HHS Public Access

Author manuscript

J Consult Clin Psychol. Author manuscript; available in PMC 2021 January 01.

Published in final edited form as:

J Consult Clin Psychol. 2020 January ; 88(1): 25–38. doi:10.1037/ccp0000451.

Personalized Prognostic Prediction of Treatment Outcome for Depressed Patients in a Naturalistic Psychiatric Hospital Setting: A Comparison of Machine Learning Approaches

Christian A. Webb, Ph.D.[1], **Zachary D. Cohen, Ph.D.**[2], **Courtney Beard, Ph.D.**[1], **Marie Forgeard, Ph.D.**[1,3], **Andrew Peckham, Ph.D.**[1], **Thröstur Björgvinsson, Ph.D., ABPP**[1]

[1]Harvard Medical School - McLean Hospital, Boston, MA

[2]Department of Psychology, University of California, Los Angeles, CA

[3]Department of Clinical Psychology, William James College, Newton, MA

Abstract

Objective: Research on predictors of treatment outcome in depression has largely derived from randomized clinical trials involving strict standardization of treatments, stringent patient exclusion criteria and careful selection and supervision of study clinicians. The extent to which findings from such studies generalize to naturalistic psychiatric settings is unclear. This study sought to predict depression outcomes for patients seeking treatment within an intensive psychiatric hospital setting, and while comparing the performance of a range of machine learning approaches.

Method: Depressed patients (N = 484; ages 18–72; 89.0% White) receiving treatment within a psychiatric partial hospital program delivering pharmacotherapy and cognitive behavioral therapy were split into a training sample and holdout sample. First, within the training sample, 51 pre-treatment variables were submitted to 13 machine learning algorithms to predict, via cross-validation, post-treatment Patient Health Questionnaire-9 depression scores. Second, the best performing modeling approach (lowest mean squared error; MSE) from the training sample was selected to predict outcome in the holdout sample.

Results: The best performing model in the training sample was elastic net regularization (ENR; MSE= 20.49, R^2 =.28), which had comparable performance in the holdout sample (MSE= 11.26, R^2 = .38). Fourteen pre-treatment variables predicted outcome. To demonstrate the translation of an ENR model to personalized prediction of treatment outcome, a patient-specific prognosis calculator is presented.

Conclusions: Informed by pre-treatment patient characteristics, such predictive models could be used to communicate prognosis to clinicians and to guide treatment planning. Identified predictors of poor prognosis may suggest important targets for intervention.

Keywords

depression; psychiatric hospital; machine learning; personalized prediction

Corresponding author: Christian A. Webb, Ph.D., Harvard Medical School & McLean Hospital, 115 Mill Street, Belmont, MA 02478., cwebb@mclean.harvard.edu, Phone: 617-855-4429, Fax: 617-855-4231.

Introduction

Research on predictors of outcome in depression treatment has largely relied on randomized clinical trials (RCTs; or single-arm clinical trials) with outpatient samples. These studies typically involve relatively stringent patient inclusion/exclusion criteria (e.g., excluding acute suicidality and various comorbidities), strict standardization of treatment procedures (e.g., manualized treatment protocol), and careful selection, training and supervision of study therapists/clinicians. Over 80% of individuals with depression are often excluded from depression trials due to failure to meet one or more inclusion criteria (Keitner, Posternak, & Ryan, 2003; Lorenzo-Luaces, Zimmerman, & Cuijpers, 2018; Zetin & Hoepner, 2007; Zimmerman, Mattia, & Posternak, 2002). Accordingly, it is unclear to what extent findings from these trials generalize to the vast majority of depressed individuals receiving treatment in naturalistic psychiatric settings.

Some of the abovementioned common characteristics of RCTs (e.g., standardized treatment, highly experienced therapists, excluding higher severity and suicidal patients) may help account for the greater variance in psychotherapy outcomes observed in naturalistic treatment settings relative to clinical trials (Beard, Stein, et al., 2016; McEvoy & Nathan, 2007; Lutz, Schiefele, Wucherpfennig, Rubel, & Stulz, 2016). In addition to greater variance in outcomes, several studies have reported more modest treatment effect sizes in psychotherapy delivered in naturalistic settings than in RCTs (Gibbons et al., 2010; Hans & Hiller, 2013; Hansen, Lambert, & Forman, 2002; Weisz, Weiss, & Donenberg, 1992; McEvoy & Nathan, 2007; but see Merrill et al, 2003; Lutz et al., 2016). Given evidence of both greater variance and poorer overall outcomes in naturalistic treatments, it is important to identify pre-treatment patient characteristics that predict relatively poor prognosis in these settings. Knowledge of which patients are likely to exhibit a poor outcome may have important clinical implications regarding treatment recommendations (e.g., a more intensive, alternative or combination treatment) and can inform more careful symptom and treatment progress monitoring (Delgadillo et al., 2018). In addition, the particular baseline patient characteristics predicting poorer prognosis may suggest important treatment targets. For example, if deficits in cognitive functioning were found to predict particularly poor depression outcomes, clinicians could consider antidepressants with pro-cognitive effects (e.g., Vortioxetine) for depressed individuals entering the clinic with marked cognitive impairments (Mahableshwarkar, Zajecka, Jacobson, Chen, & Keefe, 2015). Similarly, psychotherapeutic interventions could be selectively deployed to target other relevant pre-treatment predictors of poor outcome (e.g., sleep hygiene for insomnia, behavioral activation for anhedonia). Furthermore, the differences in outcomes, patient characteristics and treatment delivery in naturalistic settings relative to clinical trials raise questions regarding whether predictors of poor prognosis identified from trial data generalize to real-world treatment settings.

With regards to clinical trials of depression treatment, a range of clinical (e.g., higher baseline depression severity, chronic depression, higher anhedonia, earlier age of depression onset, anxiety disorders/symptoms, childhood maltreatment/trauma) and demographic (e.g., older age, less than full employment, being unmarried, less education) variables have been shown to predict worse depression outcomes (Kessler et al., 2017). Other studies have

examined predictors of outcome in naturalistic psychiatric settings (see Bohart & Wade, 2013). The majority of these studies have evaluated predictors of depression outcome in settings such as outpatient individual or group psychotherapy (Enns, Cox, & Pidlubny, 2002; Kuyken, 2004; Persons, Burns, & Perloff, 1988; Schindler, Hiller, & Witthöft, 2013; Thimm & Antonsen, 2014) or primary care (Cui, Lyness, Tang, Tu, & Conwell, 2008; Prins et al., 2011; Ronalds, Creed, Stone, Webb, & Tomenson, 1997), and have identified many of the same baseline predictors as those reported in clinical trials (e.g., higher baseline symptom severity, demographic variables).

In contrast, the literature is especially sparse with regards to predictors of outcome in intensive psychiatric settings (i.e., partial hospital programs, day programs, inpatient treatment), which patients rely on when standard outpatient treatment fails. Despite the fact that these programs deliver care at a critical juncture (e.g., when patients are either stepping up from outpatient care, or stepping down from inpatient units), very few studies have investigated pre-treatment predictors of response for depressed patients in these commonly used programs (Beard, Stein, et al., 2016; Riedel et al., 2011; Zeeck et al., 2016; also see Drymalski & Washburn, 2011). Beard and colleagues (2016) found that greater expectations of symptom improvement and fewer past hospitalizations predicted better depression outcomes in a partial hospitalization program. To increase generalizability, the only inclusion criteria in the latter study was entering treatment with clinically significant depressive symptoms (Center for the Epidemiological Studies of Depression-10 [CESD-10] Total > 10). There were no exclusion or inclusion criteria regarding diagnostic comorbidities. In one other partial hospitalization study, greater comorbidity and lower baseline motivation predicted worse depression outcomes (Zeeck et al., 2016).

To address a gap in literature, the present study sought to predict depression outcomes for patients seeking treatment within an intensive psychiatric hospital setting, and while comparing the performance of a range of machine learning approaches, which are well-suited to handling large numbers of predictor variables (in the present study, measures assessing pre-treatment patient characteristics). The bulk of the abovementioned treatment outcome prediction studies tested whether individual variables predicted depression outcome. However, any single predictor will only account for limited variance in depression outcome. Multivariable machine learning approaches allow for the combination of predictors in an effort to account for maximal variance in outcome. In addition, many machine learning approaches (e.g., random forest) can accommodate non-linear associations and higher-order interactions that may be present in the data. Finally, as a demonstration of how a predictive model could be applied to new patients entering treatment, we developed a personalized (i.e., patient-specific) “treatment outcome prognosis calculator” informed by our final model to predict estimated depression severity at hospital discharge. A prognosis calculator could have the following clinical benefits: (1) inform risk-stratification and identify patients requiring a higher level of care (Delgadillo, Huey, Bennett, & McMillan, 2017; Lorenzo-Luaces, DeRubeis, van Straten, & Tiemens, 2017), (2) identify the subset of individuals who are most likely to benefit from careful outcome monitoring (see Carlier et al., 2012; Knaup, Koesters, Schoefer, Becker, & Puschner, 2009; Shimokawa, Lambert, & Smart, 2010; Kendrick et al., 2016; Delgadillo et al., 2018), (3) highlight which specific baseline characteristics are contributing to an individual’s predicted poor prognosis, which could

guide treatment targets in an effort to improve outcome and (4) result in a more efficient and targeted allocation of limited program resources to the subset of individuals most likely to benefit from careful monitoring and clinical attention.

Method

Study Design

Participants were recruited from a behavioral health partial hospital program. To increase generalizability, no exclusion criteria were applied (e.g., all diagnostic comorbidities were allowable, and no patients were excluded on the basis of symptom severity or suicidality). We implemented a three-step modeling approach to minimize overfitting (i.e., to increase generalizability to new sets of patients) (Steps 1–2) and to yield a model applicable to *individual* patients entering partial hospital treatment (Step 3). First, within a training dataset, we constructed prognostic models based on a set of 51 clinical, demographic, and physical health variables (using 10-folds cross-validation to minimize overfitting). Informed by recent recommendations (LeDell, van der Laan, & Petersen, 2016; Rose, 2013) and empirical findings (Rosellini, Dussaillant, Zubizarreta, Kessler, & Rose, 2018), we compared the predictive performance (via mean squared error; MSE) of 13 machine learning algorithms, in addition to conventional linear regression. For example, elastic net regularization (ENR; Friedman, Hastie, & Tibshirani, 2010) is a commonly used variant of conventional regression combining the ridge and lasso penalizations to constrain coefficients among collinear variables and to minimize model overfitting. In contrast, random forests (RF; Breiman, 2001) and Bayesian additive regression trees (BART; Kapelner & Bleich, 2016) are decision-tree ensemble methods which - unlike ENR - can accommodate unspecified non-linear associations and higher-order interactions (see Algorithms below for a list of the approaches used). In sum, given that these approaches rely on different algorithms for the selection and weighting of variables, it is unclear which of these methods would yield the most accurate predictions. Thus, comparing a range of different prediction models within a training sample allowed us to select the optimal approach.

In Step 2, we selected the best performing modeling approach from the training set and applied it to a non-overlapping, fully held-out sample. Finally, as a demonstration of how a predictive model could be applied to new patients entering treatment, we developed a personalized “treatment outcome prognosis calculator” informed by our final model to predict estimated depression severity at program discharge.

Participants

The purpose of the present study is to examine predictors of treatment outcome among depressed patients receiving treatment in a naturalistic psychiatric setting delivering intensive evidence-based treatment. To increase generalizability, the only diagnostic inclusion criteria was for patients to meet DSM-IV criteria for a current episode of MDD. All comorbidities were allowed. Participants who completed fewer than three days of treatment were excluded. For the full sample (n = 484), mean pre-treatment Patient Health Questionnaire - 9 (PHQ-9; Kroenke & Spitzer, 2002) depression scores were in the “moderately severe” range (M = 17.18; SD = 4.88), with 79.1% reporting a recurrent course

of depression (mean number of previous reported depressive episodes was 5.25; *SD* = 3.85). Diagnostic comorbidity with one or more anxiety disorder was common (71.1%) in this sample, as typically seen in real-world psychiatric hospital settings (Table 1). Twenty-six percent of the sample had scores on the McLean Screening Instrument for Borderline Personality Disorder (MSI-BPD)(Zanarini et al., 2003) above the cutoff (total score > 7) suggesting a Borderline Personality Disorder (BPD) diagnosis. At admission to the program, 70% of the sample endorsed some suicidal ideation (PHQ-9 item 9, "thoughts that you would be better off dead, or of hurting yourself'). Over half (54.1%) of the sample had previously been hospitalized (inpatient) at least once for their psychiatric problems (35.5% received inpatient psychiatric treatment in the week prior to admission). On average, patients were on 2.1 (*SD* = 1.08) psychiatric medications at baseline.

Procedure

Treatment.—The present dataset includes patients who received treatment in a Behavioral Health Partial Program (BHP) at McLean Hospital from September 2015 to March 2017. Patient's attended daily (Monday-Friday) treatment sessions consisting of approximately 5 hours of clinical services each day, including CBT-based group and individual therapy, as well as pharmacotherapy. The average length of treatment (inclusive of weekends and holidays) was 12.97 days (*SD* = 4.19). During this intensive treatment period, patients met with a case manager (psychologist or social worker) several times per week who set treatment goals, determined the schedule of groups that a patient attends, and planned for aftercare. The case manager worked closely with a program psychiatrist who met individually with the patient to provide pharmacological treatment. All patients also received individual psychotherapy from a program therapist (psychologist, clinical psychology intern, postdoctoral fellow, or practicum student), who worked with the patient to practice and reinforce CBT-based skills learned in the group therapy program. The program therapist also administered a structured diagnostic interview for assessment purposes. All clinical team members met weekly to review a patient's symptom levels reported in progress monitoring questionnaires, assess treatment goals, and discuss plans for aftercare.

The majority of treatment at the BHP is delivered via CBT groups. All patients received psychoeducation about the cognitive behavioral model, and each day there was at least one group focused on identifying and challenging negative automatic thoughts, grounded in concepts originally developed by Beck and colleagues (Beck, Rush, Shaw, & Emery, 1979). Patients also attended behavioral activation (BA) skills groups throughout the week, modeled after protocols for BA for depression developed by Martell and colleagues (Martell, Dimidjian, & Herman-Dunn, 2013). Finally, patients had the opportunity to attend groups that provide more in-depth exposure to other CBT strategies such as identifying core beliefs (offered once per week), creating CBT diary cards to track thoughts and behaviors (once per week), and CBT skills for insomnia (twice per week). In addition to these core skills, group treatment also included exposure to third-wave cognitive-behavioral treatments. These included two groups per week introducing each core skill area of dialectical behavior therapy (mindfulness, emotion regulation, distress tolerance, and interpersonal effectiveness) (Linehan, 2014). Several acceptance and commitment therapy (ACT) groups were also offered on a weekly basis (for additional details, see Supplement).

Assessment of Symptoms and Diagnoses.—As part of clinical treatment at the BHP, all patients received a structured diagnostic interview administered by their program therapist on the day following admission (see Diagnostic Measure below). On their day of admission, patients completed a battery of self-report measures as part of their clinical assessment, including measures of symptoms and functioning. After the initial assessment at admission, patients completed a daily battery of questionnaires each morning before the start of treatment (8:30– 9am), including questionnaires on their day of discharge. All self-report measures were collected using REDCap, a computerized, secure, web-based platform for administering surveys (Harris et al., 2009). The local Institutional Review Board approved all study procedures.

Measures

Diagnostic Measure.—The Mini-International Neuropsychiatric Interview (MINI) (Sheehan et al., 1998) was used to diagnoses Axis I disorders from the Diagnostic and Statistical Manual of Mental Disorders-Fourth Edition (DSM-IV-TR). Diagnostic categories assessed include major depressive episodes, manic/hypomanic episodes, panic disorder, obsessive compulsive disorder, generalized anxiety disorder, social anxiety disorder, PTSD, alcohol abuse and dependence, and psychosis. The MINI can be administered in less than one hour, but nonetheless has shown good validity in comparison to other more detailed diagnostic interview instruments (Sheehan et al., 1998). In the present study, MINI interviews were conducted by program therapists on the day following admission. Program therapists completed extensive didactic training, role-plays, and observations of structured interviews prior to administering this measure.

Outcome Variable.—The Patient Health Questionnaire - 9 (PHQ-9)(Kroenke & Spitzer, 2002) is a brief self-report measure assessing symptoms of depression over the past two weeks. The PHQ-9 has demonstrated good reliability and validity in previous validation studies as a screening tool for MDD (Kroenke & Spitzer, 2002; Löwe, Kroenke, Herzog, & Gräfe, 2004). In addition, the PHQ-9 has been established as a valid measure in partial hospital settings (Beard, Hsu, Rifkin, Busch, & Björgvinsson, 2016). In the present study, the PHQ-9 showed adequate internal consistency at admission ($\alpha = .77$).

Predictor Variables.—Baseline variables included clinical measures, demographic characteristics, MDD history, comorbid diagnoses, treatment history, current psychiatric medication use, and physical health variables (For the full list of baseline variables, see Table 2 and Supplemental Methods).

Statistical Analyses

Train-Test Sample Splitting.—As the goal is to ultimately apply such a prognostic model to *new* sets of patients entering treatment, the held-out validation or test sample consisted of the most recent 20% of patients who received treatment in the BHP (i.e., of the full sample of BHP patients [$n = 484$], the 20% who most recently entered treatment were assigned to the holdout sample [$n = 97$], whereas the training sample consisted of the remaining 80% [n = 387] who received treatment prior to the patients in the holdout

sample). Importantly, the holdout sample was not used in any way during data preparation, modeling approach selection or model development.

Missing data.—Missing data imputation was conducted with the missForest (Stekhoven & Bühlmann, 2012) package in R (R Core Team, 2013), which implements a non-parametric RF-based imputation. Of the 51 baseline predictors included our models, rates of missing data were as follows: 45 variables had less than 5% missingness, 1 variable had between 5–10% missingness, and the remaining 5 variables had between 10%−18% missingness. For post-treatment PHQ-9 scores, 16.7% of participants had missing values. Missing data were imputed separately for the training sample and holdout sample (see Supplemental Methods).

Algorithms.—We implemented and compared a diverse range of machine learning approaches. Specifically, and following the approach used by Rosselini et al. (2018), we tested: (a) five penalized regressions: ENR with mixing parameter alpha (α) set to either .25, .50 or .75, as well as ridge and lasso (least absolute shrinkage and selection operator) regression (glmnet package; Friedman et al., 2010), (b) two decision-tree based algorithms: RF (randomForest package; Breiman, 2001) and BART (bartMachine package; Kapelner & Bleich, 2016), (c) three support vector regressions (SVR; kernel parameter = linear, polynomial and radial)(e1071 package; Steinwart & Christmann, 2008) and (d) two spline regressions: adaptive splines (Friedman, 1991) and adaptive polynomial splines (Stone, Hansen, Kooperberg, & Truong, 1997) (earth and polspline packages). For the purpose of comparison, we also applied conventional ordinary least squares (OLS) linear regression. Finally, super learner, a machine learning ensembling method, was implemented (SuperLearner package; Polley, LeDell, & van der Laan, 2016). Super learning assigns weights to a set of selected algorithms (in this study, the above 13 models) to develop a consolidated predictive algorithm which optimizes cross-validated MSE. See Supplemental Table 1 for details on the implementation of each algorithm.

Sample 1: Training Set.—See Figure 1 for a schematic of the analysis pipeline. In an effort to minimize overfitting and identify a model that would best generalize to a holdout sample, we ran each of the 14 algorithms within 10-folds cross-validation, such that the training dataset (n = 387) was split into ten equal size samples (Kuhn & Johnson, 2013). For each of the ten folds, models were trained on $9/10^{th}$ of the data (from the other 9 folds) and post-treatment PHQ-9 predictions were generated for that held out fold (which comprised the other $1/10^{th}$ of the training sample). Importantly, the cross-validation procedure ensures that the predictions of post-treatment PHQ-9 scores for all patients are generated from models that are constructed without the use of their own data. Model performance was compared via MSE. The 10-folds cross validation procedure was repeated 100 times. The model with the lowest MSE (i.e., mean MSE across 100 replications of 10-folds cross-validation) was selected and evaluated in the held-out validation sample. R^2 (coefficient of determination) values are reported based on the following formula: 1 - (MSE/var(y)). See supplement for r^2 values (with 95% confidence intervals) reflecting the squared correlation between predicted and observed discharge PHQ-9 scores.

Sample 2: Holdout Set.—The best performing model (i.e., lowest MSE) from the training sample was applied to the patients in the holdout sample to predict their post-treatment PHQ-9 scores. Critically, patients in the holdout set are from a non-overlapping sample that did not contribute in any way to the selection of the optimal approach or the construction of the final model used to predict their post-treatment depression scores.

Development of a Personalized Treatment Outcome Prognosis Calculator.—To demonstrate the translation of a final model to personalized prediction, we used the variables selected from the final model, and their associated parameter estimates, to program (using the Shiny R package; Chang et al., 2017) a web-based calculator predicting post-treatment PHQ-9 scores for new patients.

Results

In the full sample, depression outcomes varied widely (post-treatment PHQ-9 range = 027; M = 11.59; SD = 5.15), with 7.2% of patients completing treatment with depression scores in the “minimal” range (PHQ-9 Total = 0–4), 30.3% with “mild depression” (PHQ-9 Total = 5–9), 32.4% with “moderate depression” (PHQ-9 Total = 10–14), 23.6% with “moderately severe depression” (PHQ-9 Total = 15–19), and 6.4% with “severe depression” (PHQ-9 Total = 20–27) (see Supplemental Figures 1 and 2 for distributions of pre-treatment and post-treatment PHQ-9 scores). Thus, there was substantial variance in our dependent variable which we sought to predict based on pre-treatment patient characteristics

Training Sample.

The mean of post-treatment PHQ-9 scores in the training sample was 11.54 (SD = 5.35). Results of the 10-folds cross-validation procedure indicated that ENR (alpha=0.75) was associated with the lowest prediction error (MSE = 20.49; MAE = 3.78; R^2 = .28; see Table 3 & Figure 2). On average, the predicted post-treatment PHQ-9 scores differed from observed values by 3.78 points. Next, ENR was run on the full training sample (rather than successive 90/10 cross-validation splits) to determine a final model for implementation in the holdout sample. Tuning of ENR’s alpha and lambda parameters was performed in the training sample using the CARET package’s resampling grid search (Kuhn, 2008). Each combination of alpha (from 0 to 1 by 0.05) and lambda (from 0 to 1 by 0.05) was tested, and the optimal values were selected: alpha=0.90 and lambda=0.30 (via minimum cross-validated error criterion). See Table 4 for variables retained in the ENR model and their associated parameter estimates. Of the 51 baseline variables submitted to our final ENR model, 14 emerged as predictors of worse post-treatment prognosis: higher depression and anxiety severity, greater fatigue and difficulties concentrating, heightened BPD symptoms, more relationship problems, more pessimistic expectations of symptom improvement, prior treatment at an intensive outpatient program or partial hospital program, identifying as White, earlier age of MDD onset, comorbid diagnoses of OCD, PTSD or SAD, and a mood stabilizer prescription.

Holdout Sample.

The mean of post-treatment PHQ-9 scores in the holdout sample was 11.79 (SD = 4.27).The final ENR model (i.e., with alpha=0.90 and lambda=0.30) from the training sample was implemented in the holdout sample to predict post-treatment PHQ-9 scores in a non-overlapping group of patients (*MSE* = 11.26; *MAE* = 2.65; and R^2 = .38). On average, the predicted post-treatment PHQ-9 scores differed from observed values by 2.65 points (see Supplement Results for additional analyses). For comparison, the above model outperformed a linear regression with the sole predictor being baseline PHQ-9 (i.e., the baseline variable with the strongest correlation with post-treatment PHQ-9) (*MSE* = 12.77; *MAE* = 2.80; and R^2 = .30).

Development of a Personalized Treatment Outcome Prognosis Calculator.

A calculator generating predicted post-treatment PHQ-9 scores was developed on the basis of the baseline predictors retained in the final ENR model, and their associated parameter estimates. A user inputs the observed values for each variable for a new patient and, using the ENR model, the calculator generates predicted post-treatment PHQ-9 scores, overlaid on the distribution of scores for the full BHP sample (Figure 3, top panel). The bottom panel of Figure 3 illustrates the contribution of each variable. Features that increase risk of poor prognosis - relative to the BHP sample mean are colored red; whereas those that decrease risk are in blue.

Discussion

The present study sought to predict treatment outcomes for depressed patients in an intensive psychiatric treatment setting (i.e., partial hospital program). As expected, given the psychiatric hospital setting, the sample was characterized by elevated depression severity (mean intake PHQ-9 = 17.2; mean number of previous depressive episodes 5.3; 70% with suicidal ideation; 54% previously received inpatient psychiatric treatment) and comorbidities (71% with a comorbid anxiety disorder). Strengths of the study include: (1) comparison of a diverse array of machine learning approaches, which rely on different algorithms to generate predicted outcomes (e.g., variants of classical regression such as ENR vs. tree-ensemble approaches such as RF and BART), (2) statistical methods to minimize model overfitting and increase generalizability through the use of a training sample, with 10-folds cross-validation, and a non-overlapping holdout sample, and (3) the use of a naturalistic psychiatric sample with minimal inclusion/exclusion criteria to increase external validity to patients receiving treatment in intensive psychiatric hospital settings.

Consistent with prior work indicating high levels of variability in treatment outcome in naturalistic settings relative to RCTs (Beard, Stein, et al., 2016; McEvoy & Nathan, 2007; Lutz et al., 2016), depression outcomes varied widely in our sample and covered the full range of the PHQ-9 (0–27), providing ample variance for prediction. Of the 14 models tested, among the simplest of the machine learning approaches (ENR) yielded the best cross-validated performance in the training sample. Relative to other machine learning methods, ENR is less complex and more easily interpretable than “black box” approaches such as RF, which may include unspecified high-order interactions and non-linear associations

underlying the predicted outcomes (Song, Langfelder, & Horvath, 2013). It is important to highlight that the nature of the data and relations between variables will determine which machine learning approach yields the best performance (e.g., RF is likely to outperform ENR in the presence of unspecified non-linear associations and interactions). The fact that ENR is a relatively simple variant of conventional linear regression makes it ideal for implementation in a clinic to generate predicted outcomes for new patients (as demonstrated in Figure 3, top panel) (Delgadillo et al., 2017). Given the ease of interpretability of ENR coefficients, the presented prognosis calculator also quantifies - for each individual patient - which pre-treatment variable increases vs. decreases risk of poor outcome (Figure 3, bottom panel). The best performing model in the training sample, ENR, was then applied to the holdout sample, where the R^2 was .38. The mean absolute difference between predicted and observed PHQ-9 outcome scores was less than 3 points (2.65) on a 27-point scale. The performance of ENR in the training sample and in the holdout sample indicate that this model could provide clinicians with reasonably accurate predictions of a patient's depression status at the time of discharge on the basis of pre-treatment clinical and demographic characteristics.

There are several potential benefits of integrating a prognosis calculator into a clinical unit, such as the present partial hospital program. First, prior research in outpatient samples (Carlier et al., 2012; Knaup et al., 2009; Shimokawa et al., 2010; Kendrick et al., 2016), including a recently published large multi-site psychotherapy trial for depression and anxiety (Delgadillo et al., 2018), indicates that providing clinicians with ongoing feedback on their patients' symptom progress, and generating a "risk signal" when patients are relatively off-track, can improve treatment outcomes, *especially* for patients at risk of a poor outcome. For example, in a multisite, open-label, cluster randomized controlled psychotherapy trial for depression and anxiety, Delgadillo et al. (2018) randomly assigned therapists to an outcome feedback condition or a treatment-as-usual control group. Although there were no overall between-group differences in outcome, the subgroup of individuals at risk of a poor outcome, by virtue of symptom trajectories suggesting relatively little improvement, had significantly better depression and anxiety outcomes in the feedback condition relative to the control group. These important findings highlight that a relatively simple, low-cost and scalable procedure (i.e., patient progress monitoring) can enhance outcomes for patients at risk of a poor prognosis. Thus, our prognosis calculator could be used to identify, and subsequently closely monitor symptom progress for, patients predicted to have a poor prognosis, which may improve outcomes. Of course, this requires the integration of symptom monitoring and a system for providing feedback to clinicians (e.g., secure, web-based outcome monitoring chart). Second, prognosis calculators can inform risk-stratification and treatment recommendations. Specifically, patients predicted to be at high risk of poor outcome based on pre-treatment characteristics may be better-suited to a higher level of care. Indeed, two recent studies in outpatient samples found that patients predicted (via penalized regression) to have a poor prognosis had better outcomes if initially assigned to a relatively high-intensity treatment protocol relative to low intensity treatment (Delgadillo et al., 2017; Lorenzo-Luaces et al., 2017). In contrast to treatment matching based on statistically-based prognostic indices, treatment selection based on clinical judgment has proved less successful (Van Straten, Tiemens, Hakkaart, Nolen, & Donker,

2006). This is consistent with findings indicating that clinician's prognostic judgment of patient outcomes tend to be inaccurate (ĝisdottir et al., 2006; Grove & Meehl, 1996), often failing to identify patients who exhibit poor treatment outcomes (Hannan et al., 2005). The present study was conducted in a sample receiving treatment in a higher level of care (i.e., partial hospital program) than the Delgadillo et al. (2017) or Lorenzo-Luaces et al. (2017) studies. However, patients predicted to have a particular poor prognosis on the basis of our prognostic model may be better suited to a higher level of care such as a course of inpatient treatment before stepping down to a lower level of psychiatric care (e.g., partial hospital program). Ultimately, research is needed to test whether such patients do in fact have better outcomes in an inpatient setting than a partial hospital program. Third, in addition to generating predicted post-treatment depression severity scores, our model also provides a patient-specific variable contribution plot which specifies which baseline characteristics are contributing to the individual's predicted poor prognosis, which could guide treatment targets in an effort to improve outcome. Two individuals may have the same predicted post-treatment outcome but differ substantially in their values on the baseline variables contributing to their prognoses. For the example patient presented in Figure 3, a treatment team could use such pre-treatment information to target the most relevant identified risk factors via psychotherapy (e.g., assign patient to DBT groups for heightened BPD symptoms, target negative cognitions underlying pessimistic treatment outcome expectancies, sleep hygiene for fatigue due to insomnia). Targeted pharmacological interventions could also be considered (e.g., an antidepressant medication with pro-cognitive effects, such as Vortioxetine (Mahableshwarkar et al., 2015), for heightened concentration problems or Bupropion for fatigue (Papakostas et al., 2006)). Finally, and related to each of the above points, clinical resources are limited. A prognosis calculator could be used for more efficient and targeted allocation of limited resources to the subset of individuals most likely to benefit from careful monitoring and clinical attention.

Several of the predictors (i.e., higher depression and anxiety severity, greater fatigue, heightened BPD symptoms, more relationship problems, earlier age of MDD onset, comorbid diagnoses of OCD, PTSD or SAD) that emerged in this study are consistent with findings from a recent review of predictors of treatment outcome in depression (Kessler et al., 2017). Many of these studies were testing whether individual variables predict outcome. In the present study, we used a multivariable approach to build our model, as any single variable will only account for limited variance in outcome. It is noteworthy that of all depressive symptoms relatively poor concentration and fatigue were the two that emerged as predictors of poor prognosis. These two symptoms predicted poorer prognosis above and beyond the contribution of total depression severity, highlighting the importance of considering individual depressive symptoms in predictive efforts (Fried & Nesse, 2015). Fatigue and concentration difficulties are recognized as among the most challenging symptoms to treat and most common residual symptoms (Nierenberg et al., 2010; Targum & Fava, 2011). Moreover, concentration and fatigue are among the symptoms most strongly linked with impairment in psychosocial functioning (Fried & Nesse, 2014). In addition, patients with several anxiety comorbidities (OCD, OCD, PTSD), BPD symptomatology, and interpersonal difficulties had poorer outcomes, which suggest potentially important

treatment targets for those exhibiting elevations in these symptoms on their variable contribution plots.

A linear regression analysis where the only predictor was baseline PHQ-9, performed surprisingly well (*MSE* = 12.77; *MAE* = 2.80). The performance of the baseline depression model is due to the relatively strong correlation between baseline and discharge PHQ-9 scores in this partial hospital setting (r = .53, for the full BHP sample), likely attributable to the short-term treatment stay (mean = 13 days) and severity of these hospitalized patients relative to most treatment outcome prediction studies which rely on outpatient samples. In other words, baseline depression accounts for 28% (r^2 = .28) of the variance in depression outcome. This is compared to RCTs of outpatient treatment where the r^2 is substantially lower, such as in the large STAR*D (Rush et al., 2004) and GENDEP trials (Uher et al., 2009; Uher, Tansey, Malki, & Perlis, 2012) (mean r^2 = .16 across both trials) and the recently completed EMBARC trial (r^2 = .09) (Trivedi et al., 2016; Webb et al., 2018). These findings highlight that treatment-related factors (e.g., length of treatment) and patient characteristics (e.g., baseline severity or chronicity of depression) may have a meaningful influence on the relative predictive strength of baseline patient characteristics such as total depression score. It is also important to note that the difference in variance accounted for between the elastic net model and the PHQ-9 linear regression model (r^2 of .39 - .31 = 8%, see Supplement) exceeds a clinically significant threshold for a predictor of depression outcome (r^2 > 6.3% in the simulation study by Uher et al., 2012).

As clinical needs vary across different treatment contexts, it is important for research to be guided by what outcomes are most relevant to clinicians for their patient population and setting (e.g., predicting functional impairment at discharge, length of treatment stay, or risk of suicidal behavior on an inpatient unit or following hospital discharge). Future studies may also benefit from directly comparing the accuracy of machine learning models relative to predictions generated by clinicians themselves. Are machine learning algorithms more accurate than clinician predictions and, if so, does presenting this information give clinicians more “buy-in” for integrating algorithms into their routine intake assessments and treatment planning? Finally, the present modeling approach, and resulting risk calculator, focused on “prognostic” rather than “prescriptive” (i.e., treatment group by predictor variable interactions) predictors of outcome, the latter of which are more informative for treatment selection (Cohen & DeRubeis, 2018; Webb et al., 2018). The fact that virtually all patients received both pharmacological treatment and CBT (i.e., 97% received pharmacological treatment and all patients received both individual and group CBT) precluded the testing of prescriptive predictors. Although combined pharmacological and behavioral treatment is typical in naturalistic settings, future studies might examine predictors of outcome in settings in which patients are assigned to *separate* treatments (e.g., antidepressant medications *vs.* CBT) in order to test for prescriptive predictors.

Several limitations of the present study should be noted. With regards to the generalizability of our findings, most of the predictors that were retained in our final elastic net model have also been shown to predict outcome in outpatient samples (Kessler et al., 2017). Nevertheless, the findings emerged within a specific treatment context: a partial hospital day program providing CBT and pharmacotherapy. Accordingly, we would expect our findings

to be more likely to generalize to similar partial hospital programs than to outpatient therapy or inpatient units. The current partial hospital program likely administers more high-quality assessments than other settings given the clinical measurement initiatives at the hospital level and the clinical research program embedded in this unit. However, many other health care settings now routinely administer brief measures, such as the PHQ-9, GAD-7, BASIS-24, and quality of life measures, including primary care and other partial hospital programs (Drymalski & Washburn, 2011). Hospital settings and providers treating individuals with Medicare/Medicaid are required to use evidence-based assessments. Regarding the average treatment length, the partial hospital program in the current study is similar in structure and length to other partial hospital programs across the United States (see AABH definition ("Partial Hospital Programs," 2019) & descriptions in Drymalski & Washburn 2011)). Treatment length is also similar to the average length of inpatient psychiatric hospitalization (e.g., 10 days) (Lee, Rothbard, & Noll, 2012). It is also important to note that our inclusion criteria requiring a diagnosis of MDD excluded patients experiencing depressive symptoms but who did not meet diagnostic threshold for major depression (e.g., those with dysthymic disorder). In summary, research is needed to test the extent to which the present findings extend to other treatment settings, as this is the first study, to our knowledge, to use machine learning approaches to generate a patient-specific prognosis calculator for depressed patients treated in a naturalistic psychiatric unit. Importantly, a primary impetus for the present study was that the existing literature examining predictors of treatment outcome may not generalize to more acute settings (inpatient, residential, partial hospital units). These treatment settings represent important and highly utilized - yet substantially understudied - levels of clinical care focused on more severe and chronic psychiatric illness. There are over 400 partial hospital programs in the United States alone, yet these programs have received limited empirical attention (Forgeard, Beard, Kirakosian, & Bjorgvinsson, 2018). Given the use of both a cross-validated training sample and a holdout sample, as well very limited inclusion/exclusion criteria, we expect our findings may be generalizable to many of these programs. Second, we focused primarily on self-report predictor variables given that they could be reasonably integrated into a naturalistic psychiatric clinic. However, it is unclear to what extent behavioral (e.g., blunted reward learning or deficits in cognitive control (Webb et al., 2018)) or neuroimaging (e.g., rostral anterior cingulate cortex activity (Pizzagalli et al., 2018)) variables may provide incremental predictive validity and account for meaningfully greater variance in outcomes. Third, measurement error in both predictor and outcome variables and "population drift" (e.g., change in hospital staff or treatment protocol) can substantially influence model performance and attenuate predictor-outcome associations (Hand, 2006; Uher et al., 2012). Fourth, although the comparison of the predictive performance of elastic net vs. OLS multiple regression yielded meaningful differences in the training sample (R^2 of .28 vs. .14, respectively), the latter comparison in the hold-out yielded a small difference in R^2 (.38 vs. . 34, respectively; see supplement for additional analyses). The generally lower MSE and higher R^2 values in the holdout relative to the training sample could be due to the lower variance in the outcome variable in the holdout sample. Given the inconsistency in the relative advantage of elastic net over OLS multiple regression in the training vs. holdout samples, it unclear to what extent the advantage of machine learning approaches (in the latter case, elastic net) over multiple regression is clinically meaningful. Fifth, within the

training sample, the present study compared a set of algorithms used in a recent study (Rosellini et al. 2018). The “winning” model was then tuned in the full training sample prior to implementing it in the holdout sample. An alternative approach could have been to tune each model within the 10-folds cross-validation model comparison procedure, and/or select a different set of approaches expected to perform well in this context. Sixth, given that Type I error may be elevated for elastic net models using the minimum cross-validated error criterion to select lambda values (Waldmann, Mészáros, Gredler, Fuerst, & Sölkner, 2013), not all of the predictors retained (Table 2) may be true predictors. These limitations notwithstanding, the present study demonstrates the use of machine learning in predicting treatment outcome in a naturalistic psychiatric hospital setting, with findings translated into a patient-specific prognosis calculator. Future studies are needed to examine the generalizability of such prognostic models to other common treatment contexts for depressed patients.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding

Funding for the study was provided by the Behavioral Health Partial Program and McLean Hospital. The first author (Webb) was partially supported by K23 MH108752, R01 MH116969, and a NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation. Cohen was supported in part by a grant from MQ: Transforming mental health MQDS16/72. The opinions and assertions contained in this article should not be construed as reflecting the views of the sponsors.

References

AEgisdóttir S, White MJ, Spengler PM, Maugherman AS, Anderson LA, Cook RS, … Rush JD. (2006). The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction. The Counseling Psychologist, 34(3), 341–382. 10.1177/0011000005285875

Beam AL, & Kohane IS. (2018). Big Data and Machine Learning in Health Care. JAMA, 319(13), 1317–1318. 10.1001/jama.2017.18391 [PubMed: 29532063]

Beard C, Hsu KJ, Rifkin LS, Busch AB, & Bjorgvinsson T. (2016). Validation of the PHQ-9 in a psychiatric sample. Journal of Affective Disorders, 193, 267–273. [PubMed: 26774513]

Beard C, Stein AT, Hearon BA, Lee J, Hsu KJ, & Bjorgvinsson T. (2016). Predictors of Depression Treatment Response in an Intensive CBT Partial Hospital. Journal of Clinical Psychology, 72(4), 297–310. 10.1002/jclp.22269 [PubMed: 26934333]

Beck AT, Rush JA, Shaw BF, & Emery G. (1979). Cognitive therapy of depression. New York: Guilford Press.

Bergquist SL, Brooks GA, Keating NL, Landrum MB, & Rose S. (2017). Classifying Lung Cancer Severity with Ensemble Machine Learning in Health Care Claims Data. In Doshi-Velez F, Fackler J, Kale D, Ranganath R, Wallace B, & Wiens J. (Eds.), Proceedings of Machine Learning Research (Vol. 68, pp. 25–38).

Bohart AC, & Wade AG. (2013). The client in psychotherapy. Bergin and Garfield’s Handbook of Psychotherapy and Behavior Change, 6, 219–257.

Breiman L. (2001). Random Forests. Machine Learning, 45(1), 5–32.

Buuren S, & Groothuis-Oudshoorn K. (2011). mice: Multivariate imputation by chained equations in R. Journal of Statistical Software, 45(3).

Cameron IM, Cunningham L, Crawford JR, Eagles JM, Eisen SV, Lawton K, … Hamilton RJ. (2007). Psychometric properties of the BASIS-24© (Behaviour and Symptom Identification Scale-Revised) Mental Health Outcome Measure. International Journal of Psychiatry in Clinical Practice, 11(1), 36–43. [PubMed: 24941274]

Carlier IVE, Meuldijk D, Vliet IMV, Fenema EV, Wee N. J. A. V. der, & Zitman F. (2012). Routine outcome monitoring and feedback on physical or mental health status: Evidence and theory. Journal of Evaluation in Clinical Practice, 18(1), 104–110. [PubMed: 20846319]

Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J, RStudio, … R), R. C. T. (tar implementation from. (2017). shiny: Web Application Framework for R (Version 1.0.5). Retrieved from https://CRAN.R-project.org/package=shiny

Cohen ZD, & DeRubeis RJ. (2018). Treatment Selection in Depression. Annual Review of Clinical Psychology, 14(1), 209–236. 10.1146/annurev-clinpsy-050817-084746

Cui X, Lyness JM, Tang W, Tu X, & Conwell Y. (2008). Outcomes and Predictors of Late-Life Depression Trajectories in Older Primary Care Patients. The American Journal of Geriatric Psychiatry, 16(5), 406–415. [PubMed: 18448851]

Delgadillo J, de Jong K, Lucock M, Lutz W, Rubel J, Gilbody S, … McMillan D. (2018). Feedback-informed treatment versus usual psychological treatment for depression and anxiety: A multisite, open-label, cluster randomised controlled trial. The Lancet. Psychiatry, 5(7), 564–572. 10.1016/S2215-0366(18)30162-7 [PubMed: 29937396]

Delgadillo J, Huey D, Bennett H, & McMillan D. (2017). Case complexity as a guide for psychological treatment selection. Journal of Consulting and Clinical Psychology, 85(9), 835–853. 10.1037/ccp0000231 [PubMed: 28857592]

Devilly GJ, & Borkovec TD. (2000). Psychometric properties of the credibility/expectancy questionnaire. Journal of Behavior Therapy and Experimental Psychiatry, 31(2), 73–86. [PubMed: 11132119]

Drymalski WM, & Washburn JJ. (2011). Sudden gains in the treatment of depression in a partial hospitalization program. Journal of Consulting and Clinical Psychology, 79(3), 364–368. 10.1037/a0022973 [PubMed: 21381809]

Enns MW, Cox BJ, & Pidlubny SR. (2002). Group Cognitive Behaviour Therapy for Residual Depression: Effectiveness and Predictors of Response. Cognitive Behaviour Therapy, 31(1), 31–40. 10.1080/16506070252823643

Forgeard M, Beard C, Kirakosian N, & Bjorgvinsson T. (2018). Research in Partial Hospital Settings. In Practice-Based Research: A Guide for Clinicians (1st ed., Vol. 1, pp. 212–240).

Fried EI, & Nesse RM. (2014). The Impact of Individual Depressive Symptoms on Impairment of Psychosocial Functioning. PLOS ONE, 9(2), e90311.

Fried EI, & Nesse RM. (2015). Depression sum-scores don't add up: Why analyzing specific depression symptoms is essential. BMC Medicine, 13, 72 10.1186/s12916-015-0325-4 [PubMed: 25879936]

Friedman JH. (1991). Multivariate Adaptive Regression Splines. The Annals of Statistics, 19(1), 1–67.

Friedman JH, Hastie T, & Tibshirani R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1–22. [PubMed: 20808728]

Gibbons CJ, Fournier JC, Stirman SW, DeRubeis RJ, Crits-Christoph P, & Beck AT. (2010). The clinical effectiveness of cognitive therapy for depression in an outpatient clinic. Journal of Affective Disorders, 125(1), 169–176. [PubMed: 20080305]

Grove WM, & Meehl PE. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. Psychology, Public Policy, and Law, 293–323.

Hand DJ. (2006). Classifier Technology and the Illusion of Progress. Statistical Science, 21(1), 1–14. Retrieved from JSTOR. [PubMed: 17906740]

Hannan C, Lambert MJ, Harmon C, Nielsen SL, Smart DW, Shimokawa K, & Sutton SW. (2005). A lab test and algorithms for identifying clients at risk for treatment failure. Journal of Clinical Psychology, 61(2), 155–163. 10.1002/jclp.20108 [PubMed: 15609357]

Hans E, & Hiller W. (2013). Effectiveness of and dropout from outpatient cognitive behavioral therapy for adult unipolar depression: A meta-analysis of nonrandomized effectiveness studies. Journal of Consulting and Clinical Psychology, 81(1), 75–88. [PubMed: 23379264]

Hansen NB, Lambert MJ, & Forman EM. (2002). The psychotherapy dose-response effect and its implications for treatment delivery services. Clinical Psychology: Science and Practice, 9(3), 329–343. 10.1093/clipsy/9.3.329

Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, & Conde JG. (2009). Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. Journal of Biomedical Informatics, 42(2), 377–381. 10.1016/jjbi.2008.08.010 [PubMed: 18929686]

Kapelner A, & Bleich J. (2016). bartMachine: Machine learning with Bayesian Additive Regression Trees. Journal of Statistical Software, 70(8), 1–40.

Keitner GI, Posternak MA, & Ryan CE. (2003). How Many Subjects With Major Depressive Disorder Meet Eligibility Requirements of an Antidepressant Efficacy Trial? The Journal of Clinical Psychiatry, 64(9), 1091–1093. [PubMed: 14628985]

Kendrick T, El-Gohary M, Stuart B, Gilbody S, Churchill R, Aiken L, … Moore M. (2016). Routine use of patient reported outcome measures (PROMs) for improving treatment of common mental health disorders in adults. Cochrane Database of Systematic Reviews, (7). 10.1002/14651858.CD011119.pub2

Kessler RC, van Loo HM, Wardenaar KJ, Bossarte RM, Brenner LA, Ebert DD,… Zaslavsky AM. (2017). Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. Epidemiology and Psychiatric Sciences, 26(1), 22–36. 10.1017/S2045796016000020 [PubMed: 26810628]

Knaup C, Koesters M, Schoefer D, Becker T, & Puschner B. (2009). Effect of feedback of treatment outcome in specialist mental healthcare: Meta-analysis. The British Journal of Psychiatry, 195(1), 15–22. 10.1192/bjp.bp.108.053967 [PubMed: 19567889]

Kroenke K, & Spitzer RL. (2002). The PHQ-9: A new depression diagnostic and severity measure. Psychiatric Annals, 32(9), 509–515. 10.3928/0048-5713-20020901-06

Kuhn M. (2008). Caret package. Journal of Statistical Software, 28(5), 1–26. [PubMed: 27774042]

Kuhn M, & Johnson K. (2013). Applied Predictive Modeling. Retrieved from //www.springer.com/us/book/9781461468486

Kuyken W. (2004). Cognitive therapy outcome: The effects of hopelessness in a naturalistic outcome study. Behaviour Research and Therapy, 42(6), 631–646. [PubMed: 15081881]

LeDell E, van der Laan MJ, & Petersen M. (2016). AUC-Maximizing Ensembles through Metalearning. The International Journal of Biostatistics, 12(1), 203–218. [PubMed: 27227721]

Lee S, Rothbard AB, & Noll EL. (2012). Length of inpatient stay of persons with serious mental illness: Effects of hospital and regional characteristics. Psychiatric Services (Washington, D.C.), 63(9), 889–895. 10.1176/appi.ps.201100412

Linehan M. (2014). DBT Skills Training Manual. Guilford Publications.

Lorenzo-Luaces L, DeRubeis RJ, van Straten A, & Tiemens B. (2017). A prognostic index (PI) as a moderator of outcomes in the treatment of depression: A proof of concept combining multiple variables to inform risk-stratified stepped care models. Journal of Affective Disorders, 213, 78–85. 10.1016/jjad.2017.02.010 [PubMed: 28199892]

Lorenzo-Luaces L, Zimmerman M, & Cuijpers P. (2018). Are studies of psychotherapies for depression more or less generalizable than studies of antidepressants? Journal of Affective Disorders, 234, 8–13. 10.1016/jjad.2018.02.066 [PubMed: 29522947]

Löwe B, Kroenke K, Herzog W, & Gräfe K. (2004). Measuring depression outcome with a brief self-report instrument: Sensitivity to change of the Patient Health Questionnaire (PHQ-9). Journal of Affective Disorders, 81(1), 61–66. 10.1016/S0165-0327(03)00198-8 [PubMed: 15183601]

Lutz W, Schiefele A-K, Wucherpfennig F, Rubel J, & Stulz N. (2016). Clinical effectiveness of cognitive behavioral therapy for depression in routine care: A propensity score based comparison between randomized controlled trials and clinical practice. Journal of Affective Disorders, 189, 150–158. 10.1016/jjad.2015.08.072 [PubMed: 26433763]

Mahableshwarkar AR, Zajecka J, Jacobson W, Chen Y, & Keefe RS. (2015). A Randomized, Placebo-Controlled, Active-Reference, Double-Blind, Flexible-Dose Study of the Efficacy of Vortioxetine on Cognitive Function in Major Depressive Disorder. Neuropsychopharmacology. 10.1038/npp.2015.52

Martell CR, Dimidjian S, & Herman-Dunn R. (2013). Behavioral activation for depression: A clinician's guide. Guilford Press.

McEvoy PM, & Nathan P. (2007). Effectiveness of cognitive behavior therapy for diagnostically heterogeneous groups: A benchmarking study. Journal of Consulting and Clinical Psychology, 75(2), 344–350. 10.1037/0022-006X.75.2.344 [PubMed: 17469892]

Merrill KA, Tolbert VE, & Wade WA. (2003). Effectiveness of cognitive therapy for depression in a community mental health center: A benchmarking study. Journal of Consulting and Clinical Psychology, 71(2), 404–409. [PubMed: 12699035]

Nierenberg AA, Husain MM, Trivedi MH, Fava M, Warden D, Wisniewski SR, … Rush AJ. (2010). Residual symptoms after remission of major depressive disorder with citalopram and risk of relapse: A STAR*D report. Psychological Medicine, 40(1), 41–50. 10.1017/S0033291709006011 [PubMed: 19460188]

Papakostas GI, Nutt DJ, Hallett LA, Tucker VL, Krishen A, & Fava M. (2006). Resolution of Sleepiness and Fatigue in Major Depressive Disorder: A Comparison of Bupropion and the Selective Serotonin Reuptake Inhibitors. Biological Psychiatry, 60(12), 1350–1355. https://doi.org/10.10167j.biopsych.2006.06.015 [PubMed: 16934768]

Partial Hospital Programs. (2019, April 26). Retrieved April 26, 2019, from AABH website: https://www.aabh.org/partial-hospitalization-progra

Persons JB, Burns DD, & Perloff JM. (1988). Predictors of dropout and outcome in cognitive therapy for depression in a private practice setting. Cognitive Therapy and Research, 12(6), 557–575. 10.1007/BF01205010

Pizzagalli DA, Webb CA, Dillon DG, Tenke CE, Kayser J, Goer F, … Trivedi M. (2018). Pretreatment Rostral Anterior Cingulate Cortex Theta Activity in Relation to Symptom Improvement in Depression: A Randomized Clinical Trial. JAMA Psychiatry. 10.1001/jamapsychiatry.2018.0252

Polley EC, LeDell E, & van der Laan MJ. (2016). SuperLearner: Super Learner Prediction (Version 2.0–23) [R Package].

Prins MA, Verhaak PF, Hilbink-Smolders M, Spreeuwenberg P, Laurant MG, van der Meer K, … Bensing JM. (2011). Outcomes for depression and anxiety in primary care and details of treatment: A naturalistic longitudinal study. BMC Psychiatry, 11(1), 180 10.1186/1471-244X-11-180 [PubMed: 22099636]

R Core Team. (2013). R: A language and environment for statistical computing. Retrieved from http://www.R-project.org

Riedel M, Moller H-J, Obermeier M, Adli M, Bauer M, Kronmuller K, … Seemuller F. (2011). Clinical predictors of response and remission in inpatients with depressive syndromes. Journal of Affective Disorders, 133(1), 137–149. [PubMed: 21555156]

Ronalds C, Creed F, Stone K, Webb S, & Tomenson B. (1997). Outcome of anxiety and depressive disorders in primary care. The British Journal of Psychiatry, 171(5), 427–433. [PubMed: 9463600]

Rose S. (2013). Mortality Risk Score Prediction in an Elderly Population Using Machine Learning. American Journal of Epidemiology, 177(5), 443–452. [PubMed: 23364879]

Rosellini AJ, Dussaillant F, Zubizarreta JR, Kessler RC, & Rose S. (2018). Predicting posttraumatic stress disorder following a natural disaster. Journal of Psychiatric Research, 96, 15–22. 10.1016/jjpsychires.2017.09.010 [PubMed: 28950110]

Rush AJ, Fava M, Wisniewski SR, Lavori PW, Trivedi MH, Sackeim HA, … for the STAR*D Investigators Group. (2004). Sequenced treatment alternatives to relieve depression (STAR*D): Rationale and design. Controlled Clinical Trials, 25(1), 119–142. [PubMed: 15061154]

Schindler A, Hiller W, & Witthöft M. (2013). What Predicts Outcome, Response, and Drop out in CBT of Depressive Adults? A Naturalistic Study. Behavioural and Cognitive Psychotherapy, 41(3), 365–370. 10.1017/S1352465812001063 [PubMed: 23211066]

Shah AD, Bartlett JW, Carpenter J, Nicholas O, & Hemingway H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. American Journal of Epidemiology, 179(6), 764–774. [PubMed: 24589914]

Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, … Dunbar GC (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. The Journal of Clinical Psychiatry, 59 Suppl 20, 22–33;quiz 34–57.

Shimokawa K, Lambert MJ, & Smart DW. (2010). Enhancing treatment outcome of patients at risk of treatment failure: Meta-analytic and mega-analytic review of a psychotherapy quality assurance system. Journal of Consulting and Clinical Psychology, 78(3), 298–311. 10.1037/a0019247 [PubMed: 20515206]

Song L, Langfelder P, & Horvath S. (2013). Random generalized linear model: A highly accurate and interpretable ensemble predictor. BMC Bioinformatics, 14, 5. [PubMed: 23323760]

Spitzer RL, Kroenke K, Williams JBW, & Löwe B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. Archives of Internal Medicine, 166(10), 1092–1097. 10.1001/archinte.166.10.1092 [PubMed: 16717171]

Steinwart I, & Christmann A. (2008). Support Vector Machines. Springer Science & Business Media.

Stekhoven DJ, & Bühlmann P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112–118. [PubMed: 22039212]

Stone CJ, Hansen M, Kooperberg C, & Truong YK. (1997). The use of polynomial splines and their tensor products in extended linear modeling (with discussion). Annals of Statistics, 25, 1371–1470.

Targum SD, & Fava M. (2011). Fatigue as a Residual Symptom of Depression. Innovations in Clinical Neuroscience, 5(10), 40–43.

Thimm JC, & Antonsen L. (2014). Effectiveness of cognitive behavioral group therapy for depression in routine practice. BMC Psychiatry, 14(1), 292. [PubMed: 25330912]

Trivedi MH, McGrath PJ, Fava M, Parsey RV, Kurian BT, Phillips ML, … Weissman MM. (2016). Establishing moderators and biosignatures of antidepressant response in clinical care (EMBARC): Rationale and design. Journal of Psychiatric Research, 78, 11–23. 10.1016/jjpsychires.2016.03.001 [PubMed: 27038550]

Uher R, Maier W, Hauser J, Marusic A, Schmael C, Mors O, … McGuffin P. (2009). Differential efficacy of escitalopram and nortriptyline on dimensional measures of depression. The British Journal of Psychiatry, 194(3), 252–259. [PubMed: 19252156]

Uher R, Tansey KE, Malki K, & Perlis RH. (2012). Biomarkers predicting treatment outcome in depression: What is clinically significant? Pharmacogenomics, 13(2), 233–240. 10.2217/pgs.11.161 [PubMed: 22256872]

Van Straten A, Tiemens B, Hakkaart L, Nolen WA, & Donker MCH. (2006). Stepped care vs. matched care for mood and anxiety disorders: A randomized trial in routine practice. Acta Psychiatrica Scandinavica, 113(6), 468–476. [PubMed: 16677223]

Waldmann P, Mészáros G, Gredler B, Fuerst C, & Sölkner J. (2013). Evaluation of the lasso and the elastic net in genome-wide association studies. Frontiers in Genetics, 4, 1–11. 10.3389/fgene.2013.00270

Webb CA, Trivedi MH, Cohen ZD, Dillon DG, Fournier JC, Goer F, … Pizzagalli DA. (2018). Personalized prediction of antidepressant v. placebo response: Evidence from the EMBARC study. Psychological Medicine, 1–10.

Weisz JR, Weiss B, & Donenberg GR. (1992). The lab versus the clinic. Effects of child and adolescent psychotherapy. The American Psychologist, 47(12), 1578–1585. [PubMed: 1476328]

Zanarini MC, Vujanovic AA, Parachini EA, Boulanger JL, Frankenburg FR, & Hennen J. (2003). A screening measure for BPD: The McLean Screening Instrument for Borderline Personality Disorder (MSI-BPD). Journal of Personality Disorders, 17(6), 568–573. [PubMed: 14744082]

Zeeck A, von Wietersheim J, Weiss H, Scheidt CE, Volker A, Helesic A, … Hartmann A. (2016). Prognostic and prescriptive predictors of improvement in a naturalistic study on inpatient and day hospital treatment of depression. Journal of Affective Disorders, 197, 205–214. 10.1016/jjad.2016.03.039 [PubMed: 26995464]

Zetin M, & Hoepner CT. (2007). Relevance of Exclusion Criteria in Antidepressant Clinical Trials: A Replication Study. Journal of Clinical Psychopharmacology, 27(3), 295. [PubMed: 17502778]

Zimmerman M, Mattia JI, & Posternak MA. (2002). Are subjects in pharmacological treatment trials of depression representative of patients in routine clinical practice? The American Journal of Psychiatry, 159(3), 469–473. [PubMed: 11870014]

Public Health Significance Statement

Knowledge of which patients are likely to exhibit a poor outcome may have important clinical implications regarding treatment recommendations (e.g., a more intensive, alternative or combination treatment) and can inform more careful symptom and treatment progress monitoring. In the present study, we used machine learning to develop predictions of treatment outcome for depressed individuals seeking treatment in a “real-world” psychiatric hospital clinic. A prognosis calculator was developed which generates personalized predictions of treatment outcome for individual depressed patients.

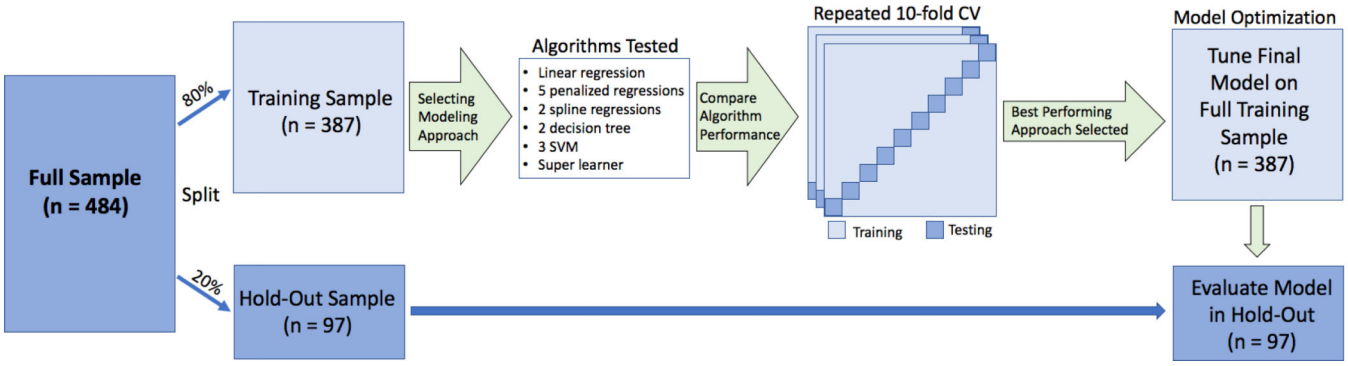

Figure 1.

Schematic of the analysis pipeline. The full sample (n = 484) was split into a training sample (80%) and a holdout validation sample (20%). We compared the performance of 13 machine learning algorithms (in addition to conventional linear regression) via 100 iterations of 10-folds cross-validation. The best performing model (lowest mean squared error; MSE) was subsequently tuned prior to implementation - without modification - in the hold-out validation sample.

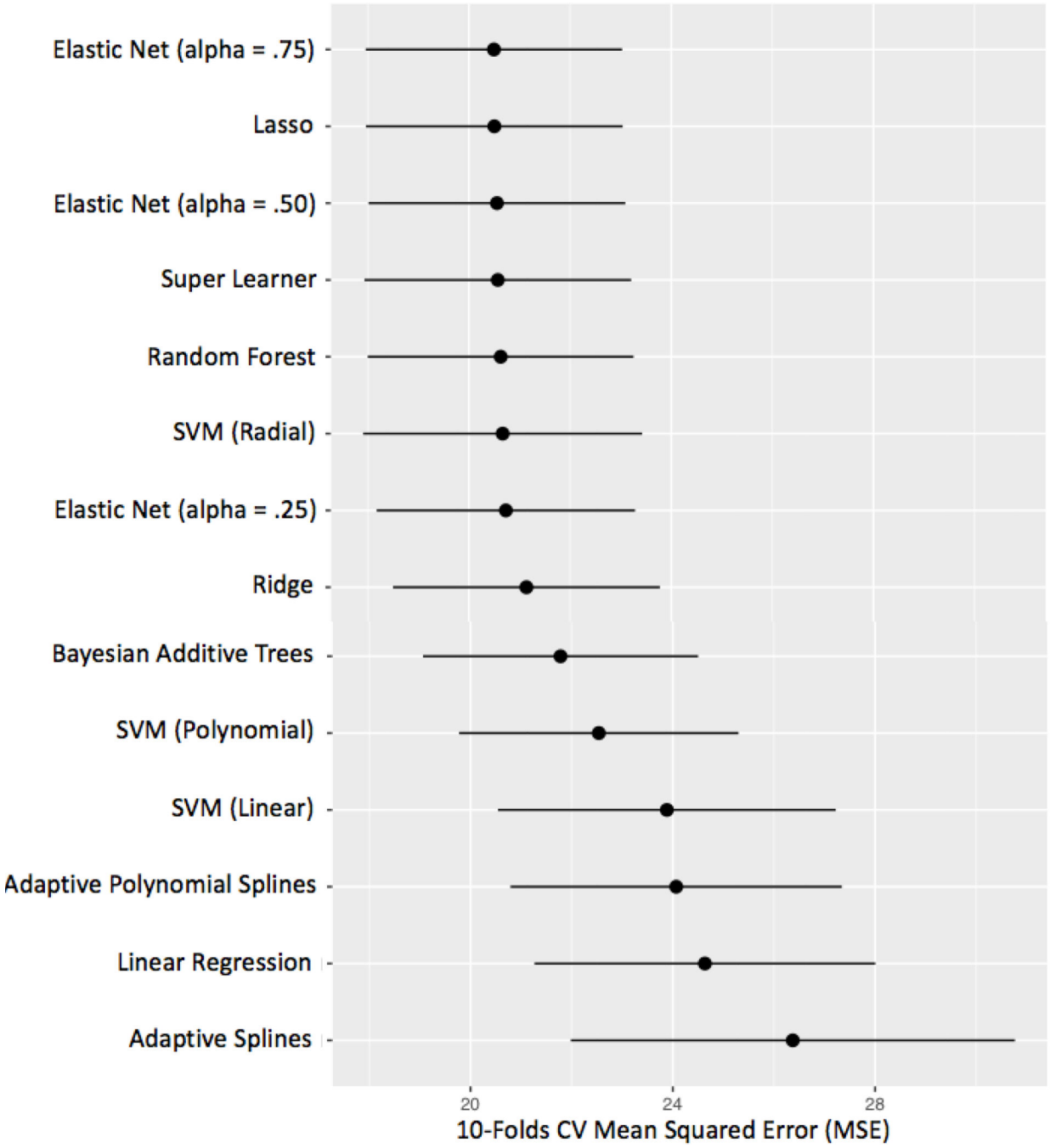

Figure 2.
Performance of the cross-validated algorithms within the training sample, displaying mean squared error (MSE) and standard error (SE) for each approach.

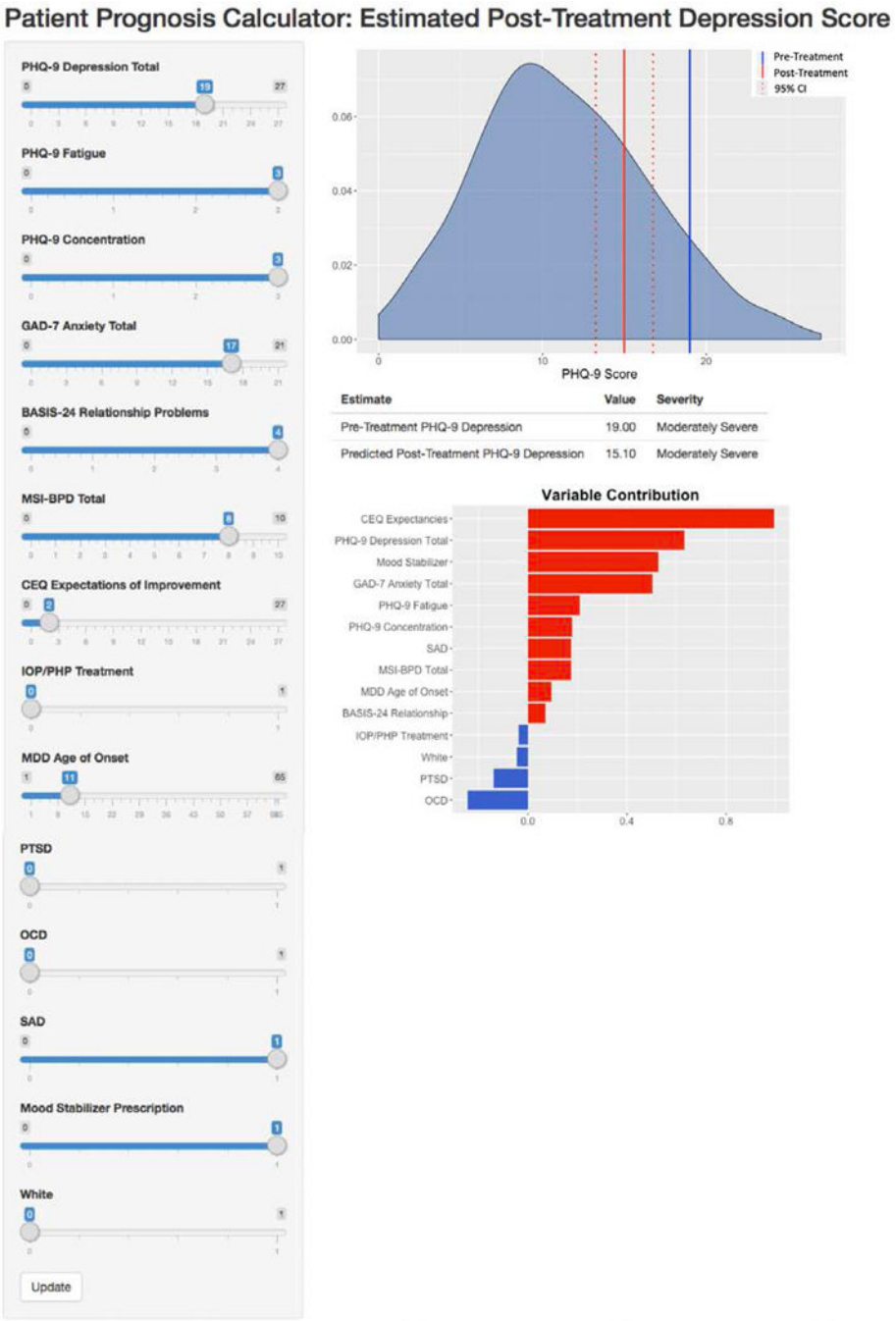

Figure 3.

Calculator generating estimated post-treatment PHQ-9 scores (with bootstrapped 95% confidence interval). A user inputs the observed values for each predictor variable for a new patient. Based on the final ENR model, the calculator generates predicted post-treatment PHQ-9 scores, which are overlaid on the distribution of outcome scores for the full sample (n = 484). The bottom panel illustrates the contribution of each variable to the predicted score for this patient. The variable contribution plot is generated by multiplying each individual's standardized scores (mean = 0; SD = 1) on each continuous predictor with the

beta coefficients from the final ENR model. Features that increase risk of poor prognosis - relative to the sample mean -are colored red; whereas those that decrease risk are in blue.

Table 1.

Demographic and Clinical Characteristics of the Sample.

Sample Characteristics			
	N	**%**	***Missing(%)***
Biological Sex			
Female	290	59.9	*0(0.0)*
Male	194	40.1	*0(0.0)*
Gender			
Female	282	58.3	*3(0.6)*
Male	192	39.7	*3(0.6)*
Non-Binary	7	1.4	*3(0.6)*
Race			
Native American or Alaskan Native	5	1.0	*0(0.0)*
Asian	35	7.2	*0(0.0)*
Black	18	3.7	*0(0.0)*
Native Hawaiian or Pacific Islander	4	0.8	*0(0.0)*
White	431	89.0	*0(0.0)*
Other	12	2.5	*0(0.0)*
Do not know	2	0.4	*0(0.0)*
Ethnicity			
Non-Latinx	458	94.6	*3(0.6)*
Latinx	23	4.8	*3(0.6)*
Education			
8th grade or less	0	0.0	*2(0.4)*
Some high school	3	0.6	*2(0.4)*
High school/GED	32	6.6	*2(0.4)*
Some college/Associate’s degree	170	35.1	*2(0.4)*
4-year college graduate	127	26.2	*2(0.4)*
Post-college education	150	31.0	*2(0.4)*
Employment			
Current student	150	31.0	*4(0.8)*
Unemployed (if not a current student)	122	25.2	*4(0.8)*
Employed part-time (if not a current student)	54	11.2	*4(0.8)*
Employed full-time (if not a current student)	150	31.0	*4(0.8)*
Marital Status			
Never Married	295	61.0	*2(0.4)*
Separated, divorced, or widowed	53	10.9	*2(0.4)*
Married or living with partner	134	27.7	*2(0.4)*
Past history of psychiatric hospitalization			
Yes	262	54.1	*10(2.1)*
No	212	43.8	*10(2.1)*
Current Diagnoses (DSM-IV-TR)			

Sample Characteristics			
	N	**%**	***Missing(%)***
Major Depressive Episode	484	100.0	*0(0.0)*
Bipolar Disorder (history of mania/hypomania)	105	21.7	*4(1.2)*
Psychotic Disorder	10	2.1	*23(4.8)*
Generalized Anxiety Disorder	230	47.5	*31(6.4)*
Social Anxiety Disorder	195	40.3	*8(1.7)*
Post-Traumatic Stress Disorder	83	17.1	*19(3.9)*
Panic Disorder	115	23.8	*8(1.7)*
Obsessive-Compulsive Disorder	82	16.9	*11(2.3)*
Alcohol Abuse or Dependence	100	20.7	*18(3.7)*
Medication			
Any antidepressant medication			
Primary antidepressant type: SSRI/SNRI	264	54.5	*0(0.0)*
Primary antidepressant type: Tricyclic	7	1.4	*0(0.0)*
Primary antidepressant type: MAOI	3	0.6	*0(0.0)*
Primary antidepressant type: Tetracyclic	33	6.8	*0(0.0)*
Primary antidepressant type: Other	63	13.0	*0(0.0)*
Other medication			
Primary antianxiety medication type: benzodiazepine	183	37.8	*0(0.0)*
Primary antianxiety medication type: other	42	8.7	*0(0.0)*
Primary antipsychotic medication type: typical	8	1.7	*0(0.0)*
Primary antipsychotic medication type: atypical	167	34.5	*0(0.0)*
Primary mood stabilizer type: lithium	28	5.8	*0(0.0)*
Primary mood stabilizer type: anti-epileptic	117	24.2	*0(0.0)*
Primary mood stabilizer type: other	9	1.9	*0(0.0)*
Stimulant/ADHD medication	83	17.1	*0(0.0)*
Sleep medication	15	3.1	*0(0.0)*
	M	*SD*	*Missing(%)*
Age (in years)	34.0	13.3	*0(0.0)*
Duration of treatment (including weekends/holidays)	13.0	4.2	*0(0.0)*
Total number of diagnoses	2.9	1.35	*37(7.6)*

Table 2.

Baseline predictors submitted to models, including estimates of internal consistency for multi-item scales.

Clinical Measures	Demographics	MDD History & Comorbidities
PHQ-9 Total * (α = .77)	Sex (male/female)	MDD single episode or recurrent
PHQ-9 Interest/Pleasure	Age	MDD number of episodes
PHQ-9 Depressed Mood	Race (white/other) *	MDD age of onset *
PHQ-9 Insomnia/Hypersomnia	Education level	Panic Disorder
PHQ-9 Fatigue *	Employment status (part- or full- time/unemployed)	Agoraphobia
PHQ-9 Appetite	Marital Status (Married or living with partner/other)	Social Anxiety Disorder *
PHQ-9 Self-Esteem/Guilt	**Treatment History**	Post-Traumatic Stress Disorder *
PHQ-9 Concentration *	Referral source (inpatient/outpatient)	Obsessive Compulsive Disorder *
PHQ-9 Psychomotor symptoms	Prior inpatient treatment (yes/no)	Generalized Anxiety Disorder
PHQ-9 Suicidality	Prior IOP/PHP treatment (yes/no) *	Alcohol Dependence
BASIS-24 Self-Harm (α = .83)	Psychiatric hospitalization within the past 6 months	**Physical Health**
BASIS-24 Emotional Lability (α = .70)	(yes/no)	Physical Health Rating
BASIS-24 Psychosis (α = .68)	**Psychiatric Medications**	Weight
BASIS-24 Substance Abuse (α = .73)	Antidepressant (yes/no)	Body Mass Index
BASIS-24 Relationships * (α = .75)	Antianxiety (yes/no)	Blood Pressure (Systolic)
MSI-BPD Total * (α = .71)	Mood stabilizer (yes/no) *	Blood Pressure (Diastolic)
Suicide Risk (low/mid/high)	Antipsychotic (yes/no)	Waist Circumference
GAD-7 Total * (α = .86)	Stimulant (yes/no)	
CEQ Expectancy * (α = .90)		
CEQ Credibility (α = .81)		

Note. Baseline predictors retained in the final elastic net model are listed with an asterisk. Measures and previous validation studies: PHQ-9 = 9-item Patient Health Questionnaire (Kroenke & Spitzer, 2002); BASIS-24 = 24-item Behavior and Symptom Identification Scale (Cameron et al., 2007); GAD-7 = 7-item Generalized Anxiety Disorder scale (Spitzer, Kroenke, Williams, & Löwe, 2006); MSI-BPD = McLean Screening Instrument for Borderline Personality Disorder (Zanarini et al., 2003); CEQ = Credibility/Expectancy questionnaire (Devilly & Borkovec, 2000); IOP = Intensive outpatient program; PHP = Partial hospital program; MDD = Major Depressive Disorder.

* Predictors selected by final elastic net model.

Table 3.

Performance of the cross-validated algorithms in the training sample

Type of Algorithm	Algorithm	MSE	SE	MAE	R^2	SL Wgt
Conventional OLS regression	*Linear regression*	24.64	1.72	4.05	.14	-
Penalized regression (i.e., regularization)	*Ridge regression*	21.14	1.35	3.84	.26	-
	Elastic net (alpha = 0.25)	20.73	1.30	3.81	.27	-
	Elastic net (alpha = 0.50)	20.56	1.29	3.79	.28	-
	Elastic net (alpha = 0.75)	20.49	1.29	3.78	.28	-
	Lasso regression	20.50	1.30	3.78	.28	.32
Spline regression	*Adaptive splines*	26.37	2.24	4.11	.08	-
	Adaptive polynomial splines	24.07	1.67	4.00	.16	.18
Decision tree	*Random forests*	20.63	1.34	3.79	.28	.22
	Bayesian additive trees	21.78	1.39	3.90	.24	-
Support vector regression	*SVR (linear)*	23.89	1.70	3.98	.16	-
	SVR (polynomial)	22.54	1.67	3.99	.21	-
	SVR (radial)	20.67	1.41	3.80	.28	.27
Ensembling method	*Super learner*	20.57	1.35	3.79	.28	N/A

Note. MSE = Mean Square Error; SE = Standard Error of MSE; MAE: Mean Absolute Error; R^2: 1 - (MSE/var(y)); SL WGT = Super Learner weighting; OLS = Ordinary Least Squares; SVR = Support Vector Regression.

Table 4.

Baseline variables retained in final elastic net model

Variable	*B*
(Intercept)	11.99
PHQ-9 Depression Total	1.82
PHQ-9 Fatigue	0.25
PHQ-9 Concentration	0.17
GAD-7 Anxiety Total	0.59
BASIS-24 Relationship Problems	0.01
MSI-BPD Total	0.12
CEQ Expectations of Improvement	-0.44
Prior IOP or PHP Treatment	0.04
MDD Age of Onset	-0.12
Social Anxiety Disorder	0.33
Obsessive Compulsive Disorder	0.44
Post-Traumatic Stress Disorder	0.25
Mood Stabilizer Prescription	1.01
White (Yes/No)	0.02

Note. PHQ-9 = 9-item Patient Health Questionnaire; GAD-7 = 7-item Generalized Anxiety Disorder scale; MSI-BPD = McLean Screening Instrument for Borderline Personality Disorder; CEQ = credibility/expectancy questionnaire; IOP = Intensive outpatient program; PHP = Partial hospital program; MDD = Major Depressive Disorder.