelihood	.72	.80	.74	.51	.70	.86	.72	.60
Ridge	.72	.81	.76	.67	.69	.84	.76	.72
Random forests	.70	.80	.71	.68	.68	.83	.71	.72
Support vector machine	.72	.81	.70	.66	.69	.82	.58	.74
Feature selection								
Single predictor	.64	.77	.58	.49	.69	.76	.59	.63
Manual selection	.68	.82	.67	.61	.69	.84	.67	.49
Forward selection	.71	.81	.76	.50	.70	.84	.77	.49
Backward deletion	.71	.81	.75	.50	.70	.86	.74	.49
Stepwise selection	.71	.81	.75	.50	.70	.84	.77	.49
Lasso	.71	.82	.76	.70	.70	.85	.75	.64

Note. Depressive = depressive disorders; Substance = substance use disorders; Thought = thought disorders; Somatic = somatic disorders; RC = Restructured Clinical; VA = Veterans Affairs.

well in external validation samples as in the training sample (sometimes even better). The reason for this outlying finding, which violates the shrinkage trend observed with all other prediction criteria examined in this study, is not entirely clear. Of note, the predictor pool in the present study included a particular MMPI-2-RF scale, SUB (Substance Abuse, which also shares items with related externalizing RC4 and BXD), that maps very closely onto the substance use-related prediction criteria at both conceptual and zero-order empirical levels. It is therefore also possible, if not likely, that this especially close association further enhanced prediction generalizability in the present study (though as indicated by other results, discussed in greater detail below, strong theoretical association alone appears insufficient to maximize model generalizability). Generalizability across settings was likely further bolstered by similarities in the distributions of substance abuse-related prediction criteria between settings.

Table 10
Rank Sum Scores for External Validation Performance of Classification Models Based on AUC

Modeling approach	Feature set	Rank sum score
Forward selection	Feature selection	14
Lasso	Feature selection	15
Stepwise selection	Feature selection	17
Ridge	All substantive scales	18
Backward deletion	Feature selection	22
Maximum likelihood	All substantive scales	28
Ridge	RC Scales	43
Support vector machines	All substantive scales	44
Maximum likelihood	RC Scales	46
Manual selection	Feature selection	47
Random forests	All substantive scales	57
Random forests	RC Scales	64
Support vector machines	RC Scales	69
Single predictor	Feature selection	69

Note. AUC = area under the curve; RC = Restructured Clinical.

K-fold cross-validation performance estimates in the present study, though generally more conservative than training data estimates of performance, nevertheless typically hewed closely to those training data estimates. In cases for which a significant drop in model performance was observed when porting models from training data to external validation data, the *k*-fold cross-validation estimates of model performance tended not to reflect the extent to which predictive power was attenuated across settings, indicating that such estimates should not necessarily be taken as good estimates of generalizable predictive power across settings. In a notable exception to this pattern, some of the newer machine learning algorithms considered in the present study (namely, random forests and support vector regression/machines) evidenced extreme overfitting to the training data, predicting data points in the training sample with perfect or near-perfect accuracy. However, these models' cross-validated performance estimates fell roughly in line with those of the other methods considered, as did their external validation performance estimates. This result is neither unexpected nor problematic, as those predictive methods were explicitly designed to predict new data points, and their training data performances are therefore not particularly meaningful. The results of the present study indicate that *k*-fold cross-validation model performance is insufficient to predict generalizable model performance across settings, although one could reasonably infer, as is often the case in machine learning studies, that it is a good estimate of model performance within the training setting (i.e., if new observations were drawn from the same population or process that generated the training sample).

The results of the present study do not point to a simple rule for anticipating the extent to which any given model will generalize across settings. Certainly, any a priori assumption that a model will demonstrate predictive power in an applied setting similar to its

predictive power in its training data is unwarranted. Any general assumption that a model will have no practical predictive validity outside of its training setting is also unjustified, given that many models in the present study demonstrated meaningful predictive power in the external validation sample, even when substantial performance shrinkage was evident. How, then, should one estimate the expected performance of a model in a new, potential applied setting? The simplest and most effective option is to test the model on known data from that setting, if possible. It generally takes significantly fewer observations to test than to appropriately train a model, making model testing potentially more economically feasible than model development in a new setting. Barring that, a clinician must calibrate their expectations more informally, considering information such as the potential loss of predictive power indicated by past research (e.g., the present study), shifts in base rates and other distributional characteristics between settings, and qualitative differences between the settings. Depending on the nature and quantity of available information, a researcher can also apply a formal corrective procedure, such as model updating, to account for differences between settings, even when lacking sufficient data to either formally test an existing model or to train a new one (see the Future Directions section below). The extent to which a loss in predictive power between settings is acceptable is ultimately a matter of professional judgment that will depend on both the model and the purpose to which it is applied.

The Predictive Value of Parsimony, or Lack Thereof

The results of the present study are inconsistent with conventional recommendations that psychology researchers prioritize parsimony in model construction, at least when maximizing the model's predictive power is of primary interest. In the present study, models demonstrated a systematic tendency toward stronger predictive performance when they were permitted to incorporate more features, not fewer, at the training stage. In fact, the most parsimonious models (i.e., those employing only a single predictor or a small set of manually specified predictors) were often among the least predictively powerful when compared with more complex alternatives. Furthermore, models constructed using only a circumscribed, static scale set (the MMPI-2-RF RC Scales) as predictors were generally outperformed by models permitted to simultaneously incorporate predictive information from the full pool of potential predictors (all MMPI-2-RF substantive scales). These findings are perhaps especially surprising considering that increased complexity is generally associated with increased risk of overfitting.

In the present study, models benefitted from a large training sample size (much larger than one encounters in many psychology studies), which enabled estimation of seemingly stable model parameters, even when a relatively large number of predictors were retained. Indeed, it appears that when the sample size offers sufficient statistical power in a psychology study, and when one wishes to enhance the generalizable predictive power of the model developed, one could consider including all the potential predictors that the sample size would support. These criteria will not always be met, of course. When the number of potential predictors is relatively large given the number of training data observations (when the dataset is *wide*, in data science parlance), parsimony may be a statistical necessity, and in other cases, model developers may very reasonably wish to balance predictive power with interpretability.

Optimal Feature Selection Methods for Predictive Modeling in Psychological Assessment

Perhaps one of the most striking findings from the current study is that automated feature selection methods (the lasso and stepwise-family methods) produced some of the most consistently powerful predictive models, relative to other models considered, when validated on external data, whereas models whose features were selected based on theory were generally among the poorer-performing models. These results appear to strongly contradict usual best practice feature selection guidelines in psychology research, which both advocate for theory-based feature selection and condemn most automated approaches. Results of the present study also appear to devalue a certain practice in psychology, particularly common in the assessment literature: namely, the tendency to focus on identification and interpretation of applied single-predictor models (e.g., single-scale interpretation), even in the presence of other potential ancillary predictors. In other words, single-predictor models, even when empirically selected, did not generalize well across settings relative to multiple-predictor models in the present study.

An inspection of the predictor weights and other variable importance metrics from the more complex models constructed in the present study provides some insight into the reason complex and empirically selected models tended to outperform more parsimonious and theory-based models. When selecting features based on theory, which may not be specific enough to meaningfully guide prediction, researchers tend to focus on predictors with strong, positive associations with the target criterion. However, many of the larger and more complex models in the present study appear to have optimized their predictive power by detecting and appropriately weighting not only strong convergent features, but also negatively associated features, useful ancillary features with weaker theoretical associations to the criterion, and suppressor variables that a human researcher would not necessarily have thought to include. Though not explicitly designed to do so, the complex and automatically selected models in the present study achieved relatively good predictive performance by learning to discriminate between some of the institutional pseudo-taxa in the sample (i.e., the latent groups representing alternative pathways into mental health treatment, such as suicidality or psychosis; concept credited to Tellegen in Meehl, 1992[15]; see also Grove, 1991) and by detecting important features that a human might have overlooked. For example, prediction of delusions and thought disorders was augmented by not only considering MMPI-2-RF scales positively associated with thought dysfunction, such as RC6 (Persecutory Ideation) and RC8 (Aberrant Experiences) but also

[15] Meehl (1992) writes:

> In taxometrics one has the problem of pseudo-taxa, of data sets that may behave taxonically when examined by whatever taxometric method, but that are in some sense spurious, artifactual, not 'real entities.' This concern was raised most forcibly on the local scene by Auke Tellegen and has therefore come to be designated in Minnesota circles as the *Tellegen Case*.

by negatively weighting features more distinctively associated with internalizing problems (e.g., depression), such as RCd (Demoralization). One might reasonably hypothesize that such modeling approaches are similarly capable of detecting and leveraging important but perhaps difficult to anticipate statistical associations between variables in other applied settings.

The stepwise-type approaches considered in the present study likely benefited from the large size of the training sample. As such, these statistically greedy methods, which require statistical power similar to the power needed by an OLS/MLE model incorporating *all* potential predictors (see Babyak, 2004), are unlikely to perform as well under some conditions of lower statistical power (one of the principal reasons such methods have often performed poorly in past research). Indeed, in the current study, stepwise-type approaches failed to produce stable, generalizable parameter estimates when statistical power was limited by the low training data base rate of one of prediction criteria; namely, somatic disorder diagnosis. Of note, however, models trained using the lasso and other regularization methods tend to require less statistical power than models parameterized using traditional OLS/MLE optimization, including stepwise-type approaches.[16] As such, the lasso appears to be a viable and user-friendly alternative for regression with feature selection even, perhaps, when low statistical power would not permit use of stepwise-type approaches. In the case of prediction of somatic disorder diagnosis in the present study, the lasso did indeed produce a relatively stable and generalizable predictive model, whereas models produced using stepwise-type methods appeared to fare no better than chance when applied to the external validation data.

Selecting a Predictive Modeling Approach in Psychological Assessment

The results of the present study can be used to inform the selection of a predictive modeling approach when working with psychometric data. There is perhaps no conclusive process for determining the extent to which a particular method is likely to outperform another without direct performance data, given the variability in predictive performance for each method across prediction criteria. In the present study, the author produced rank orderings for modeling approaches by aggregating relative external validation performances across predictions for each method (see Tables 7 and 10), though this method does not fully convey the magnitude of the differences in model performance and could therefore lead to over- or underestimates of the magnitude of such differences. Indeed, observed differences in absolute model fit for competing models appeared, in some cases, very small (e.g., accurately detecting 1% more cases of a target diagnostic group). In machine learning studies, particularly those in which multiple competing models predicting the same criterion are compared, this is not unusual, and preference is usually given to the model or modeling approach whose performance suggests the potential for greatest generalizability, even when its advantage over competitors appears minor (and, everything else being equal, preference is generally given to more parsimonious models). When applying comparative methods in the domain of psychology research, it is perhaps best left to the researcher (or the consumer of research) to determine whether the statistical differences observed between models are substantial enough to warrant selection of one model over another. In psychology, even very small effect size differences can be clinically significant in the long run, and such differences can be very meaningful for a subset of those affected by the prediction (Funder & Ozer, 2019).

Similarly, because the differences in performance between some of the methods evaluated in some cases are rather slight, in applied practice the extent to which any difference would be considered meaningful would depend not only on quantitative differences in model fit, but also on the context of prediction, which would include consideration of the nature of the prediction criterion as well as the manner in which prediction would be used to guide decision-making. For example, in predicting a high-stakes criterion such as suicide risk or sex offender recidivism, prior to making an important clinical decision or formulating a psycholegal opinion, a psychologist might favor any prediction method that offers a robust advantage over another, even if that advantage is relatively small in terms of absolute model fit. This important point is perhaps easy to overlook in the present study, given that the prediction criteria are somewhat arbitrary, and the models described herein were developed for methodological analyses, not for direct applied use. Regardless, if a researcher wishes to develop a prediction model using a less robust method than available alternatives, the burden is arguably on that researcher to justify their selection of modeling approach.

Although the methods used in the present study are relatively simple to apply when compared with advanced machine learning methods (indeed, that is one reason they were selected), some of the present methods are undoubtedly unfamiliar to most psychology researchers and are at least modestly more complex to use properly than conventional linear or logistic regression. The complexity, or perceived complexity, of these techniques will therefore surely influence some researchers' decisions to use them or not. The results of the present study do not indicate that conventional prediction techniques are invalid for generating generalizable predictions; on the contrary, in the present data they performed moderately well even when compared with much newer machine learning approaches. Furthermore, in many cases any reasonably valid prediction model is preferable to none at all. As such, conventional modeling methods remain viable in at least some applied prediction contexts in assessment psychology; however, a researcher who chooses to employ such a method should do so with the understanding that more robust and generalizable alternatives are likely available.

Comparison of model optimization methods. As just noted, conventional OLS/MLE methods performed moderately well overall. However, regularized methods (i.e., the ridge and the lasso) were remarkable for fairly consistently producing models of comparable or greater generalizable predictive power than those produced by OLS/MLE methods. This is perhaps especially noteworthy given that such regularized models take the familiar form of linear or logistic regression models and can be interpreted similarly by anyone so inclined. The relative advantage of regularized regression over OLS/MLE regression appeared greater in the pres-

[16] Regularized regression uses fewer effective degrees of freedom than OLS/MLE regression because the shrinkage penalty allows the model to reclaim degrees of freedom from the full OLS/MLE model parameters (i.e., the predictor weights).

ent study as the number of potential predictors increased (e.g., when using all Substantive Scales rather than simply just the RC Scales).

Some researchers have recently advocated for greater use of count regression models over linear regression models in psychological assessment research, as many of the regression criteria used in that research domain take the form of count variables, and many such criteria follow skewed, nonnormal distributions in the training data that would technically violate the assumptions of conventional linear regression. Negative binomial regression was selected as a representative count regression modeling approach in the present study and was tested against linear regression methods. Perhaps surprisingly, negative binomial models tended to perform relatively poorly when compared with conventional linear models, at least when evaluated based on RMSE. The most direct potential explanation for this finding appears to be that despite taking the form of count variables, the regression criteria in the present study nevertheless fit poorly to the negative binomial distribution and are better approximated using linear combinations of predictors. Thus, broad recommendations for count regression modeling replacing linear modeling in assessment research appear at least partly unwarranted. At the very least, these results indicate a need to consider and, ideally, test alternative modeling approaches before selecting a negative binomial (or other count regression) model as a final predictive approach when working with psychometric data.

Effectiveness of advanced machine learning predictive algorithms. Unexpectedly, the random forests and support vector prediction models developed and applied in the present study demonstrated underwhelming, if somewhat uneven, generalizable predictive performances relative to many of the other approaches considered. This is surprising given that those methods are highly regarded in the machine learning literature (see, e.g., Efron & Hastie, 2016; Hastie et al., 2009) and have demonstrated top-of-the-line predictive performance in past comparative studies in other substantive domains (Caruana & Niculescu-Mizil, 2006; Fernández-Delgado et al., 2014). In light of the novelty of this finding, it would be premature to conclude that the present study refutes volumes of past research supporting the effectiveness of those predictive algorithms. At least two possible explanations could help account for the discrepancy between expected and observed performance for the random forests and support vector models. First, the present study strongly emphasized the off-the-shelf performance of these newer machine-learning algorithms. Thus, the relatively poor performance of random forests and support vector models may reflect an overreliance on default tuning variable values, in which case, those methods' performance could perhaps be improved through more careful calibration of tuning parameters (a possible direction for future research).

A second possible explanation for the relatively modest performance of advanced machine learning algorithms in the present study, not mutually exclusive with the one just discussed, is that those methods are not optimal for applied predictions using psychometric data. Consider, for example, that effect sizes in psychology (including generalizable model performance) tend to be rather low compared with those seen in many other fields in which machine learning is applied. It is possible that the low signal-to-noise ratio in the present data reduced the random forests and support vector algorithms' ability to learn effectively from the data to such an extent that those methods were unable to distinguish themselves from more conventional approaches, though it is notable that random forests and support vector methods fared no better in predicting substance abuse variables (which generally evidenced stronger associations with predictors) than they did in predicting other variables. In addition, it is possible that those methods are not especially robust to cross-setting validation when the training and validation settings are dissimilar, meaning that they may suffer from some of the same practical limitations to generalizability affecting more conventional approaches to statistical prediction.

Limitations and Future Directions

The present study evaluated only a small fraction of the available methods used in areas of applied statistical prediction, and selection of methods was biased toward those considered especially popular or simple to apply. Furthermore, only a small subset of the vast number of possible psychometric predictors (the MMPI-2-RF Substantive Scales) and prediction criteria were considered, and these were evaluated in only a few clinical settings. As such, a good deal of additional comparative research is needed to evaluate the generalizability of other predictive methods, to understand the predictive associations between other feature pools and prediction criteria, and to evaluate the generalizability of those predictions across other settings. Although the factors that influence the manner and extent to which statistical prediction models vary between settings are poorly understood at this time, they are, in theory, possible to systematically study.

To support such research, it would be beneficial to identify a mutually consensual set of practically important prediction criteria that can be meaningfully evaluated between settings. This would provide a structured research framework that would allow a systematic evaluation of how prediction of those criteria differs across settings and predictive methods and ultimately bolster our ability to meaningfully predict. To maximize the utility of these predictive models, their prediction criteria should be closely tied to real-world decision making.

The applied statistics literature offers methods for correcting for differences between settings to minimize potential shrinkage. For example, a simple and reasonably effective approach is to update the model intercept to reflect differences in the average value of the criterion variable between settings. In this case, the only information that must be gathered from the applied setting is the average value of the criterion. Bayesian updating, which allows for reestimation of some or all model parameters, taking into account information from both settings, is a more advanced approach, though this typically requires more detailed information. For an overview of these and other model updating methods, see Steyerberg (2009). Like any other prediction models, these updated models can be tested empirically to evaluate the validity of the correction procedures and the extent to which they enhance model generalizabilty.

Alternative versions of many of the algorithms used in the present study exist, some of them augmented by other statistical features. For example, regularization was considered in the present study only as it applies to linear and logistic regression; however, regularization has also been used to enhance the generalizability of other forms of statistical prediction, including support vector machines and other forms of regression (e.g., negative binomial

regression). The effectiveness of these methods with respect to generalizable prediction of psychological variables has not yet been assessed.

The negative binomial models developed as part of the present study's regression analyses performed relatively poorly compared with many of the other regression models considered. However, this does not mean that other count modeling approaches (of which there are many) will necessarily perform similarly poorly, or that the negative binomial model is entirely inappropriate for working with count variables in psychology. Future comparative studies could evaluate alternative forms of count regression that may better fit the prediction criteria, including zero-inflated and zero-truncated methods, hurdle models, and advanced methods such as generalized Poisson regression (Consul & Famoye, 1992) or COM-Poisson regression (Sellers & Shmueli, 2010).

One noteworthy finding from the present study was that predictive generalizability was generally enhanced when the model was permitted to incorporate more, rather than less predictive information (either through automated multivariable feature selection or through inclusion of numerous potentially important predictors), assuming that the training data support a model with that level of complexity. One implication of this finding, combined with clinical observations in the field of psychological assessment that assessment findings are more valid when the clinician is able to incorporate information from multiple sources, is that statistical prediction may be enhanced when driven by multimethod assessment data. For example, rather than generating a predictive model using only data from a particular self-report test (even a broadband test such as the MMPI-2-RF), such a model could also potentially incorporate, for example, data from additional tests, demographic data, encoded behavioral observations, interview responses, and informant report data. Additionally, prediction could perhaps be enhanced by inclusion of validity indicators (whether embedded or independent), which could also enable meaningful prediction even for individuals with so-called invalid protocols. Test data need not take the form of scale scores; for example, some predictive models using item-level data rather than scale scores may benefit from a more granular approach to prediction while potentially reducing prediction error due to variance accounted for by items not meaningfully predictive of the criterion variable. When the number of items is very large, this approach is likely feasible only when using a comparably large sample size, feature selection, and/or a dimension reduction technique.

The statistical models developed in the present study were trained to predict criteria that were not operationalized optimally and likely contained significant unreliability. For example, clinical diagnoses were not standardized using a structured diagnostic interview, and record review form scales were limited by the information that had been included in clinical records when data collection was completed. Although these issues reflect common limitations of clinical research, and are therefore consistent with the present study's overall goal of evaluating the generalizability of the types of predictive models that are most likely to be generated from clinical field data, the unreliability inherent in these prediction criteria likely attenuated the predictive power of this study's models. In future research oriented toward development of useful applied predictive models, such models should ideally be trained on the most reliable and valid data feasible.

Summary and Conclusions

The results of the present study carry significant implications for the practice of predictive modeling using psychometric data. By applying a model evaluation framework rooted in fundamental machine learning practice, the author appraised the generalizability of a range of predictive modeling approaches, including methods used both in psychology research and in modern machine learning. Results support a longstanding concern in psychology that statistical prediction models tend to overfit their calibration data. Such models may nevertheless demonstrate acceptable predictive power in new settings, though the extent to which a loss of predictive power across settings is acceptable will depend in practice upon the context of prediction. The extent to which any given model will generalize to a new setting is difficult to anticipate without actually testing it in the new setting, as even *k*-fold cross-validation, a machine learning tool designed to provide a relatively unbiased estimate of model performance in the training setting, appeared incapable of predicting significant performance shrinkage (or lack thereof) across settings in the present study.

The present study's findings appear to raise significant questions regarding some conventional recommendations for predictive modeling practices in psychology. Perhaps most importantly, the results of this study suggest that psychological theory may be overvalued as a means for selecting predictors to include in statistical models, at least when generalizable model performance is a significant concern. In fact, in the present study, models whose predictors were selected based on theory were typically among the poorest performing with respect to generalizability of prediction. Of course, the lackluster performance of manually selected models in the present study may reflect not a problem with theory-based predictor selection in general, but with the author's ability to skillfully apply theory to predictor selection, or perhaps even with the theory itself. In the present study, theory-based predictor selection was guided by extensive familiarity with the MMPI-2-RF as well as by familiarity with the clinical personality assessment literature. Nevertheless, any number of theoretically plausible alternative models could be considered, some of which may outperform the author's own. However, the question of whether theory-driven approaches to predictor selection will generally tend to outperform automatic or static predictor selection in similar research contexts is ultimately an empirical one, and the early data provided by the present study weighs against theory-based approaches. As previously noted, it appears unlikely that a researcher could consistently anticipate a priori the sorts of associations in the data that appeared to give more complex and automatically selected models an edge in generalizability in the present study.

Results of the present study also contradict concerns in the psychology literature that multivariate models whose predictors are selected automatically, such as by stepwise-type methods, are almost necessarily excessively overfit to their calibration data, relative to other modeling approaches. On the contrary, in the present study, models whose features were selected automatically in this fashion tended to evidence stronger generalizable predictive power than models whose features were selected in other ways (e.g., based on theory). Model parsimony also was not advanta-

geous, as conventional wisdom in psychology research would suggest. Indeed, results of the present study indicate that including as many potentially important predictors as are supported by the sample size of the training data may be an effective way to enhance the generalizable predictive power of a statistical model. Traditional ordinary least squares linear regression and maximum likelihood estimation logistic regression appear modestly robust to loss of generalizability across settings. However, models produced by regularized (penalized) regression appeared even more generalizable, and because such models can be interpreted in a similar fashion to traditional regression models, they may be especially viable alternatives to conventional regression methods.[17]

As just discussed, the results of the present study indicate that theory is perhaps a poorer basis for feature selection than conventional guidelines would suggest. However, even if such a position were adopted by the psychology community, theory would continue to play an important role in psychological assessment research. The author notes that although the pool of potential predictors used in the present study (the MMPI-2-RF Substantive Scales) is broad in scope, these predictors were not included in the study (nor, indeed, in the MMPI-2-RF itself) arbitrarily or randomly. Rather, this predictor set was used because the MMPI-2-RF scales cover a broad range of complementary psychological constructs, each potentially relevant, both conceptually and empirically, to the prediction of psychological or behavioral variables, and thus potentially to the prediction criteria in this study. In the larger context of applied predictive modeling, application of theory is necessary to guide development of potential predictors. It is needed to identify potential constructs for measurement and prediction, and it is needed to guide the development and refinement of methods for measuring those constructs. Furthermore, findings from predictive studies in which psychological constructs are measured lead to improved understanding of those constructs and of the nomological net as a whole, leading to improvements in measurement, which further enhance prediction, leading to improvements in theory, and so on, in an iterative, continual growth process that benefits theory, measurement, and prediction (Cronbach & Meehl, 1955).

Although it is the author's position, bolstered in part by the results of the present study, that psychologists would likely benefit substantially from greater use of machine learning-informed methods for developing, evaluating, and applying predictive models, such a change would constitute a significant paradigm shift in psychology research and practice (see also Yarkoni & Westfall, 2017). Components of this paradigm shift could include, but are certainly not limited to, considering a greater range of statistical modeling approaches (including some that are presently unfamiliar to most psychology researchers), shifting focus from training data performance to test data performance, relaxing or abandoning some conventional modeling principles when they are not supported by external validation data, and accepting statistical prediction as a valid basis for decision-making in applied psychology. To accomplish such a shift wholesale appears improbable, given the depth with which some countervailing methodological principles are rooted in the tradition of psychological science. As the present empirical study demonstrates, however, basic machine learning methods and principles can be readily applied to psychological data and do not necessarily require an excessive amount of foundational background knowledge to comprehend or require access to exclusive software or hardware to utilize.[18] The principle benefit of the proposed paradigm shift, even if such a shift occurs gradually, through thoughtful but tentative studies of applied predictive modeling in psychology, is the potential to substantially improve real-world decision-making using psychological data across all domains in which such data are collected.

[17] Implementations of regularized regression can be found in many popular statistical software packages, including SPSS.

[18] All predictive modeling in the present study was carried out on a home computer using R (R Core Team, 2015), a widely used, free, and open-source statistical software package. Extensive tutorials for using R and for developing and evaluating predictive models within R are freely available on the internet and are also found in many excellent statistics texts. The R code used to develop and evaluate the models described in this study can be found in this paper's online supplemental materials.

References

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., & Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist, 34,* 341–382. http://dx.doi.org/10.1177/0011000005285875

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & B. F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest, Hungary: Academiai Kiado.

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental health disorders (DSM–III–R)*. Washington, DC: Author.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders: DSM–IV*. Washington, DC: Author.

Anderson, N. B., & the APA Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist, 61,* 271–285. http://dx.doi.org/10.1037/0003-066X.61.4.271

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine, 66,* 411–421.

Ben-Porath, Y. S. (2012a). Addressing challenges to MMPI-2-RF-based testimony: Questions and answers. *Archives of Clinical Neuropsychology, 27,* 691–705. http://dx.doi.org/10.1093/arclin/acs083

Ben-Porath, Y. S. (2012b). *Interpreting the MMPI-2-RF*. Minneapolis, MN: University of Minnesota Press.

Ben-Porath, Y. S., & Tellegen, A. (2008). *MMPI-2-RF: Manual for administration, scoring and interpretation*. Minneapolis, MN: University of Minnesota Press.

Breiman, L. (1996a). Bagging predictors. *Machine Learning, 24,* 123–140. http://dx.doi.org/10.1007/BF00058655

Breiman, L. (1996b). Heuristics of instability and stabilization in model selection. *Annals of Statistics, 24,* 2350–2383. http://dx.doi.org/10.1214/aos/1032181158

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*. Boca Raton, FL: CRC press.

Browne, M. W. (2000). Cross-Validation Methods. *Journal of Mathematical Psychology, 44,* 108–132. http://dx.doi.org/10.1006/jmps.1999.1279

Busemeyer, J. R., & Wang, Y. M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology, 44,* 171–189. http://dx.doi.org/10.1006/jmps.1999.1282

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *MMPI-2: Manual for administration and scoring*. Minneapolis, MN: Pearson.

Caruana, R., & Niculescu-Mizil, A. (2006, June). *An empirical comparison of supervised learning algorithms*. In *Proceedings of the 23rd international conference on Machine learning* (pp. 161–168). ACM.

Consul, P., & Famoye, F. (1992). Generalized Poisson regression model. *Communications in Statistics Theory and Methods, 21,* 89–109. http://dx.doi.org/10.1080/03610929208830766

Copas, J. B. (1983). Regression, prediction and shrinkage. *Journal of the Royal Statistical Society Series B. Methodological, 45,* 311–355. http://dx.doi.org/10.1111/j.2517-6161.1983.tb01258.x

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302. http://dx.doi.org/10.1037/h0040957

Darlington, R. B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin, 69,* 161–182. http://dx.doi.org/10.1037/h0025471

Darlington, R. B. (1990). *Regression and linear models*. New York, NY: McGraw-Hill.

Derksen, S., & Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology, 45,* 265–282. http://dx.doi.org/10.1111/j.2044-8317.1992.tb00992.x

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., & Weingessel, A. (2005). Misc functions of the department of statistics (e1071), TU Wien. Retrieved from https://cran.r-project.org/web/packages/e1071/index.html

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. In M. C. Mozer, M. I. Jordan, & T. Petsche (Eds.), *Advances in neural information processing systems* (Vol. 9, pp. 155–161). Cambridge, MA: MIT Press.

Efron, B., & Hastie, T. (2016). *Computer age statistical inference* (Vol. 5). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781316576533

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems. *Journal of Machine Learning Research, 15,* 3133–3181.

Freund, Y., & Schapire, R. E. (1995, March). A decision-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory* (pp. 23–37). Berlin, Germany: Springer.

Friedman, J. H. (1997). On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data mining and knowledge discovery, 1,* 55–77.

Friedman, J., Hastie, T., & Tibshirani, R. (2009). glmnet: Lasso and elastic-net regularized generalized linear models. Retrieved from https://cran.r-project.org/web/packages/glmnet/index.html

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science, 2,* 156–168. http://dx.doi.org/10.1177/2515245919847202

Grove, W. M. (1991). Validity of taxometric inferences based on cluster analysis stopping rules. In D. Cichetti & W. M. Grove (Eds.), *Thinking clearly about psychology: Essays in honor of Paul Everett Meehl* (pp. 313–329). Minneapolis, MN: University of Minnesota Press.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law, 2,* 293–323. http://dx.doi.org/10.1037/1076-8971.2.2.293

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12,* 19–30. http://dx.doi.org/10.1037/1040-3590.12.1.19

Gunn, S. R. (1998). Support vector machines for classification and regression. *ISIS Technical Report, 14*.

Harrell, F. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis*. Amsterdam, The Netherlands: Springer. http://dx.doi.org/10.1007/978-3-319-19425-7

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference and prediction* (2nd ed.). New York, NY: Springer. http://dx.doi.org/10.1007/978-0-387-84858-7

Hilbe, J. M. (2011). *Negative binomial regression*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511973420

Hilbe, J. M. (2014). *Modeling count data*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139236065

Hoerl, A. E., & Kennard, R. W. (1981). Ridge regression—1980: Advances, algorithms, and applications. *American Journal of Mathematical and Management Sciences, 1,* 5–83. http://dx.doi.org/10.1080/01966324.1981.10737061

Hunsley, J., & Mash, E. J. (2007). Evidence-based assessment. *Annual Review of Clinical Psychology, 3,* 29–51. http://dx.doi.org/10.1146/annurev.clinpsy.3.022806.091419

Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment, 15,* 446–455. http://dx.doi.org/10.1037/1040-3590.15.4.446

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An introduction to statistical learning*. New York, NY: Springer.

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab-An S4 package for kernel methods in R. *Journal of Statistical Software, 11,* 1–20.

Kearns, M., & Valiant, L. (1994). Cryptographic limitations on learning Boolean formulae and finite automata. *Journal of the Association for Computing Machinery, 41,* 67–95. http://dx.doi.org/10.1145/174644.174647

Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCAI, 14,* 1137–1145.

Kotov, R., Krueger, R. F., Watson, D., Achenbach, T. M., Althoff, R. R., Bagby, R. M., . . . Zimmerman, M. (2017). The hierarchical taxonomy of psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal of Abnormal Psychology, 126,* 454–477. http://dx.doi.org/10.1037/abn0000258

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software, 28,* 1–26. http://dx.doi.org/10.18637/jss.v028.i05

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26). New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4614-6849-3

Le Cessie, S., & Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied Statistics, 41,* 191–201. http://dx.doi.org/10.2307/2347628

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

Linn, R. L. (1968). Range restriction problems in the use of self-selected groups for test validation. *Psychological Bulletin, 69,* 69–73.

Linn, R. L., Harnisch, D. L., & Dunbar, S. B. (1981). Corrections for range restriction: An empirical investigation of conditions resulting in conservative corrections. *Journal of Applied Psychology, 66,* 655–663.

Mattera, D., & Haykin, S. (1999). Support vector machines for dynamic reconstruction of a chaotic system. In B. Schölkopf, C. J. C. Burges, & A. J. Smola (Eds.), *Advances in kernel methods— Support vector learning* (pp. 211–242). Cambridge, MA: MIT Press.

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. http://dx.doi.org/10.1037/11281-000

Meehl, P. E. (1956). Wanted—A good cook-book. *American Psychologist, 11,* 263–272. http://dx.doi.org/10.1037/h0044164

Meehl, P. E. (1957). When shall we use our heads instead of the formula? *Journal of Counseling Psychology, 4,* 268–273. http://dx.doi.org/10.1037/h0047554

Meehl, P. E. (1986). Causes and effects of my disturbing little book. *Journal of Personality Assessment, 50,* 370–375. http://dx.doi.org/10.1207/s15327752jpa5003_6

Meehl, P. E. (1992). Factors and taxa, traits and types, differences of degree and differences in kind. *Journal of Personality, 60,* 117–174. http://dx.doi.org/10.1111/j.1467-6494.1992.tb00269.x

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52,* 194–216. http://dx.doi.org/10.1037/h0048070

Miller, A. (2002). *Subset selection in regression*. Boca Raton, FL: CRC Press. http://dx.doi.org/10.1201/9781420035933

Mosier, C. I. (1951). The need and means of cross validation. I. Problems and designs of cross-validation. *Educational and Psychological Measurement, 11,* 5–11.

Müller, K.-R., Smola, A., Rätsch, G., Schölkopf, B., Kohlmorgen, J., & Vapnik, V. (1997). Predicting time series with support vector machines. In W. Gerstner, A. Germond, M. Hasler, & J.-D. Nicoud (Eds.), *Artificial Neural Networks ICANN'97* (Vol. 1327, pp. 999–1004). Berlin, Germany: Springer Lecture Notes in Computer Science. http://dx.doi.org/10.1007/BFb0020283

Ozer, D. J. (1985). Correlation and the coefficient of determination. *Psychological Bulletin, 97,* 307–315. http://dx.doi.org/10.1037/0033-2909.97.2.307

R Core Team. (2015). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., & Firth, D. (2013). MASS: Support functions and datasets for venables and Ripley's MASS. Retrieved from https://cran.r-project.org/web/packages/MASS/index.html

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

Sellers, K. F., & Shmueli, G. (2010). A flexible regression model for count data. *The Annals of Applied Statistics, 4,* 943–961. http://dx.doi.org/10.1214/09-AOAS306

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing, 14,* 199–222. http://dx.doi.org/10.1023/B:STCO.0000035301.49549.88

Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems, 9,* 155–161.

Statnikov, A., Wang, L., & Aliferis, C. F. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics, 9,* 319. http://dx.doi.org/10.1186/1471-2105-9-319

Steyerberg, E. (2009). *Clinical prediction models: A practical approach to development, validation, and updating*. New York, NY: Springer Science & Business Media. http://dx.doi.org/10.1007/978-0-387-77244-8

Stitson, M., Gammerman, A., Vapnik, V., Vovk, V., Watkins, C., & Weston, J. (1999). Support vector regression with ANOVA decomposition kernels. In B. Schölkopf, C. J. C. Burges, & A. J. Smola (Eds.), *Advances in kernel methods—Support vector learning* (pp. 285–292). Cambridge, MA: MIT Press Cambridge.

Tellegen, A., & Ben-Porath, Y. S. (2008). *MMPI-2-RF: Technical manual*. Minneapolis, MN: University of Minnesota Press.

Tellegen, A., Ben-Porath, Y. S., McNulty, J. L., Arbisi, P. A., Graham, J. R., & Kaemmer, B. (2003). *The MMPI-2 Restructured Clinical (RC) scales: Development, validation, and interpretation*. Minneapolis, MN: University of Minnesota Press.

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

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM, 27,* 1134–1142. http://dx.doi.org/10.1145/1968.1972

Were, K., Bui, D. T., Dick, Ø. B., & Singh, B. R. (2015). A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. *Ecological Indicators, 52,* 394–403. http://dx.doi.org/10.1016/j.ecolind.2014.12.028

Wickham, H. (2017). The tidyverse (R package ver. 1.1. 1). Retrieved from https://cran.r-project.org/web/packages/tidyverse/index.html

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment*. Menlo Park, CA: Addison Wesley.

Wright, A. G., Pincus, A. L., & Lenzenweger, M. F. (2012). An empirical examination of distributional assumptions underlying the relationship between personality disorder symptoms and personality traits. *Journal of Abnormal Psychology, 121,* 699–706. http://dx.doi.org/10.1037/a0029042

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

Received October 21, 2019
Revision received January 14, 2020
Accepted January 14, 2020 ■