

Original Article

Cite this article: Puac-Polanco V *et al* (2023). Development of a model to predict antidepressant treatment response for depression among Veterans. *Psychological Medicine* **53**, 5001–5011. <https://doi.org/10.1017/S0033291722001982>

Received: 25 February 2022

Revised: 8 June 2022

Accepted: 13 June 2022

First published online: 15 July 2022

Key words:


Antidepressant medication; clinical decision support; depression; machine learning; treatment response; Veterans Health Administration

Author for correspondence:

Ronald C. Kessler,

E-mail: kessler@hcp.med.harvard.edu

Development of a model to predict antidepressant treatment response for depression among Veterans

Victor Puac-Polanco¹, Hannah N. Ziobrowski¹, Eric L. Ross^{2,3,4}, Howard Liu^{1,5}, Brett Turner^{1,5,6}, Ruifeng Cui^{7,8}, Lucinda B. Leung^{9,10}, Robert M. Bossarte^{5,11}, Corey Bryant¹², Jutta Joormann¹³, Andrew A. Nierenberg^{4,14}, David W. Oslin^{15,16}, Wilfred R. Pigeon^{5,17}, Edward P. Post^{12,18}, Nur Hani Zainal¹, Alan M. Zaslavsky¹, Jose R. Zubizarreta^{1,19,20}, Alex Luedtke^{21,22}, Chris J. Kennedy^{3,4}, Andrea Cipriani²³, Toshiaki A. Furukawa²⁴ and Ronald C. Kessler¹ 

Abstract

Background. Only a limited number of patients with major depressive disorder (MDD) respond to a first course of antidepressant medication (ADM). We investigated the feasibility of creating a baseline model to determine which of these would be among patients beginning ADM treatment in the US Veterans Health Administration (VHA).

Methods. A 2018–2020 national sample of $n = 660$ VHA patients receiving ADM treatment for MDD completed an extensive baseline self-report assessment near the beginning of treatment and a 3-month self-report follow-up assessment. Using baseline self-report data along with administrative and geospatial data, an ensemble machine learning method was used to develop a model for 3-month treatment response defined by the Quick Inventory of Depression Symptomatology Self-Report and a modified Sheehan Disability Scale. The model was developed in a 70% training sample and tested in the remaining 30% test sample.

Results. In total, 35.7% of patients responded to treatment. The prediction model had an area under the ROC curve (s.e.) of 0.66 (0.04) in the test sample. A strong gradient in probability (s.e.) of treatment response was found across three subsamples of the test sample using training sample thresholds for high [45.6% (5.5)], intermediate [34.5% (7.6)], and low [11.1% (4.9)] probabilities of response. Baseline symptom severity, comorbidity, treatment characteristics (expectations, history, and aspects of current treatment), and protective/resilience factors were the most important predictors.

Conclusions. Although these results are promising, parallel models to predict response to alternative treatments based on data collected before initiating treatment would be needed for such models to help guide treatment selection.

Introduction

Major depressive disorder (MDD) has high prevalence and high impairment (GBD 2019 Diseases and Injuries Collaborators, 2020). The two primary first-line MDD treatments are psychotherapy and antidepressant medication (ADM; Qaseem, Barry, & Kansagara, 2016). ADM is the more common treatment despite most patients preferring psychotherapy (McHugh, Whitton, Peckham, Welge, & Otto, 2013) due to lower cost and wider availability (Hockenberry, Joski, Yarbrough, & Druss, 2019). But some MDD patients do not respond to ADMs (Cipriani *et al.*, 2018; Kazdin *et al.*, 2021; Little, 2009) but do to psychotherapy or an ADM-psychotherapy combination. However, the latter treatments often are provided only after months of unsuccessful ADM treatment (Day *et al.*, 2021). A meaningful proportion of patients drop out before receiving other treatments (Larson *et al.*, 2021). A strategy to predict likelihood of responding before initiating ADM treatment could be of value.

Many multivariable models have been developed, typically using machine learning (ML) methods (Chekroud *et al.*, 2021; Ermers, Hagoort, & Scheepers, 2020; Lee *et al.*, 2018), to predict depression treatment response. Most such models can be faulted, though, for (i) low external validity because of restriction to clinical trial samples; (ii) focus on biomarkers infeasible to use in routine clinical practice; (iii) including many fewer predictors than documented in the literature; or (iv) suboptimal analytic methods.

The current report presents results of a study designed to address these problems by analyzing an observational sample of patients recruited near beginning ADM treatment and administered an extensive baseline battery of self-report questions to assess predictors of ADM treatment response found in previous studies. The patients were followed for 3 months to

assess treatment response. The data were analyzed using a state-of-the-art stacked generalization ML method.

Materials and methods

Sample

As detailed elsewhere (Puac-Polanco *et al.*, 2021), a probability sample of patients beginning MDD outpatient treatment was selected from Veterans Health Administration (VHA) electronic health records (EHRs) December 2018–June 2020. Inclusion criteria were: (i) beginning first outpatient MDD treatment in the past year; and (ii) receiving ADM prescription and/or psychotherapy referral. Exclusions were: (i) 12-month suicide attempt; (ii) lifetime diagnoses of bipolar disorder, nonaffective psychosis, dementia, intellectual disability, autism, Tourette's disorder, stereotyped movement disorder, or borderline intellectual functioning; (iii) lifetime prescriptions of mood stabilizers or antipsychotic medications (online Supplementary Table S1). The exclusion of 12-month suicide attempts was made because such patients in VHA are placed on a high-risk list that leads to intensive case management, making the experiences of these patients unrepresentative of the more general patient population.

Recruitment letters were sent to 55 106 provisionally eligible patients the day after their first outpatient visit. The letter described the study purposes and the requirements of self-report web- or phone-based baseline assessment taking 45 min and at 3 months follow-up taking 20 min, with compensation of \$50 and \$25, respectively, for the two assessments. A team member then attempted to contact each patient over the next week (three call attempts). Of the 17 000 reached, 6298 agreed to participate and 4164 completed the baseline assessment (online Supplementary Fig. S1). At baseline, 809 respondents had received an ADM prescription without referral to psychotherapy and were otherwise eligible. The 660 of these 809 who completed the 3-month assessment are the focus of this report. The protocol was approved by the Institutional Review Board of Syracuse VA Medical Center, Syracuse, New York.

Measures

Treatment response

Two-week depressive symptoms were assessed with the 16-item Quick Inventory of Depression Symptomatology Self-Report (QIDS-SR; Rush *et al.*, 2003). A modified version of the Sheehan Disability Scale (SDS; Leon, Olfson, Portera, Farber, & Sheehan, 1997) was used to assess role impairment by asking patients how much depression interfered with the ability to work, participate in family and home life, or participate in social activities in the past 2 weeks on a 0–10 visual analog scale with response options of *not at all* (0), *mildly* (1–3), *moderately* (4–6), *markedly* (7–9), and *extremely* (10) (Cronbach's $\alpha = 0.85$).

Treatment response was defined as either (i) a 3-month QIDS-SR score no more than half its baseline value or (ii) a baseline SDS score of 4–10 in *any* role impairment domain along with a 3-month SDS score of 0–3 in *all* such domains. A similar composite definition of ADM treatment response was used in previous research (Huang *et al.*, 2018; Wang *et al.*, 2018; Zilcha-Mano *et al.*, 2021).

Predictors

Numerous recent reports have carried out reviews or meta-analyses of research on baseline predictors of response to

individual types of depression treatment (e.g. Furukawa *et al.*, 2021; Noma *et al.*, 2019) or treatment in general pooled across multiple treatment types (e.g. Buckman *et al.*, 2021a, 2021b, 2021c, 2022), which are referred to collectively as *prognostic* predictors. Other reviews have examined baseline variables that interact significantly with treatment type to predict outcomes (Maj *et al.*, 2020; Perlman *et al.*, 2019; Perna, Alciati, Daccò, Grassi, & Caldirola, 2020), which are referred to as *prescriptive* predictors. Predictors from all important domains of either prescriptive or prognostic predictors were included in our baseline questionnaire or abstracted from EHRs or government small-area geospatial databases linked to patient residential addresses. Included here were six domains involving the episodes (symptom frequency, severity, subtypes, clinical staging, psychiatric comorbidities, functioning, and quality of life), two others involving stressors (early environmental exposures, recent environmental stressors), and three involving personality/cognition (personality scales, neurocognition, dysfunctional cognitive schemas). A separate domain of 'protective/resilience factors' assessed patient psychological characteristics (e.g. coping styles, self-reported psychological resilience) and environmental resources (e.g. access to supportive social relationships; access to material resources). Two other domains included information about comorbid physical disorders and family history of psychopathology.

We also included information about socio-demographics and treatment characteristics associated in previous research with differential depression treatment response (Constantino, Vislà, Coyne, & Boswell, 2018; Kraus, Kadriu, Lanzenberger, Zarate, & Kasper, 2019). Treatment characteristics included self-reports about current expectations, which were assessed as of the time of baseline rather than asking patients to recall their expectations prior to making their initial treatment contact. This time frame is relevant because, as noted above, the baseline assessment was carried out only after treatment started. Treatment characteristics also included patient self-reports about past treatment experiences and EHR data on treatment histories and ADM types prescribed in the current treatment. The latter were classified as norepinephrine-dopamine reuptake inhibitors (NDRI), serotonin antagonist reuptake inhibitors (SARI), serotonin modulator and stimulators, serotonin-norepinephrine reuptake inhibitors (SNRI), selective serotonin reuptake inhibitors (SSRI), tricyclic antidepressants, and tetracyclic antidepressants. We also included a dummy variable for ADMs suggested as most effective in controlled trials (i.e. escitalopram, mirtazapine, paroxetine, sertraline, venlafaxine) (Cipriani *et al.*, 2018; Kazdin *et al.*, 2021; Little, 2009). Other dummy variables were included for typical combinations of ADMs with baseline symptoms (e.g. trazodone with sleep disturbance, duloxetine with severe physical pain). We also recorded whether the treatment provider was the patient's regular primary care physician, someone else in the same primary care office, or someone at a mental health specialty clinic.

Categorical variables were coded as dummy indicators. Quantitative variables were standardized to a mean of 0 and variance of 1 and discretized into quintiles to create stabilized predictors and nested dichotomies. These transformations resulted in 2768 potential predictors (online Supplementary Tables S2–S4). Item-level missingness was handled by single imputation carried out in the total sample before defining separate training and test samples, with missing values imputed to the mode for dichotomous and categorical variables and to the mean for ordinal and interval variables.

Analysis methods

The R program *sbw* (Zubizarreta, Li, Allouah, & Greifer, 2021) was used to make weighting adjustments for: (i) discrepancies in baseline EHR variables between eligible VHA patients and the 809 baseline respondents and (ii) discrepancies in baseline survey variables between the 660 3-month follow-up respondents and nonrespondents (Zubizarreta, 2015).

The Super Learner (SL) stacked generalization ML method (Polley, LeDell, Kennedy, Lendle, & van der Laan, 2021) was used to develop a prediction model in the weighted sample of 3-month respondents. SL generates predictions from a weighted combination of conventional and flexible ML algorithms in an ensemble. Our SL specification used 10-fold cross-validation (10F-CV) to generate a weighted composite that performs at least as well in expectation as the best algorithm in the ensemble (Polley, Rose, & van der Laan, 2011). The appeal of stacked generalization over single algorithms is improved predictive accuracy by virtue of combining results across algorithms that include a wide range of functional forms (Polley et al., 2011). Consistent with recommendations (LeDell, van der Laan, & Petersen, 2016), a diverse set of algorithms was included in the SL ensemble (online Supplementary Table S5). Some prior computational psychiatric studies have used similar stacked generalization procedures (Karrer et al., 2019; Ziobrowski et al., 2021a).

We estimated the SL model in a stratified (by the outcome variable) random 70% training sample ($n = 462$) and validated it in the remaining 30% test sample ($n = 198$). Prediction strength, defined as area under the receiver operating characteristic curve [AUC (ROC)], was compared across a wide range of hyperparameter settings for each algorithm in the 10F-CV sample (online Supplementary Table S5). Predictors were selected independently in each 10F-CV fold with a range of constraints on predictor number using lasso regression (Park & Casella, 2008) for linear models and Bayesian additive regression trees (Chipman, George, & McCulloch, 2010) for nonadditive models. Comparisons of AUC (ROC) estimated in the full training sample and 10F-CV allowed determination of how much each learner (i.e. combination of number of allowed predictors and hyperparameter values for a given algorithm) was overfitting and CV prediction strength. A subset of learners with balance between these two criteria was selected for the final SL ensemble. Once the final SL model was estimated, 10F-CV was used for model calibration in the 10F-CV sample based on isotonic regression (Lindhiem, Petersen, Mentch, & Youngstrom, 2020).

Models were assessed in the test sample by how well predicted probability of treatment response ranked patients on observed response (i.e. discrimination). The AUC (ROC) and the AUC of the precision recall curve [AUC (PRC)] were compared for the SL and a simpler benchmark lasso regression model whose penalty parameter was selected via internal cross-validation, both estimated in the training sample and applied to the test sample. Operating characteristics in the test sample were then inspected across quantiles of predicted probability of response. Operating characteristics included conditional and cumulative *sensitivity* (SN; the proportion of all patients responding to treatment who were in the quantile) and *positive predictive value* (PPV; the prevalence of treatment response in the decile). A locally estimated scatterplot smoothed calibration curve (Austin & Steyerberg, 2014) with 0.75 bandwidth was used to quantify model calibration in the test sample using the integrated calibration index (ICI) and expected calibration error (ECE) (Austin &

Steyerberg, 2019). Model fairness (Yuan, Kumar, Ahmad, & Teredesai, 2021) was evaluated by examining variation in the association of predicted probability of response with observed response across socio-demographic subgroups related to health disparities (age, sex, race/ethnicity, education) using robust Poisson regression models (Zou, 2004).

Predictor importance was examined using the model-agnostic kernel Shapley Additive Explanations (SHAP) method (Lundberg & Lee, 2017), which generates a predicted difference in outcome score for each patient based on changing one and only one predictor at a time from its observed score to the mean across all logically possible permutations of other predictors. The mean of this 'SHAP value' for a given predictor across all patients is 0. However, the mean *absolute* SHAP value provides useful information about the average importance of the predictor. A bee swarm plot of the association between the individual-level SHAP value and the observed score for a given predictor was used to describe dominant direction of association. Mean absolute SHAP values can also be aggregated across subsets of predictors by summing SHAP values across the predictors at the individual level and then calculating the mean of the absolute value of this sum. Such aggregate scores estimate the expected change in prevalence of treatment response if all predictors for all patients changed from their observed values to the mean values.

SAS statistical software, version 9.4 (SAS Institute Inc, 2013) was used for data management, estimating prevalence of treatment response, and calculating SN, PPV, and AUC. R, version 4.0.5 (R Core Team, 2021) was used to estimate the SL model and SHAP values.

Reporting

We followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines (Collins, Reitsma, Altman, & Moons, 2015) in presenting results.

Results

Sample characteristics and treatment response

Baseline QIDS-SR scores were transformed to approximate Hamilton Rating Scale of Depression (HRSD) categories using published transformation rules (Table 3 in Rush et al., 2003) to give a sense of the baseline symptom severity distribution. In total, 30.1% of patients were classified as having baseline mild depression, 35.6% moderate, 21.4% severe, and 12.9% very severe (online Supplementary Table S6). Given that the baseline assessment was not administered until after the initiation of treatment, there is a possibility that these severities were lower than if assessments had been carried out prior to beginning treatment. However, the Pearson correlation between the baseline QIDS-SR score and number of days between beginning treatment and taking the baseline assessment (median = 21 days; inter-quartile range = 14–30 days) was nonsignificant ($r = 0.013$, $p = 0.74$).

The great majority (80.7%) of patients were prescribed a single ADM, most commonly SSRIs (57.0%), SNRIs (16.8%), NDRI (15.7%), and SARIs (15.0%). The modal socio-demographic categories were between 35 and 49 years of age, male, non-Hispanic White, married, and living in a major metropolitan area. Except for age ($p = 0.009$), no statistically significant differences were found in baseline socio-demographics, clinical

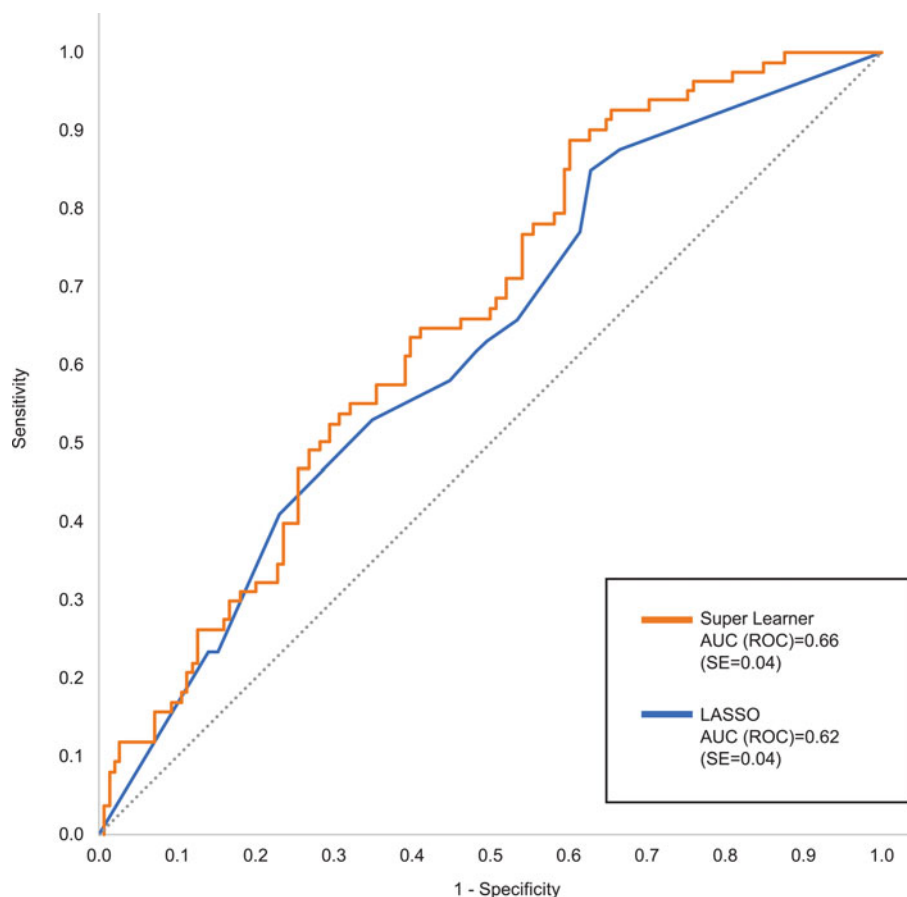


Fig. 1. Receiver operating characteristic (ROC) curve comparing Super Learner with benchmark lasso in the test sample.

severity, or ADM classes between the weighted baseline sample and the doubly weighted analytic sample. In total, 35.7% of doubly weighted test sample patients responded to treatment as of the 3-month assessment.

Model performance

The SL test sample AUC (ROC) was 0.66 compared to 0.62 for the benchmark lasso model (Fig. 1). SL had better SN than lasso for all values of specificity (SP) above 0.25. SL had much higher PPV than the lasso model for SN below 0.10 and somewhat higher PPV than lasso across most of the remaining SN range (Fig. 2). SL test sample AUC (ROC) remained 0.66 when the analysis was limited to patients classified as having at least moderate baseline symptom severity compared to AUC (ROC) of 0.60 among patients classified as having mild baseline symptom severity.

Calibration based on the isotonic regression transformation was (ICI = 0.28 and ECE = 0.34). The SL model also had comparable fairness across subgroups defined by age, sex, race/ethnicity, and education (online Supplementary Table S7).

A monotonic gradient was found in the proportion of test sample patients that responded to treatment [i.e., PPV (s.e.)] across SL model quantiles defined in the training sample. These quantiles could be collapsed without meaningful loss of information into three groups of patients (Table 1). Among patients in the first group, those with high predicted probabilities of response, 45.7% (5.5) responded to treatment. In the intermediate predicted probably group, 34.5% (7.6) responded. In the low predicted

probability group, 11.1% (4.9) responded. These predicted probabilities did not vary significantly between patients whose baseline symptom severity was mild *v.* more severe (Table 1). Although the thresholds to define these groups were for quintiles in the training sample, 50.4% of patients in the test sample fell into the high group, 30.1% the intermediate group, and 19.4% the low group.

Predictor importance

[The 2768 predictors were highly redundant, as indicated by 750 (27.1%) of them having significant univariable associations with the outcome in the training sample at the 0.05 level but only 53 (1.9%) being selected by SL (Fig. 3). Forty-six of these 53 were patient self-reports, four EHR variables, and three geospatial variables. The aggregate mean absolute SHAP value across all these predictors was 4.3%. This means the probability of treatment response would have changed by an estimated average of 4.3% if each patient's scores on all selected predictors changed from observed to sample-wide mean values.

The most important predictors were features of the episode (10 of 53 predictors), with an aggregate mean absolute SHAP value of 3.5% (81% of the total). This included the most important predictor, overall depressive symptom frequency in the 2 weeks before treatment, in addition to two other important symptom measures, frequency of being happy or at peace (third most important) and anhedonia (reverse coded, 14th most important), along with five indicators of current or recent comorbidity (4th, 15th, 22nd, 34th, 44th). These predictors were for the most part associated with reduced probability of treatment response.

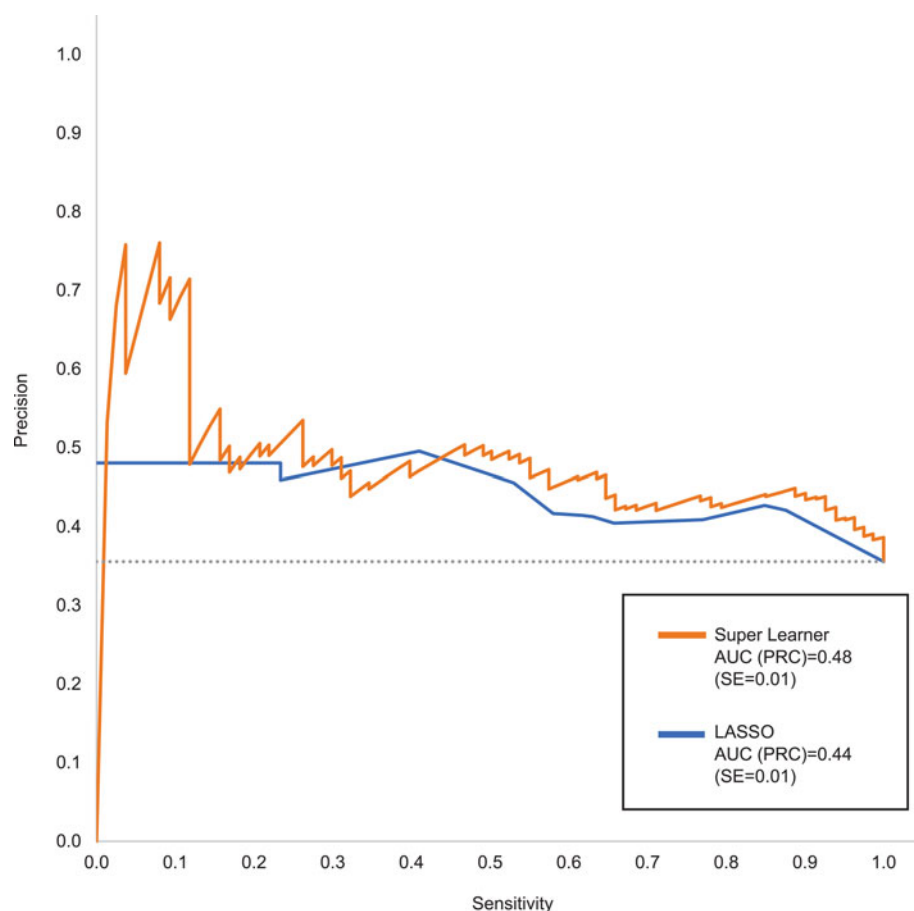


Fig. 2. Precision recall curve (PRC) comparing Super Learner with benchmark lasso in the test sample.

Table 1. Prediction of 3-month ADM treatment response in the test sample in three group defined by predicted probabilities in the training sample ($n = 198$)

	Distribution		PPV		Cumulative PPV		Sensitivity		Cumulative Sensitivity	
	%	(s.e.)	%	(s.e.)	%	(s.e.)	%	(s.e.)	%	(s.e.)
High	50.4 ^a	(4.1)	45.7 ^b	(5.5)	45.7	(5.5)	64.7	(6.7)	64.7	(6.7)
Intermediate	30.1 ^c	(3.9)	34.5 ^d	(7.6)	41.5	(4.5)	29.2	(6.6)	93.9	(2.7)
Low	19.4	(3.3)	11.1	(4.9)	35.6	(3.9)	6.0	(2.7)	100.0	–

ADM, antidepressant medication; PPV, positive predictive value (i.e. predicted proportion with treatment response); s.e., standard error; SN, sensitivity (i.e. proportion of all treatment responders).

^a80.9% among patients with mild baseline symptom severity v. 35.3% among other patients.

^b47.5% among patients with mild baseline symptom severity v. 45.7% among other patients.

^c13.7% in the subsample of patients with mild baseline symptom severity v. 38.3% among other patients.

^d30.3% among patients with mild baseline symptom severity v. 38.4% among other patients.

However, SHAP value distributions (Fig. 4) show that some associations were nonmonotonic. For example, lower-than-average but not lowest overall depressive symptom frequency was associated with highest probability of treatment response.

The second most important predictor domain involved treatment characteristics (22 of 53 selected predictors), with an aggregate mean absolute SHAP value of 1.2% (25% of the total). Included were 10 indicators of positive treatment expectation/preference (e.g. 8th most important, expectation of having a good relationship with treatment provider), all positively associated with treatment response. Another seven treatment-related predictors involved current treatment (e.g. sixth most important,

referral to a mental health specialist or psychologist carried out intake). None of the ADM types was among the important predictors. The remaining treatment-related predictors involved treatment history (e.g. 12th most important, past psychotherapy was not helpful), all positively associated with treatment response.

There were only two other important predictor domains: recent stressors and protective/resilience factors, with aggregate mean absolute SHAP values of 0.9% (26% of the total) and 0.7% (17% of the total), respectively. The most important stressors were financial (second) and high mortality rate due to drug overdose in the patient's county of residence (fifth), both negatively associated



sx, symptoms; freq, frequency; (D), dummy variable; (S), stabilized variable; ADM, antidepressant medication; tx, treatment; (RS), reverse stabilized; MDE, major depressive episode; px, patient; PCP, primary care provider; PTSD, post-traumatic stress disorder; BMI, body mass index; HRSD, Hamilton Rating Scale of Depression.

Fig. 3. Predictor importance as determined by Shapley Additive Explanation (SHAP) values for the Super Learner Model in the test sample^{†1}.

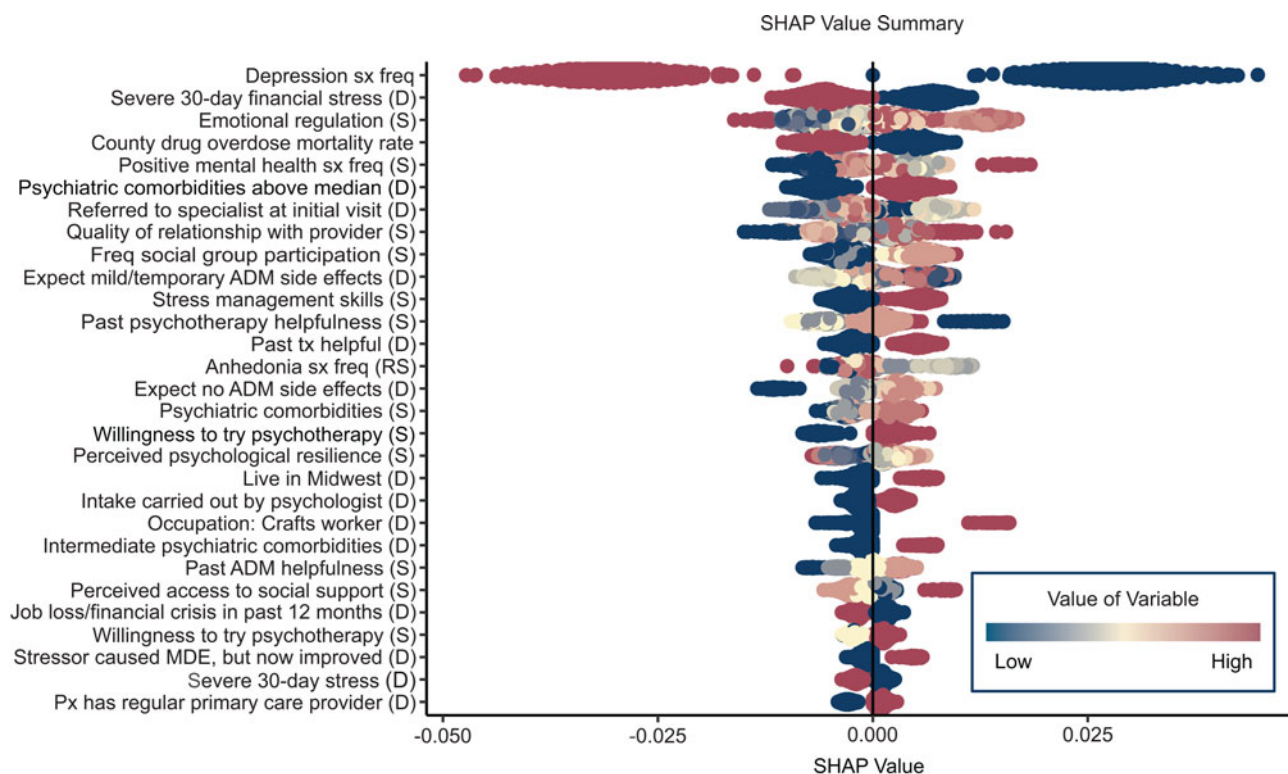
with treatment response. The protective/resilience factors included three indicators of psychological resilience and three of social support. As shown in the bee swarm plot, most of these predictors had nonmonotonic associations with the outcome due to patients with higher-than-average but not highest reported protective/resilience scores having highest probabilities of treatment response.

Discussion

The 35.7% 3-month ADM treatment response rate is comparable to previous VHA studies (Katz, Liebmann, Resnick, & Hoff, 2021) but lower than most civilian studies (Cuijpers *et al.*, 2020), presumably reflecting the greater severity/complexity of depressed Veterans than civilians (Ziobrowski *et al.*, 2021b). This highlights the potential importance of patients in the group with highest predicted probability of ADM response being more than four times as likely to respond as patients in the lowest group (45.7% *v.* 11.1%). Accurate discrimination of this sort is valuable as a first step in determining optimal treatments. However,

multiple treatment-specific models need to be developed and combined to create a precision treatment rule for optimal assignment of patients across interventions (Kessler & Luedtke, 2021). For instance, psychotherapy or ADM plus psychotherapy might be prioritized for patients with low predicted probabilities of ADM treatment response, but only in the subset of patients with higher predicted probabilities of response in treatment-specific models for psychotherapy or combined therapy.

In this respect, our model might be compared to pharmacogenomic models used to determine pre-emptively whether ADMs are likely to be effective for individual patients (Greden *et al.*, 2019). Our model performed at least as well as these pharmacogenomic models and could be implemented at a fraction of the cost of pharmacogenetic testing. The largest pharmacogenomic testing trial to date for ADM selection found that patients receiving test-congruent medications had a 12% higher probability of treatment response than patients receiving test-incongruent medications (29% *v.* 17%; Greden *et al.*, 2019), whereas the differences we found were 34.6% (45.7% *v.* 11.1%) between our high and low



sx, symptoms; freq, frequency; (D), dummy variable; (S), stabilized variable; ADM, antidepressant medication; tx, treatment; (RS), reverse stabilized; MDE, major depressive episode; px, patient; PCP, primary care provider; PTSD, post-traumatic stress disorder; BMI, body mass index; HRSD, Hamilton Rating Scale of Depression.

Fig. 4. Bee swarm plot of individual-level predictor-specific SHAP values for the most important predictors in the Super Learner model².

groups, 11.2% (45.7% *v.* 34.5%) between our top and intermediate groups, and 23.4% (34.5% *v.* 11.1%) between our intermediate and low groups.

Caution is needed in interpreting results regarding predictor importance because predictor importance rankings can be very unstable when, as in our dataset, many predictors are highly correlated (Leeuwenberg et al., 2022). Several broad results about predictor importance are nonetheless noteworthy. The most striking is that baseline clinical characteristics of the episode were by far the most important predictors. This is consistent with a recent individual-level meta-analysis of over 6000 patients in primary care depression treatment across 12 trials, where baseline depression symptom severity was by far the single most important predictor of treatment response independent of treatment type (Buckman et al., 2021b). Other significant clinical predictors in that recent meta-analysis included duration of the depressive episode before beginning treatment, comorbid panic, and duration of comorbid anxiety. We found a different set of important clinical predictors, including two secondary depressive symptom factors and several measures of psychiatric comorbidity, but, as in the meta-analysis, these were all much less important than overall baseline depression symptom frequency.

The secondary symptom factors (absence of positive emotions, anhedonia) are both central aspects of melancholic depression. Evidence in previous studies has been mixed for melancholic depression being less responsive to treatment than others (Maj et al., 2020). It is noteworthy that the baseline assessment included the other indicators of melancholia (i.e. deep feelings of despair, mood worse in the morning, early morning awakening,

psychomotor changes, weight loss, excessive guilt), but we did not attempt to define this or any other theoretical (Benazzi, 2006) or data-driven (Buckman et al., 2021a) MDD subtype beyond those that emerged in an exploratory factor analysis of symptoms in the baseline assessment. The nonadditive models in the SL ensemble would have been expected to detect interactions across these factors if a strong data-driven episode subtype existed. Nonetheless, it might be useful in future investigations to use unsupervised ML methods to explore the possibility of detecting such clusters.

The importance of treatment characteristics, the next most important predictor domain in our sample, was striking in two ways. First, ADM type was unrelated to treatment response. Second, multiple aspects of treatment history and current treatment expectations were important. Although the literature on treatment expectations is inconsistent in its measures and controls for prior experiences, our finding that both process and outcome expectations were important predictors is broadly consistent with previous studies (Laferton, Kube, Salzmann, Auer, & Shedden-Mora, 2017). This is striking given that we controlled for and found significant associations of several measures of past treatment experiences that presumably underlie expectations. Taken together, these results argue for the potential value of shared decision-making and patient-centered care for depression (Rush & Thase, 2018), for the potential value of expanding interventions to influence treatment expectations (Gruszka, Burger, & Jensen, 2019) and for the importance of including psychometrically sound and conceptually cohesive questions about treatment expectations and past treatment experiences in baseline patient assessments (e.g. Barth, Kern, Lüthi, & Witt, 2019).

The finding that recent stressors were important is broadly consistent with evidence documenting effects of stressful life experiences on depression treatment response (Buckman *et al.*, 2022). The fact that financial stress was the second most important predictor was especially striking and is consistent with prior studies showing that unemployment and low household income are top predictors of low ADM treatment response (Lee *et al.*, 2018). The findings that baseline protective/resilience factors were important are also in line with much previous research (Buckman *et al.*, 2021c; Laird, Lavretsky, St Cyr, & Siddarth, 2018). The fact that some of these associations were nonmonotonic is consistent with naturalistic evidence that moderate, compared to extremely low or high, levels of emotional reactivity to stress predict low future depression severity (Santee & Starr, 2021) and that baseline self-reported resilience is sometimes significant in predicting depression treatment response only in interaction with other predictors (Choi *et al.*, 2021; Min, Lee, Lee, Lee, & Chae, 2012). These specifications might reflect the greater importance of protective/resilience factors in the subset of patients whose depressive episodes are triggered by stressful life experiences, which could be the subject of future investigation (Chromik, 2021).

Limitations

The study had several noteworthy limitations. Three of these involve external validity. First, the baseline response rate was low, although comparable to response rates in other VHA studies examining mental health outcomes (King, Beehler, Buchholz, Johnson, & Wray, 2019; Stolzmann *et al.*, 2019). However, as shown in a previous report (Puac-Polanco *et al.*, 2021), there are minimal differences between our responders and nonresponders on baseline administrative variables and equally modest differences in baseline self-reports between patients followed *v.* lost to follow-up, although response bias might nonetheless exist with respect to unmeasured variables. Second, we did not account for possible disruptions in care due to the COVID-19 pandemic, which involved 7.6% of study patients who completed assessments after February 2020. Third, although the model was validated in a separate test sample, it was not tested in an external validation sample. Nor is it clear whether findings would generalize to non-VHA patients.

A separate set of limitations involve design decisions that could have biased results. One of these is that study recruitment and assessment occurred only after the initial visit, during which time symptoms might have decreased, leading to distortion in our estimates of associations between baseline symptoms and treatment response. As reported above, the association between time between initiating treatment and completing the baseline assessment was unrelated to baseline QIDS-SR scores, somewhat reducing this concern, but it is nonetheless important that future replications and extensions of our work are carried out with baseline assessment administered before treatment selection is made. Another limitation that might have biased results was the use of a very large set of predictors, which could have resulted in overfitting even though we used procedures to minimize this possibility.

A final set of noteworthy limitations involves the measures. The predictors excluded information about military experiences that might have led to the depression, and the outcomes were based on self-reports rather than clinical interviews.

Strengths

The study also had several strengths, including an observational sample with greater external validity than clinical trial samples, a rich baseline predictor set that included a wide range of variables found in previous research to be prognostic predictors of depression treatment response, and use of a rigorous ML method to develop the model.

Conclusions

Within the context of these limitations, we found that a model to predict ADM treatment response could be developed based largely on a battery of self-report questions along with some administrative variables from EHRs and geospatial databases. The model had modest overall prediction strength but nonetheless provided enough discrimination across three broad groups of patients to have potential value in informing depressed patients pre-emptively about their likelihood of responding to ADM as part of a patient-centered shared decision-making process. The model had good calibration and fairness with respect to key indicators of health disparities. Our findings would need to be replicated in a sample where the baseline assessment occurred before the beginning of treatment, the model streamlined, and parallel models built for predicted response to other types of treatment before results could be useful. In addition, parallel models combined across different treatments would be needed to determine best treatment options for particular patients (Kessler & Luedtke, 2021).

Supplementary material. The supplementary material for this article can be found at <https://doi.org/10.1017/S0033291722001982>

Financial support. This research was supported by the Office of Mental Health Services and Suicide Prevention and Center of Excellence for Suicide Prevention (Bossarte), the National Institute of Mental Health of the National Institutes of Health (R01MH121478, Kessler), the United States Department of Veterans Affairs Health Services Research & Development Service Career Development Award (IK2 HX002867, Leung), the PCORI Project Program Award (ME-2019C1-16172, Zubizarreta), and the Advanced Fellowship from the VISN 4 Mental Illness Research, Education, & Clinical Center (MIRECC, Cui, Oslin). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, the US Department of Veterans Affairs, or the United States Government.

Conflict of interest. In the past 3 years, Dr Kessler was a consultant for Datastat, Inc., Holmusk, RallyPoint Networks, Inc., and Sage Therapeutics. He has stock options in Mirah, PYM, and Roga Sciences. Dr Pigeon consulted for CurAegis Technologies and received clinical trial support from Pfizer, Inc. and Abbvie, Inc. Dr Zubizarreta consulted for Johnson & Johnson Real World Data Analytics. Dr Cipriani is supported by the National Institute for Health Research (NIHR) Oxford Cognitive Health Clinical Research Facility, by an NIHR Research Professorship (grant RP-2017-08-ST2-006), by the NIHR Oxford and Thames Valley Applied Research Collaboration and by the NIHR Oxford Health Biomedical Research Centre (grant BRC-1215-20005); he has also received research, educational, and consultancy fees from INCiPiT (Italian Network for Paediatric Trials), CARIPLO Foundation and Angelini Pharma. The views expressed are those of the authors and not necessarily those of the UK National Health Service, the NIHR, or the UK Department of Health. The remaining authors report no conflict of interest.

Ethical standards. All procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

Notes

- 1 See online Supplementary Table S8 for descriptions of the predictor labels.
2 See online Supplementary Table S8 for descriptions of the predictor labels.

¹Department of Health Care Policy, Harvard Medical School, Boston, MA, USA; ²Department of Psychiatry, McLean Hospital, Belmont, MA, USA; ³Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA; ⁴Department of Psychiatry, Harvard Medical School, Boston, MA, USA; ⁵Center of Excellence for Suicide Prevention, Canandaigua VA Medical Center, Canandaigua, NY, USA; ⁶Harvard T.H. Chan School of Public Health, Boston, MA, USA; ⁷Department of Veterans Affairs, VISN 4 Mental Illness Research, Education and Clinical Center, VA Pittsburgh Health Care System, Pittsburgh, PA, USA; ⁸Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA; ⁹Center for the Study of Healthcare Innovation, Implementation, and Policy, VA Greater Los Angeles Healthcare System, Los Angeles, CA, USA; ¹⁰Division of General Internal Medicine and Health Services Research, UCLA David Geffen School of Medicine, Los Angeles, CA, USA; ¹¹Department of Behavioral Medicine and Psychiatry, West Virginia University, Morgantown, WV, USA; ¹²Center for Clinical Management Research, VA Ann Arbor, Ann Arbor, MI, USA; ¹³Department of Psychology, Yale University, New Haven, CT, USA; ¹⁴Department of Psychiatry, Dauten Family Center for Bipolar Treatment Innovation, Massachusetts General Hospital, Boston, MA, USA; ¹⁵VISN 4 Mental Illness Research, Education, and Clinical Center, Corporal Michael J. Crescenzo VA Medical Center, Philadelphia, PA, USA; ¹⁶Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA; ¹⁷Department of Psychiatry, University of Rochester Medical Center, Rochester, NY, USA; ¹⁸Department of Medicine, University of Michigan Medical School, Ann Arbor, MI, USA; ¹⁹Department of Statistics, Harvard University, Cambridge, MA, USA; ²⁰Department of Biostatistics, Harvard University, Cambridge, MA, USA; ²¹Department of Statistics, University of Washington, Seattle, WA, USA; ²²Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA; ²³Department of Psychiatry, University of Oxford, Oxford, UK and ²⁴Department of Health Promotion and Human Behavior, School of Public Health, Kyoto University Graduate School of Medicine, Kyoto, Japan

References

- Austin, P. C., & Steyerberg, E. W. (2014). Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Statistics in Medicine*, 33(3), 517–535. doi: 10.1002/sim.5941
- Austin, P. C., & Steyerberg, E. W. (2019). The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Statistics in Medicine*, 38(21), 4051–4065. doi: 10.1002/sim.8281
- Barth, J., Kern, A., Lüthi, S., & Witt, C. M. (2019). Assessment of patients' expectations: Development and validation of the Expectation for Treatment Scale (ETS). *BMJ Open*, 9(6), e026712. doi: 10.1136/bmjopen-2018-026712
- Benazzi, F. (2006). Various forms of depression. *Dialogues in Clinical Neuroscience*, 8(2), 151–161. doi: 10.31887/DCNS.2006.8.2/fbenazzi
- Buckman, J. E. J., Cohen, Z. D., O'Driscoll, C., Fried, E. I., Saunders, R., Ambler, G., ... Pilling, S. (2021a). Predicting prognosis for adults with depression using individual symptom data: A comparison of modelling approaches. *Psychological Medicine*, Advance online publication. doi: 10.1017/S0033291721001616
- Buckman, J. E. J., Saunders, R., Arundell, L. L., Oshinowo, I. D., Cohen, Z. D., O'Driscoll, C., ... Pilling, S. (2022). Life events and treatment prognosis for depression: A systematic review and individual patient data meta-analysis. *Journal of Affective Disorders*, 299, 298–308. doi: 10.1016/j.jad.2021.12.030
- Buckman, J. E. J., Saunders, R., Cohen, Z. D., Barnett, P., Clarke, K., Ambler, G., ... Pilling, S. (2021b). The contribution of depressive 'disorder characteristics' to determinations of prognosis for adults with depression: An individual patient data meta-analysis. *Psychological Medicine*, 51(7), 1068–1081. doi: 10.1017/S0033291721001367
- Buckman, J. E. J., Saunders, R., O'Driscoll, C., Cohen, Z. D., Stott, J., Ambler, G., ... Pilling, S. (2021c). Is social support pre-treatment associated with prognosis for adults with depression in primary care? *Acta Psychiatrica Scandinavica*, 143(5), 392–405. doi: 10.1111/acps.13285
- Chekroud, A. M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., ... Choi, K. (2021). The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry*, 20(2), 154–170. doi: 10.1002/wps.20882
- Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), 266–298. doi: 10.1214/09-AOAS285
- Choi, W., Kim, J. W., Kang, H. J., Kim, H. K., Kang, H. C., Lee, J. Y., ... Kim, J. M. (2021). Synergistic effects of resilience and serum ghrelin levels on the 12-week pharmacotherapeutic response in patients with depressive disorders. *Journal of Affective Disorders*, 295, 1489–1493. doi: 10.1016/j.jad.2021.09.039
- Chromik, M. (2021). Making SHAP Rap: Bridging local and global insights through interaction and narratives. In C. Ardito, R. Lanzilotti, A. Malizia, H. Petrie, A. Piccinno, G. Desolda & K. Inkpen (Eds.), *Human-computer interaction - INTERACT 2021* (pp. 641–651). Cham: Springer.
- Cipriani, A., Furukawa, T. A., Salanti, G., Chaimani, A., Atkinson, L. Z., Ogawa, Y., ... Geddes, J. R. (2018). Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: A systematic review and network meta-analysis. *Lancet*, 391(10128), 1357–1366. doi: 10.1016/S0140-6736(17)32802-7
- Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD statement. *The British Journal of Surgery*, 102(3), 148–158. doi: 10.1002/bjs.9736
- Constantino, M. J., Visla, A., Coyne, A. E., & Boswell, J. F. (2018). A meta-analysis of the association between patients' early treatment outcome expectation and their posttreatment outcomes. *Psychotherapy*, 55(4), 473–485. doi: 10.1037/pst0000169
- Cuijpers, P., Noma, H., Karyotaki, E., Vinkers, C. H., Cipriani, A., & Furukawa, T. A. (2020). A network meta-analysis of the effects of psychotherapies, pharmacotherapies and their combination in the treatment of adult depression. *World Psychiatry*, 19(1), 92–107. doi: 10.1002/wps.20701
- Day, E., Shah, R., Taylor, R. W., Marwood, L., Nortey, K., Harvey, J., ... Strawbridge, R. (2021). A retrospective examination of care pathways in individuals with treatment-resistant depression. *British Journal of Psychiatry Open*, 7(3), e101–e101. doi: 10.1192/bjo.2021.59
- Erners, N. J., Hagoort, K., & Scheepers, F. E. (2020). The predictive validity of machine learning models in the classification and treatment of major depressive disorder: State of the art and future directions. *Frontiers in Psychiatry*, 11, 472. doi: 10.3389/fpsy.2020.00472
- Furukawa, T. A., Sukanuma, A., Ostinelli, E. G., Andersson, G., Beevers, C. G., Shumake, J., ... Cuijpers, P. (2021). Dismantling, optimising and personalising internet cognitive behavioural therapy for depression: A systematic review and component network meta-analysis using individual participant data. *The Lancet Psychiatry*, 8(6), 500–511. doi: 10.1016/S2215-0366(21)00077-8
- GBD 2019 Diseases and Injuries Collaborators. (2020). Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*, 396(10258), 1204–1222. doi: 10.1016/S0140-6736(20)30925-9
- Greden, J. F., Parikh, S. V., Rothschild, A. J., Thase, M. E., Dunlop, B. W., DeBattista, C., ... Dechairo, B. (2019). Impact of pharmacogenomics on clinical outcomes in major depressive disorder in the GUIDED trial: A large, patient- and rater-blinded, randomized, controlled study. *Journal of Psychiatric Research*, 111, 59–67. doi: 10.1016/j.jpsychires.2019.01.003
- Gruszka, P., Burger, C., & Jensen, M. P. (2019). Optimizing expectations via mobile apps: A new approach for examining and enhancing placebo effects. *Frontiers in Psychiatry*, 10, 365. doi: 10.3389/fpsy.2019.00365
- Hockenberry, J. M., Joski, P., Yarbrough, C., & Druss, B. G. (2019). Trends in treatment and spending for patients receiving outpatient treatment of depression in the United States, 1998–2015. *JAMA Psychiatry*, 76(8), 810–817. doi: 10.1001/jamapsychiatry.2019.0633
- Huang, J., Wang, Y., Chen, J., Zhang, Y., Yuan, Z., Yue, L., ... Fang, Y. (2018). Clinical outcomes of patients with major depressive disorder treated with either duloxetine, escitalopram, fluoxetine, paroxetine, or sertraline. *Neuropsychiatric Disease and Treatment*, 14, 2473–2484. doi: 10.2147/ndt.s159800
- Karrer, T. M., Bassett, D. S., Derntl, B., Gruber, O., Aleman, A., Jardri, R., ... Bzdok, D. (2019). Brain-based ranking of cognitive domains to predict

- schizophrenia. *Human Brain Mapping*, 40(15), 4487–4507. doi: 10.1002/hbm.24716
- Katz, I. R., Liebmann, E. P., Resnick, S. G., & Hoff, R. A. (2021). Performance of the PHQ-9 across conditions and comorbidities: Findings from the Veterans Outcome Assessment survey. *Journal of Affective Disorders*, 294, 864–867. doi: 10.1016/j.jad.2021.07.108
- Kazdin, A. E., Wu, C.-S., Hwang, I., Puac-Polanco, V., Sampson, N. A., Al-Hamzawi, A., ... Kessler, R. C. (2021). Antidepressant use in low-middle- and high-income countries: A World Mental Health Surveys report. *Psychological Medicine*, Advance online publication. doi: 10.1017/S0033291721003160
- Kessler, R. C., & Luedtke, A. (2021). Pragmatic precision psychiatry – A new direction for optimizing treatment selection. *JAMA Psychiatry*, 78(12), 1384–1390. doi: 10.1001/jamapsychiatry.2021.2500
- King, P. R., Beehler, G. P., Buchholz, L. J., Johnson, E. M., & Wray, L. O. (2019). Functional concerns and treatment priorities among veterans receiving VHA Primary Care Behavioral Health services. *Families, Systems & Health*, 37(1), 68–73. doi: 10.1037/fsh0000393
- Kraus, C., Kadriu, B., Lanzemberger, R., Zarate Jr, C. A., & Kasper, S. (2019). Prognosis and improved outcomes in major depression: A review. *Translational Psychiatry*, 9(1), 127. doi: 10.1038/s41398-019-0460-3
- Laferton, J. A. C., Kube, T., Salzmann, S., Auer, C. J., & Shedden-Mora, M. C. (2017). Patients' expectations regarding medical treatment: A critical review of concepts and their assessment. *Frontiers in Psychology*, 8, 233–233. doi: 10.3389/fpsyg.2017.00233
- Laird, K. T., Lavretsky, H., St Cyr, N., & Siddarth, P. (2018). Resilience predicts remission in antidepressant treatment of geriatric depression. *International Journal of Geriatric Psychiatry*, 33(12), 1596–1603. doi: 10.1002/gps.4953
- Larson, S., Nemoianu, A., Lawrence, D. F., Troup, M. A., Gionfriddo, M. R., Pousti, B., ... Touya, M. (2021). Characterizing primary care for patients with major depressive disorder using electronic health records of a US-based healthcare provider. *Journal of Affective Disorders*, 300, 377–384. doi: 10.1016/j.jad.2021.12.096
- LeDell, E., van der Laan, M. J., & Petersen, M. (2016). AUC-maximizing ensembles through Metalearning. *The International Journal of Biostatistics*, 12(1), 203–218. doi: 10.1515/ijb-2015-0035
- Lee, Y., Ragugett, R. M., Mansur, R. B., Boutilier, J. J., Rosenblat, J. D., Trevizol, A., ... McIntyre, R. S. (2018). Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. *Journal of Affective Disorders*, 241, 519–532. doi: 10.1016/j.jad.2018.08.073
- Leeuwenberg, A. M., van Smeden, M., Langendijk, J. A., van der Schaaf, A., Mauer, M. E., Moons, K. G. M., ... Schuit, E. (2022). Performance of binary prediction models in high-correlation low-dimensional settings: A comparison of methods. *Diagnostic and Prognostic Research*, 6(1), 1. doi: 10.1186/s41512-021-00115-5
- Leon, A. C., Olfson, M., Portera, L., Farber, L., & Sheehan, D. V. (1997). Assessing psychiatric impairment in primary care with the Sheehan Disability Scale. *International Journal of Psychiatry in Medicine*, 27(2), 93–105. doi: 10.2190/t8em-c8yh-373n-luwd
- Lindhiem, O., Petersen, I. T., Mentch, L. K., & Youngstrom, E. A. (2020). The importance of calibration in clinical psychology. *Assessment*, 27(4), 840–854. doi: 10.1177/1073191117752055
- Little, A. (2009). Treatment-resistant depression. *American Family Physician*, 80(2), 167–172.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Retrieved from <https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf>.
- Maj, M., Stein, D. J., Parker, G., Zimmerman, M., Fava, G. A., De Hert, M., ... Wittchen, H. U. (2020). The clinical characterization of the adult patient with depression aimed at personalization of management. *World Psychiatry*, 19(3), 269–293. doi: 10.1002/wps.20771
- McHugh, R. K., Whitton, S. W., Peckham, A. D., Welge, J. A., & Otto, M. W. (2013). Patient preference for psychological vs pharmacologic treatment of psychiatric disorders: A meta-analytic review. *The Journal of Clinical Psychiatry*, 74(6), 595–602. doi: 10.4088/JCP.12r07757
- Min, J. A., Lee, N. B., Lee, C. U., Lee, C., & Chae, J. H. (2012). Low trait anxiety, high resilience, and their interaction as possible predictors for treatment response in patients with depression. *Journal of Affective Disorders*, 137(1–3), 61–69. doi: 10.1016/j.jad.2011.12.026
- Noma, H., Furukawa, T. A., Maruo, K., Imai, H., Shinohara, K., Tanaka, S., ... Cipriani, A. (2019). Exploratory analyses of effect modifiers in the antidepressant treatment of major depression: Individual-participant data meta-analysis of 2803 participants in seven placebo-controlled randomized trials. *Journal of Affective Disorders*, 250, 419–424. doi: 10.1016/j.jad.2019.03.031
- Park, T., & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(482), 681–686. doi: 10.1198/01621450800000337
- Perlman, K., Benrimoh, D., Israel, S., Rollins, C., Brown, E., Tunteng, J. F., ... Berlim, M. T. (2019). A systematic meta-review of predictors of antidepressant treatment outcome in major depressive disorder. *Journal of Affective Disorders*, 243, 503–515. doi: 10.1016/j.jad.2018.09.067
- Perna, G., Alciati, A., Daccò, S., Grassi, M., & Caldirola, D. (2020). Personalized psychiatry and depression: The role of sociodemographic and clinical variables. *Psychiatry Investigation*, 17(3), 193–206. doi: 10.30773/pi.2019.0289
- Polley, E. C., Rose, S., & van der Laan, M. J. (2011). Super learning. In M. J. van der Laan & S. Rose (Eds.), *Targeted learning: Casual inference for observational and experimental data* (pp. 43–66). New York: Springer.
- Polley, E., LeDell, E., Kennedy, C., Lendle, S., & van der Laan, M. J. (2021). Superlearner: Super learner prediction, version 2.0-28. Retrieved from <https://CRAN.R-project.org/package=SuperLearner>.
- Puac-Polanco, V., Leung, L. B., Bossarte, R. M., Bryant, C., Keusch, J. N., Liu, H., ... Kessler, R. C. (2021). Treatment differences in primary and specialty settings in veterans with major depression. *Journal of the American Board of Family Medicine*, 34(2), 268–290. doi: 10.3122/jabfm.2021.02.200475
- Qaseem, A., Barry, M. J., & Kansagara, D. (2016). Nonpharmacologic versus pharmacologic treatment of adult patients with major depressive disorder: A clinical practice guideline from the American College of Physicians. *Annals of Internal Medicine*, 164(5), 350–359. doi: 10.7326/m15-2570
- R Core Team. (2021). R: A language and environment for statistical computing. Retrieved from <https://www.R-project.org/>.
- Rush, A. J., & Thase, M. E. (2018). Improving depression outcome by patient-centered medical management. *American Journal of Psychiatry*, 175(12), 1187–1198. doi: 10.1176/appi.ajp.2018.18040398
- Rush, A. J., Trivedi, M. H., Ibrahim, H. M., Carmody, T. J., Arnow, B., Klein, D. N., ... Keller, M. B. (2003). The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): A psychometric evaluation in patients with chronic major depression. *Biological Psychiatry*, 54(5), 573–583. doi: 10.1016/s0006-3223(02)01866-8
- Santee, A. C., & Starr, L. R. (2021). Examining linear and nonlinear associations between negative emotional reactivity to daily events and depression among adolescents. *Clinical Psychological Science*, Advance online publication. doi: 10.1177/21677026211045684
- SAS Institute Inc. (2013). *SAS software* (9.4 ed.). Cary, NC: SAS Institute Inc.
- Stolzmann, K., Meterko, M., Miller, C. J., Belanger, L., Seibert, M. N., & Bauer, M. S. (2019). Survey response rate and quality in a mental health clinic population: Results from a randomized survey comparison. *The Journal of Behavioral Health Services & Research*, 46(3), 521–532. doi: 10.1007/s11414-018-9617-8
- Wang, G., You, X., Wang, X., Xu, X., Bai, L., Xie, J., ... Hu, C. (2018). Safety and effectiveness of escitalopram in an 8-week open study in Chinese patients with depression and anxiety. *Neuropsychiatric Disease and Treatment*, 14, 2087–2097. doi: 10.2147/ndt.s164673
- Yuan, M., Kumar, V., Ahmad, M. A., & Teredesai, A. (2021). Assessing fairness in classification parity of machine learning models in healthcare. Retrieved from <https://arxiv.org/abs/2102.03717>.
- Zilcha-Mano, S., Wang, X., Wajsbrot, D. B., Boucher, M., Fine, S. A., & Rutherford, B. R. (2021). Trajectories of function and symptom change in Desvenlafaxine clinical trials: Toward personalized treatment for depression. *Journal of Clinical Psychopharmacology*, 41(5), 579–584. doi: 10.1097/jcp.0000000000001435
- Ziobrowski, H. N., Kennedy, C. J., Ustun, B., House, S. L., Beaudoin, F. L., An, X., ... van Rooij, S. J. H. (2021a). Development and validation of a model to

- predict posttraumatic stress disorder and major depression after a motor vehicle collision. *JAMA Psychiatry*, 78(11), 1228–1237. doi: 10.1001/jamapsychiatry.2021.2427
- Ziobrowski, H. N., Leung, L. B., Bossarte, R. M., Bryant, C., Keusch, J. N., Liu, H., ... Kessler, R. C. (2021b). Comorbid mental disorders, depression symptom severity, and role impairment among Veterans initiating depression treatment through the Veterans Health Administration. *Journal of Affective Disorders*, 290, 227–236. doi: 10.1016/j.jad.2021.04.033
- Zou, G. (2004). A modified Poisson regression approach to prospective studies with binary data. *American Journal of Epidemiology*, 159(7), 702–706. doi: 10.1093/aje/kwh090
- Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511), 910–922. doi: 10.1080/01621459.2015.1023805
- Zubizarreta, J. R., Li, Y., Allouah, A., & Greifer, N. (2021). sbw: Stable balancing weights for causal inference and estimation with incomplete outcome data (Version 1.1.1). Retrieved from <https://cran.rstudio.com/web/packages/sbw/>.