Psychological Medicine

cambridge.org/psm

Development of a model to predict psychotherapy response for depression among Veterans

Original Article

Cite this article: Ziobrowski HN *et al* (2023). Development of a model to predict psychotherapy response for depression among Veterans. *Psychological Medicine* **53**, 3591–3600. https://doi.org/10.1017/S0033291722000228

Received: 24 November 2021
Revised: 14 January 2022
Accepted: 17 January 2022
First published online: 11 February 2022

Keywords:
Depression; machine learning; psychotherapy; treatment response; Veterans Health Administration

Author for correspondence:
Ronald C. Kessler,
E-mail: kessler@hcp.med.harvard.edu

Hannah N. Ziobrowski[1], Ruifeng Cui[2,3], Eric L. Ross[4,5,6], Howard Liu[1,7], Victor Puac-Polanco[1], Brett Turner[1,8], Lucinda B. Leung[9,10], Robert M. Bossarte[7,11], Corey Bryant[12], Wilfred R. Pigeon[7,13], David W. Oslin[14,15], Edward P. Post[12,16], Alan M. Zaslavsky[1], Jose R. Zubizarreta[1,17,18], Andrew A. Nierenberg[6,19], Alex Luedtke[20,21], Chris J. Kennedy[22] and Ronald C. Kessler[1]

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

Abstract

Background. Fewer than half of patients with major depressive disorder (MDD) respond to psychotherapy. Pre-emptively informing patients of their likelihood of responding could be useful as part of a patient-centered treatment decision-support plan.

Methods. This prospective observational study examined a national sample of 807 patients beginning psychotherapy for MDD at the Veterans Health Administration. Patients completed a self-report survey at baseline and 3-months follow-up (data collected 2018–2020). We developed a machine learning (ML) model to predict psychotherapy response at 3 months using baseline survey, administrative, and geospatial variables in a 70% training sample. Model performance was then evaluated in the 30% test sample.

Results. 32.0% of patients responded to treatment after 3 months. The best ML model had an AUC (SE) of 0.652 (0.038) in the test sample. Among the one-third of patients ranked by the model as most likely to respond, 50.0% in the test sample responded to psychotherapy. In comparison, among the remaining two-thirds of patients, <25% responded to psychotherapy. The model selected 43 predictors, of which nearly all were self-report variables.

Conclusions. Patients with MDD could pre-emptively be informed of their likelihood of responding to psychotherapy using a prediction tool based on self-report data. This tool could meaningfully help patients and providers in shared decision-making, although parallel information about the likelihood of responding to alternative treatments would be needed to inform decision-making across multiple treatments.

Introduction

Depression is a leading cause of disability worldwide (GBD, 2019 Diseases & Injuries Collaborators, 2020) and is associated with great psychological, physical, and economic burden (Herrman et al., in press). The primary first-line treatments for depression are psychotherapy, anti-depressant medication, and their combination (Qaseem, Barry, & Kansagara, 2016). Most patients initiating depression treatment prefer psychotherapy (Gelhorn, Sexton, & Classi, 2011; Leung et al., 2021; van Schaik et al., 2004), but psychotherapy is substantially more expensive

CAMBRIDGE UNIVERSITY PRESS

https://doi.org/10.1017/S0033291722000228 Published online by Cambridge University Press

and time-consuming than ADM (Koeser, Donisi, Goldberg, & McCrone, 2015; Ross, Vijan, Miller, Valenstein, & Zivin, 2019) and is also less accessible than ADM due to the fact that primary care physicians can prescribe ADMs but psychotherapy requires access to a mental health specialist.

Fewer than half of patients respond to psychotherapy alone (Blais et al., 2013; Cuijpers et al., 2021). As randomized trials show that psychotherapy and ADM have comparable aggregate effects when used alone and that combined treatment has better aggregate effects than either alone (Cuijpers et al., 2013; Kappelmann et al., 2020), the typical response to treatment nonresponse with psychotherapy would be either augmentation with an ADM or switching to an ADM. However, as depression treatment selection typically works through trial and error, patients who begin with psychotherapy often spend weeks or months trying this treatment before determining that the treatment is not working (Blais et al., 2013), at which time they can either augment or switch to ADM if they have not already dropped out of treatment. A strategy to predict patients' likelihood of responding to psychotherapy before the beginning of treatment could help avoid these delays. Such a strategy, if it could be developed, might help reduce treatment dropout and facilitate more rapid receipt of helpful treatments for patients who are unlikely to respond to psychotherapy.

Previous research has documented diverse baseline risk factors that consistently predict psychotherapy response, such as depression severity and subtypes and prior history, psychiatric comorbidity, and stressful life experiences (e.g. Bone *et al.* 2021; Coley, Boggs, Beck, & Simon, 2021; Serbanescu *et al.* 2020). However, none of these associations is strong enough to be used as a primary basis for treatment planning. Based on this fact, researchers have examined whether multivariable models that combine information across a range of individually significant predictors can improve the prediction of psychotherapy treatment response (DeRubeis et al., 2014; Huibers et al., 2015; Saunders et al., 2021). However, such models risk overfitting (van Klaveren, Balan, Steyerberg, & Kent, 2019). Machine learning (ML) methods can help protect against overfitting (Roelofs et al., 2019). Although some ML studies of psychotherapy treatment response have been carried out (Coley et al., 2021; Pearson, Pisner, Meyer, Shumake, & Beevers, 2019; Tymofiyeva et al., 2019), most were based on secondary analyses of randomized clinical trials. The latter studies typically had limited predictor sets and reduced external validity because patients unwilling to be randomized or with psychiatric comorbidities were excluded. Observational samples resolve these problems, but the few studies that tried to predict depression psychotherapy treatment response in observational samples either had limited predictor sets (Bone et al., 2021; Delgadillo & Gonzalez Salas Duhne, 2020; Tymofiyeva et al., 2019), used predictors that would not be possible to collect in a routine clinical visit (Tymofiyeva et al., 2019), or based predictions only on administrative data (Coley et al., 2021).

The current report presents the results of a study designed to address the above limitations by collecting information on a rich baseline set of potential predictors from a self-report assessment and administrative data in a prospective observational sample of Veterans Health Administration (VHA) patients initiating treatment for major depressive disorder (MDD). Patients were followed for 3 months to assess treatment response. The VHA provides a unique opportunity to study MDD treatment response because it is the largest national US healthcare delivery system integrating mental health services into primary care (Leung et al., 2019). We focus here on baseline variables known or hypothesized to predict psychotherapy treatment response. We aimed to develop a parsimonious model with a small number of predictors that could feasibly be administered in routine clinical practice.

Methods

Sample

We recruited eligible VHA patients from weekly nationally representative probability samples between December 2018 and June 2020. As we aimed to focus on incident treatment encounters, we excluded patients who in the prior 12 months before the focal visit received any MDD treatment or attempted suicide. Outpatient settings included primary care and specialty mental health clinics. Patients either had to receive a prescription for an ADM or a referral to psychotherapy in the focal visit to be eligible. Focal visits were not counted as eligible if the record noted that the patient was depressed but that watchful waiting was being used rather than treatment. The present report considers only the subset of eligible patients who were referred to psychotherapy but did not receive an ADM prescription in the focal visit. We additionally excluded patients who had any lifetime diagnosis of bipolar disorder, nonaffective psychosis, dementia, intellectual disabilities, autism, Tourette's disorder, stereotyped movement disorders, or borderline intellectual functioning, or ever received a prescription of either antimanic or antipsychotic medication (see Online Supplementary Table S1 for ICD-9-CM and ICD-10-CM codes).

As described in more detail elsewhere (Puac-Polanco et al., 2021) and shown in Online Supplementary Fig. S1, recruitment letters were mailed to 55 106 eligible patients inviting them to participate in a study of depression treatment that would require completing one self-report web- or phone-based survey at baseline (taking approximately 45 min) and another self-report survey at 3-months follow-up (taking approximately 20 min). Patients received up to three recruitment calls over the next week. A total of 17 000 patients were reached within this period, 6298 of whom agreed to participate and 4164 completed the baseline survey. Of these patients, 1554 were excluded after completing the baseline survey because they either reported being actively suicidal, did not report depression as a presenting problem, reported mania as a presenting problem, or did not report depression severity equal to at least 6 on the Quick Inventory of Depression Symptomatology Self-Report (QIDS-SR; (Rush et al., 2003)). As reported previously (Puac-Polanco et al., 2021), patients who completed the baseline questionnaire were, on average, slightly older than non-respondents and somewhat more likely to be female, non-Hispanic White, and currently married (with odds-ratios ranging between 1.2 and 1.7), but multivariate associations with participation were weak [area under the receiver operating characteristic curve (AUC) = 0.59].

Among the remaining 2609 baseline respondents, 989 received psychotherapy without ADM and 807 of the latter completed the 3-month follow-up survey. These are the patients included in the present report. Patients were compensated $50 and $25 for completing the baseline and 3-month surveys, respectively. The Institutional Review Board of Syracuse VA Medical Center, Syracuse, New York, approved these procedures. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional

https://doi.org/10.1017/S0033291722000228 Published online by Cambridge University Press

committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. We followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) reporting guidelines (Collins, Reitsma, Altman, & Moons, 2015) for reporting analyses designed to develop predictive models.

Measures

Treatment response

Self-reports of depressive symptom severity and role impairment were assessed at baseline and 3-months. Depressive symptoms were assessed with the 16-item QIDS-SR (Rush et al., 2003) (Cronbach's $\alpha = 0.675$), which asks about symptom severity in the past 2 weeks using a 0–3 response scale with embedded labels for each item (e.g. between *not feel sad* and *sad nearly all the time* for depressed mood). Role impairment due to depression was assessed with a modified version of the Sheehan Disability Scale (Leon, Olfson, Portera, Farber, & Sheehan, 1997), in which patients rated how much their depression interfered with their ability to work, participate in family and home life, and participate in social activities in the past 2 weeks on a labeled 0–10 visual analog scale of *not at all* (0), *mildly* (1–3), *moderately* (4–6), *markedly* (7–9), and *extremely* (10) (Cronbach's $\alpha = 0.847$).

Patients were classified as responding to treatment if they met any of the following criteria in their 3-month follow-up assessments: (1) had a QIDS-SR score of 0–5 (indicating 'remission' of depressive symptoms (Rush et al., 2006)), (2) their QIDS-SR score was half or less of its baseline value, or (3) they had a baseline score of 4 or more (i.e. moderate-severe) in *one or more* role impairment domains and a 3-month score of 0–3 (i.e. none-mild) on *all* role impairment domains.

Predictors

A recent review by Maj et al. (2020) recommended considering 14 risk factor domains to personalize depression treatment: symptom profile, clinical subtypes, severity, clinical staging, early environmental exposures, recent environmental stressors, family history, functioning and quality of life, physical comorbidities, personality, antecedent and concomitant psychiatric conditions, protective factors/resilience, neurocognition, and dysfunctional cognitive schemas. We included predictors from each of these 14 domains as well as from two additional domains that have been shown in previous research to predict depression treatment response: sociodemographics and treatment characteristics, the latter including information about prior treatment history as well as current expectations and preferences. In total, 2810 baseline predictors were included in our analysis, derived from the baseline survey, administrative records, and geospatial data based on patient residence (Online Supplementary Tables S2–S4). These 2810 predictors included transformations of the same variables, as described in more detail in Online Supplementary Tables S2–S4. Categorical variables were indicator-coded with dummy variables. Quantitative variables were standardized to a mean of 0 and variance of 1 for use in linear algorithms (described below) and discretized into ventiles for use in tree-based algorithms (described below) to avoid overfitting and to reduce computation time.

Analysis methods

Analysis was limited to patients who completed both the baseline and 3-month follow-up self-report surveys. As detailed in a previous report (Puac-Polanco et al., 2021), we used the R program *sbw* (Zubizarreta, Li, Allouah, & Greifer, 2021) to account for potential selection bias from nonresponse by creating stable balancing weights to adjust for significant differences between baseline respondents and the full target sample on significant predictors of non-response (Zubizarreta, 2015). This procedure was then repeated in the weighted follow-up sample, applying a second *sbw* to adjust for discrepancies in baseline survey predictors between respondents in the follow-up sample and those lost to follow-up. These doubly weighted data were used in the analysis.

A ML model was developed based on the doubly weighted data to predict psychotherapy response. Rather than use a single algorithm, which has been done in previous ML studies of MDD treatment response, we used the Super Learner (SL) stacked generalization method (Polley, LeDell, Kennedy, Lendle, & van der Laan, 2021, May 10) to pool results across a library of multiple algorithms. This was done using a weight for each algorithm derived from a training sample (described below) via 10-fold cross-validation. The composite predicted outcome score based on this weight is guaranteed in expectation to perform at least as well as the best component algorithm in the library in terms of a prespecified criterion (Polley, Rose, & van der Laan, 2011), which in our case was non-negative least squares. Consistent with recommendations (LeDell, van der Laan, & Petersen, 2016; Naimi & Balzer, 2018), we included a diverse set of algorithms in the library to capture nonlinearities and interactions and reduce the risk of model misspecification (Kabir & Ludwig, 2019). These included linear algorithms (logistic regression, regularized regression, spline and polynomial spline regressions, support vector machines) and tree-based algorithms (boosting and bagging ensemble trees and Bayesian additive regression trees) (Online Supplementary Table S5). Similar stacking procedures have been used in prior computational psychiatric studies (Karrer et al., 2019; Ziobrowski et al., 2021a).

Hyperparameters were tuned by including individual algorithms multiple times in the library with different hyperparameter values. This tuning method allowed SL to weight relative importance across this range rather than using an external grid search or random search method. Feature selection was independently conducted in each 10-fold cross-validation training sample. To increase the feasibility of implementation in clinical practice and to reduce overfitting, we explored 2 feature reduction methods: least absolute shrinkage and selection operator penalized regression (lasso; Park and Casella, 2008) and variable importance ranking from Bayesian additive regression trees (BART; Chipman, George, and McCulloch, 2010). We then compared the predictive accuracy of the SL with a simpler lasso penalized regression model to see how much, if at all, SL improved prediction.

Models were estimated in a training sample that included 70% of patients selected with stratification to have the same distribution on a wide range of predictors and the outcome as the total sample. The remaining 30% of patients were used to evaluate model accuracy. We used a locally estimated scatterplot smoothed calibration curve (Austin & Steyerberg, 2014) to quantify calibration of predicted outcome probabilities from our best model in the test sample using the integrated calibration index (ICI) and expected calibration error (ECE) (Austin & Steyerberg, 2019; Naeini, Cooper, & Hauskrecht, 2015).

Model evaluation in the 30% test sample was carried out by examining the association between the predicted probability of treatment response and observed response across a range of cut-

https://doi.org/10.1017/S0033291722000228 Published online by Cambridge University Press

points derived from the training sample distribution. We evaluated model fairness, defined as whether a model performance was comparable across important segments of the population (Yuan, Kumar, Ahmad, & Teredesai, 2021), by examining variation in the association of predicted probability of response with observed response across socio-demographic subgroups (age, sex, race/ethnicity, and education) using robust Poisson regression models (Zou, 2004). Lastly, we assessed predictor importance by examining standardized model coefficients from predictors selected by the final ML model.

Data were managed and the outcome prevalence and AUC were calculated using SAS statistical software, version 9.4 (SAS Institute Inc, 2013). ML models were estimated in R, version 4.0.5 (R Core Team, 2021).

Results

Sample characteristics and treatment response

The mean QIDS-SR score of depression symptom severity at baseline was 12.9 among the total weighted sample. When we transformed QIDS-SR scores into Hamilton Depression Rating Scale criteria, 32.3% of patients met criteria for mild depression, 33.9% for moderate depression, 18.6% for severe depression, and 15.2% for very severe depression. Most patients were male, non-Hispanic White, married, and living in a major metro area (Table 1). There were no statistically significant differences in baseline socio-demographics or depression severity between patients who completed the baseline and 3-month surveys *v.* those who completed the baseline but not the 3-month survey. In the total weighted sample, 32.0% [standard error (s.e.) = 2.0] of patients responded to psychotherapy after 3 months of treatment. For our three criteria of responding to treatment, 7.9% of patients met the criteria for remission, 14.1% of patients had a QIDS-SR score at 3-months that was half or less of their baseline score, and 23.5% of patients showed improvement in role functioning.

Model performance

The AUC (s.e.) of the SL ensemble model in the test sample was 0.648 (0.039). However, the simpler lasso model had slightly better performance, with AUC (s.e.) of 0.652 (0.038). We consequently focused on the lasso model. This model had good calibration in the test sample [mean (s.e.) ICI, 0.056 (0.005); mean (s.e.) ECE, 0.054 (0.004)] (Fig. 1) as well as comparable prediction accuracy in terms of fairness across subgroups defined by age, sex, race/ethnicity, and education (Online Supplementary Table S6).

Fifty percent of patients in the top tertile of predicted probability of treatment response did, in fact, respond to treatment (Table 2). In comparison, 23.5 and 21.1%, respectively, of patients in the second and third tertiles responded to treatment.

Predictor importance

A total of 43 predictors were selected by the lasso model. Figure 2 displays these predictors, which are sorted from the strongest (defined by associations of predictors standardized to have a mean of 0 and variance of 1.0 with logits of the dichotomous outcome) at the top to weakest at the bottom. Positive associations indicated higher likelihoods of treatment response, whereas negative associations indicated lower likelihoods of treatment response. The great majority of predictors ($n = 39$) were based on patient self-reports rather than administrative or geo-spatial data. Predictors came mainly from the risk factor domains of antecedent and concomitant psychiatric conditions ($n = 7$), clinical staging ($n = 6$), treatment characteristics ($n = 6$), protective factors/resilience ($n = 5$), and socio-demographics ($n = 5$). The top 5 predictors were: having a greater cognitive reappraisal score, being aged 74+, having a longer drive time to VHA facility, having greater concerns about ADMs, and being in the current depressive episode for more than 3+ months before seeking treatment. The first three of these predictors were associated with increased likelihoods of treatment response, whereas the latter two were associated with decreased likelihoods of treatment response.

Discussion

Our finding that fewer than one-third of patients with MDD responded to psychotherapy after 3 months is lower than response rates observed in other observational studies (Blais et al., 2013) and randomized clinical trials (Cuijpers et al., 2021) from non-Veteran samples, where 40–50% of patients responded to psychotherapy. However, response rates for MDD treatment have been found to be similarly low for Veterans in other studies (Katz, Liebmann, Resnick, & Hoff, 2021), possibly due to the particularly high burden of psychiatric comorbidities and impairment in this population (Ziobrowski et al., 2021b). This low treatment response rate highlights the need to develop clinical tools that can support patients in treatment decision-making. To this end, our model is of potential value in finding that the one-third of patients with the highest predicted probability of treatment response had an observed probability of response (50%) more than twice than those of patients in the lower tertiles (23–21%).

Although this level of discrimination is both statistically and substantively significant, it is not strong enough to be the primary arbiter of treatment selection. Nor does the model provide information on which other treatments would be optimal for a given patient (i.e. ADM-alone, combined ADM and psychotherapy, some other therapy). Our model could be useful, though, in the context of a broader shared decision-making conversation that informs patients and providers about a patient's likelihood of responding to psychotherapy. Such a tool could help guide patients with a low likelihood of response toward considering alternative treatments options, thus averting the costs and morbidity of ineffective psychotherapy monotherapy. Conversely, patients with a high likelihood of response could be reassured about deferring ADMs, thus limiting their potential for somatic side effects. This approach would be similar in concept to pharmacogenomic testing, in which patients' genetic information is used to pre-emptively identify specific ADMs that are more *v.* less likely to cause side effects or be effective (Greden et al., 2019). Notably, our model performs as well as or better than pharmacogenomic testing in terms of its predictive power, as indicated by the fact that in the largest trial to date of pharmacogenomic testing for ADM selection, patients receiving test-congruent *v.* test-incongruent medications had 29% *v.* 17% treatment response rates (Greden et al., 2019).

Caution is needed in interpreting the results reported above about predictor importance because these results do not reflect causal relationships and can be unstable when, as in our dataset, many of the predictors in the full set used to select the final

https://doi.org/10.1017/S0033291722000228 Published online by Cambridge University Press

Table 1. Distribution of socio-demographic characteristics, baseline depression severity, and treatment response among the full baseline sample, analytic sample, and patients lost to follow-up

	Weighted for baseline non-response						Also weighted for loss to follow-up			
	Baseline sample[a] (*n* = 989)		Analytic sample[b] (*n* = 807)		Patients lost to follow-up[c] (*n* = 182)		Analytic sample[b] (*n* = 807)		Differences between the analytic sample and patients lost to follow-up[d]	
	%	(s.e.)	%	(s.e.)	%	(s.e.)	%	(s.e.)	χ^2	df
Age									5.36	3
18–34	22.2	(1.4)	21.0	(1.5)	27.6	(3.4)	22.1	(1.9)		
35–49	26.6	(1.4)	26.2	(1.6)	28.7	(3.4)	25.5	(1.9)		
50–59	19.0	(1.3)	19.4	(1.4)	17.0	(2.8)	18.7	(1.7)		
60+	32.2	(1.5)	33.4	(1.7)	26.8	(3.4)	33.8	(2.1)		
Sex									0.53	1
Female	20.3	(1.3)	19.9	(1.4)	22.4	(3.2)	19.0	(1.7)		
Male	79.7	(1.3)	80.1	(1.4)	77.6	(3.2)	81.0	(1.7)		
Race/ethnicity									2.94	3
Non-Hispanic White	63.1	(1.6)	63.6	(1.8)	61.0	(3.8)	68.4	(2.0)		
Non-Hispanic Black	18.9	(1.3)	17.9	(1.4)	23.5	(3.4)	15.3	(1.5)		
Hispanic	9.9	(1.0)	10.2	(1.1)	8.3	(2.2)	8.6	(1.0)		
Other	8.1	(0.9)	8.3	(1.0)	7.2	(2.0)	7.7	(1.2)		
Marital status									4.88	4
Currently married	52.6	(1.6)	53.7	(1.8)	47.2	(3.8)	53.2	(2.2)		
Divorced	24.2	(1.4)	22.9	(1.5)	30.0	(3.5)	22.3	(1.8)		
Separated	4.2	(0.7)	4.0	(0.7)	5.3	(1.8)	3.9	(0.8)		
Widowed	3.1	(0.6)	3.2	(0.6)	2.6	(1.2)	2.8	(0.6)		
Never married	16.0	(1.2)	16.2	(1.3)	14.9	(2.7)	17.9	(1.8)		
Census region									4.17	3
Northeast	12.1	(1.0)	11.9	(1.1)	13.4	(2.5)	15.0	(1.7)		
Midwest	19.5	(1.3)	18.8	(1.4)	22.6	(3.1)	18.0	(1.6)		
South	47.4	(1.6)	47.2	(1.8)	48.3	(3.8)	46.1	(2.2)		
West	21.0	(1.3)	22.1	(1.5)	15.7	(2.8)	20.9	(1.6)		
Urbanicity									1.88	2
Major metro	83.2	(1.2)	83.9	(1.3)	80.1	(3.0)	78.3	(2.0)		
Urban	15.2	(1.1)	14.7	(1.2)	17.5	(2.8)	19.4	(1.9)		
Rural	1.6	(0.4)	1.4	(0.4)	2.4	(1.2)	2.3	(0.8)		
% of population below 1.5 × of poverty line									3.59	3
1st quartile (low % with low income)	21.3	(1.3)	20.4	(1.5)	25.3	(3.3)	20.5	(1.8)		
2nd quartile	24.6	(1.4)	24.5	(1.5)	25.0	(3.3)	26.5	(2.0)		
3rd quartile	26.3	(1.4)	26.1	(1.6)	27.0	(3.4)	25.0	(1.8)		
4th quartile (high % with low income)	27.8	(1.5)	28.9	(1.6)	22.7	(3.2)	27.9	(1.9)		
Baseline depression severity									4.46	3
Mild	32.2	(1.5)	33.4	(1.7)	26.4	(3.3)	32.3	(2.0)		
Moderate	33.7	(1.5)	33.7	(1.7)	33.6	(3.6)	33.9	(2.1)		

(*Continued*)

https://doi.org/10.1017/S0033291722000228 Published online by Cambridge University Press

Table 1. (*Continued.*)

	Weighted for baseline non-response						Also weighted for loss to follow-up			
	Baseline sample[a] (*n* = 989)		Analytic sample[b] (*n* = 807)		Patients lost to follow-up[c] (*n* = 182)		Analytic sample[b] (*n* = 807)		Differences between the analytic sample and patients lost to follow-up[d]	
	%	(s.e.)	%	(s.e.)	%	(s.e.)	%	(s.e.)	χ^2	df
Severe	19.3	(1.3)	18.4	(1.4)	23.7	(3.3)	18.6	(1.7)		
Very severe	14.8	(1.2)	14.5	(1.3)	16.3	(2.9)	15.3	(1.7)		
Treatment response	–		–		–		32.0	(2.0)	–	–

DF, degrees of freedom; s.e., standard error.

[a] Patients who received psychotherapy and responded to the baseline survey.

[b] Patients who received psychotherapy and responded to both the baseline and 3-month surveys.

[c] Patients who received psychotherapy and responded to the baseline but not the 3-month survey.

[d] None of the χ^2 tests is significant at the 0.05 level, two-sided test. $p = 0.15–0.47$.

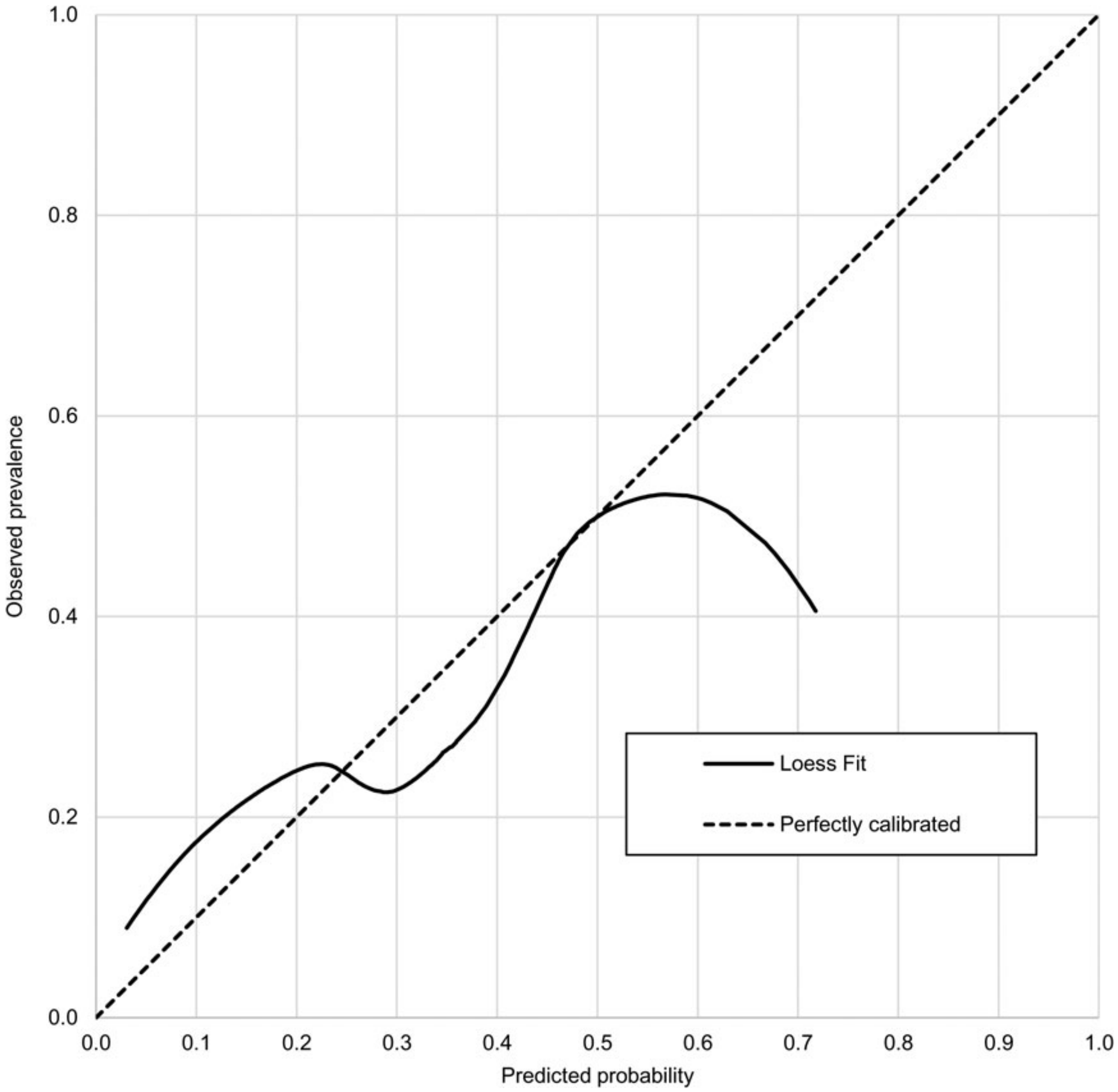

Fig. 1. Locally estimated scatterplot smoothing (LOESS) calibration curve for the predicted probability of psychotherapy response in the test sample for the Least Absolute Shrinkage and Selection Operator Feature Selection (lasso) model.

Note. The integrated calibration index is 0.056, and the expected calibration error is 0.054.

Table 2. Prediction of psychotherapy treatment response in the test sample of 241 patients using the lasso model

	Treatment response		
	Observed		
Predicted risk distribution[a]	%	(s.e.)	(*n*)
High (39.2–83.4)	50.0	(7.1)	(82)
Intermediate (21.6–39.2)	23.5	(4.6)	(91)
Low (1.3–21.6)	21.1	(5.0)	(68)

CI, confidence interval; lasso, least absolute shrinkage and selection operator.
[a]Defined by tertiles of predicted probability of treatment response in the training sample of 566 patients.

lasso predictors are highly inter-related (Leeuwenberg et al., 2022). Nonetheless, several results are noteworthy. First, nearly all top predictors were self-report measures, suggesting that patient self-report data may be more useful for predicting psychotherapy treatment response than administrative or geo-spatial data. Moreover, these self-report variables would be feasible to collect in a primary care visit. Second, some of the top predictors (e.g. recency of TBI and age at baseline) may be specific to Veterans who served in Iraq and Afghanistan. Third, several variables about socio-demographics and treatment characteristics were among the most important predictors. This is noteworthy because socio-demographics and treatment characteristics were not among the categories included by Maj et al. (2020) as salient risk factor domains that should be considered in efforts to personalize depression treatment. Fourth, while most of the selected variables are not modifiable (e.g. age, lifetime histories of mental disorders, personality characteristics), several are potentially modifiable in psychotherapy treatment, such as variables related to emotion regulation. Fifth, the model did not select any variables related to depression severity, clinical subtypes, or family history of depression, and selected only 1 variable related to depression symptoms (late insomnia). Although treatment decisions may be based on these factors in practice, these results show that these variables are not the most predictive of psychotherapy response among depressed patients in the VHA system.

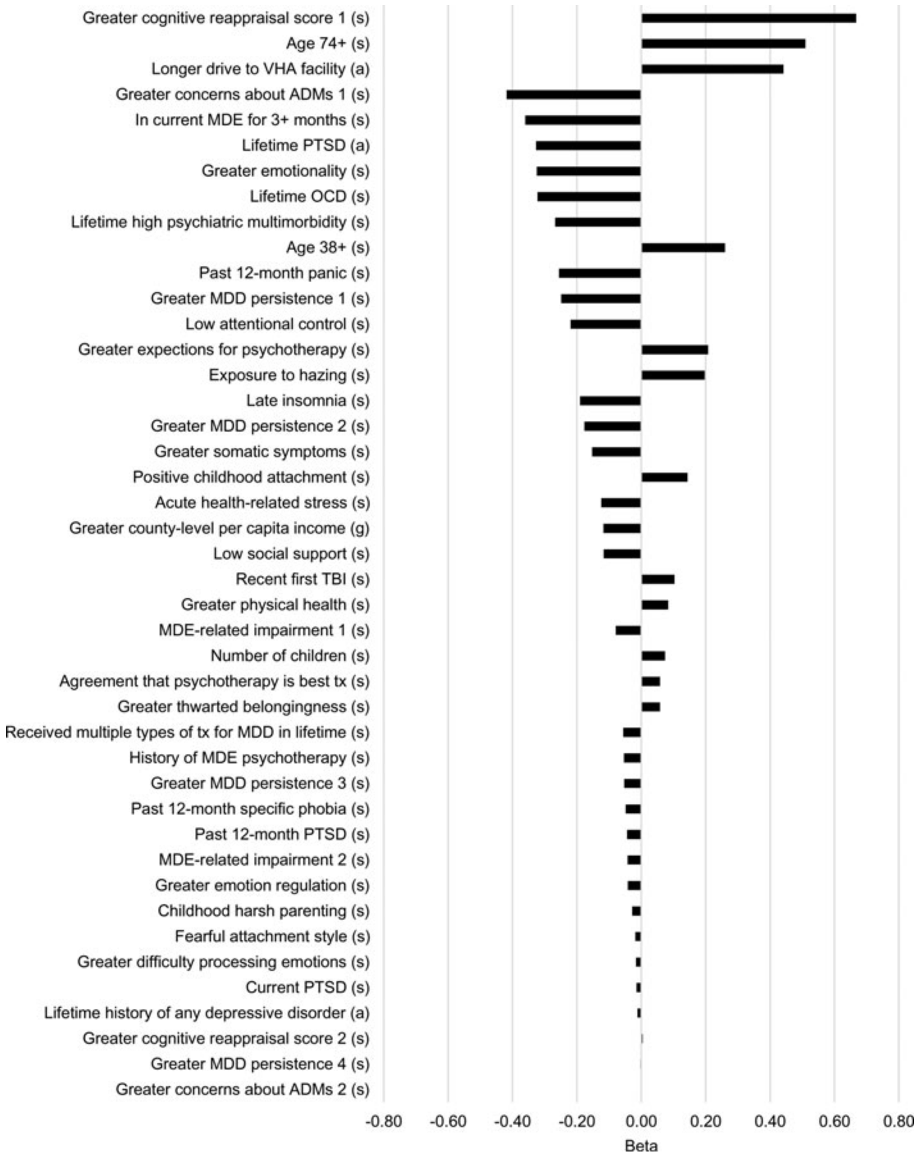

Fig. 2. Predictor importance determined by logits of standardized predictors with the dichotomous outcome based on the Least Absolute Shrinkage and Selection Operator Feature Selection (lasso) model.
Note: Each predictor is standardized to have a mean of 0 and variance of 1.0. (a) = administrative variable, (g) = geo-spatial variable, (s) = survey variable. A more detailed description of the cut-points for the variables selected by the lasso model can be found in the supplementary information. ADM, anti-depressant medication; MDE, major depressive episode; MDD, major depressive disorder; OCD, obsessive compulsive disorder; PTSD, posttraumatic stress disorder; TBI, traumatic brain injury; tx, treatment; VHA, Veterans Health Administration.

https://doi.org/10.1017/S0033291722000228 Published online by Cambridge University Press

The study has several strengths, including the large sample size, the rich and diverse set of predictors from self-reports, administrative records, and geo-spatial data, and rigorous ML methods used to develop the model and help reduce potential overfitting. However, there are also several limitations to note. First, the baseline survey response rate was low, although similar to rates reported in other studies examining mental health outcomes among VHA patients (King, Beehler, Buchholz, Johnson, & Wray, 2019; Stolzmann et al., 2019). We previously reported (Puac-Polanco et al., 2021) that there were minimal differences between responders and non-responders with regard to baseline administrative variables and we found here equally modest baseline self-report differences between baseline respondents who were followed and those lost to follow-up, both of which were adjusted for in the weighted analyses. However, we had no way to determine if response bias exists with respect to unmeasured variables. Second, our outcome measures were based on brief validated self-report scales rather than clinical interviews. Third, patients included those who had mild baseline QIDS-SR scores, whereas many other studies require baseline sores of at least moderate severity. It is noteworthy, though, that baseline symptom severity was not among the important predictors, which means that this broad definition of sample eligibility might not have influenced results. Fourth, psychotherapy response was assessed only up through 3 months of treatment. It is possible that some patients improved later and that some defined as responding at 3 months had recurrences of more severe baseline symptoms shortly thereafter. Fifth, it is unclear whether our findings are generalizable to non-VHA patients. Sixth, we did not account for possible disruptions in care due to the COVID-19 pandemic, but 92.4% of study patients completed assessments before March 2020. Seventh, with more patients receiving telehealth care since the start of the COVID-19 pandemic, the important predictors observed in this analysis may have since changed. Lastly, our predictive model only provides information on patients' likelihood of responding to psychotherapy in the absence of ADMs. The model cannot tell us which alternative treatments would be optimal for a given patient nor the magnitude of benefit a patient would be expected to attain by receiving an alternative treatment.

Conclusions

We found that a parsimonious model to predict psychotherapy treatment response for depression can be developed using a battery of self-report questions along with some administrative variables in electronic health records and geospatial variables. This model could be used to inform depressed patients pre-emptively about their likelihood of responding to psychotherapy as part of a patient-centered treatment decision-making process. Our findings should be replicated before such a model is implemented in practice. More elaborate models are also needed to compare predicted probabilities of treatment response at the patient level across different types of treatment to determine the best treatment option for particular patients (Kessler & Luedtke, 2021).

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0033291722000228.

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). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, the U.S. Department of Veterans Affairs, or the United States Government.

Conflicts 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. The remaining authors report no conflict of interest.

Ethical standards. The authors assert that 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.

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

Blais, M. A., Malone, J. C., Stein, M. B., Slavin-Mulford, J., O'Keefe, S. M., Renna, M., & Sinclair, S. J. (2013) Treatment as usual (TAU) for depression: A comparison of psychotherapy, pharmacotherapy, and combined treatment at a large academic medical center. *Psychotherapy*, *50*(1), 110–118. doi:10.1037/a0031385

Bone, C., Simmonds-Buckley, M., Thwaites, R., Sandford, D., Merzhvynska, M., Rubel, J., … Delgadillo, J. (2021) Dynamic prediction of psychological treatment outcomes: Development and validation of a prediction model using routinely collected symptom data. *The Lancet Digital Health*, *3*(4), e231–e240. doi:10.1016/s2589-7500(21)00018-2

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

Coley, R. Y., Boggs, J. M., Beck, A., & Simon, G. E. (2021) Predicting outcomes of psychotherapy for depression with electronic health record data. *Journal of Affective Disorders Reports*, *6*(100198). doi:10.1016/j.jadr.2021.100198

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. *British Journal of Surgery*, *102*(3), 148–158. doi:10.1002/bjs.9736

Cuijpers, P., Berking, M., Andersson, G., Quigley, L., Kleiboer, A., & Dobson, K. S. (2013) A meta-analysis of cognitive-behavioural therapy for adult depression, alone and in comparison with other treatments. *Canadian Journal of Psychiatry*, *58*(7), 376–385. doi:10.1177/070674371305800702

Cuijpers, P., Karyotaki, E., Ciharova, M., Miguel, C., Noma, H., & Furukawa, T. A. (2021) The effects of psychotherapies for depression on response, remission, reliable change, and deterioration: A meta-analysis. *Acta Psychiatrica Scandinavica*, *144*(3), 288–299. doi:10.1111/acps.13335

Delgadillo, J., & Gonzalez Salas Duhne, P. (2020) Targeted prescription of cognitive-behavioral therapy versus person-centered counseling for depression using a machine learning approach. *Journal of Consulting and Clinical Psychology*, *88*(1), 14–24. doi:10.1037/ccp0000476

DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L. A., & Lorenzo-Luaces, L. (2014) The personalized advantage Index: Translating research on prediction into individualized treatment recommendations. A demonstration. *PLoS ONE*, *9*(1), e83875. doi:10.1371/journal.pone.0083875

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. *Lancet*, *396*(10258), 1204-1222. doi:10.1016/s0140-6736(20)30925-9

Gelhorn, H. L., Sexton, C. C., & Classi, P. M. (2011) Patient preferences for treatment of major depressive disorder and the impact on health outcomes: A systematic review. *Primary Care Companion for CNS Disorders*, *13*(5). doi:10.4088/PCC.11r01161

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

Herrman, H., Patel, V., Kieling, C., Berk, M., Buchweitz, C., Cuijpers, P., … Wolpert, M. (in press) Time for united action on depression: A lancet-world psychiatric association commission. *The Lancet*.

Huibers, M. J., Cohen, Z. D., Lemmens, L. H., Arntz, A., Peeters, F. P., Cuijpers, P., & DeRubeis, R. J. (2015) Predicting optimal outcomes in cognitive therapy or interpersonal psychotherapy for depressed individuals using the personalized advantage Index approach. *PLoS ONE*, *10*(11), e0140771. doi:10.1371/journal.pone.0140771

Kabir, M. F., & Ludwig, S. A. (2019) Enhancing the performance of classification using super learning. *Data-Enabled Discovery and Applications*, *3*(1), 5. doi:10.1007/s41688-019-0030-0

Kappelmann, N., Rein, M., Fietz, J., Mayberg, H. S., Craighead, W. E., Dunlop, B. W., … Kopf-Beck, J. (2020) Psychotherapy or medication for depression? Using individual symptom meta-analyses to derive a symptom-oriented therapy (SOrT) metric for a personalised psychiatry. *BMC Medicine*, *18*(1), 170. doi:10.1186/s12916-020-01623-9

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

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

Koeser, L., Donisi, V., Goldberg, D. P., & McCrone, P. (2015) Modelling the cost-effectiveness of pharmacotherapy compared with cognitive-behavioural therapy and combination therapy for the treatment of moderate to severe depression in the UK. *Psychological Medicine*, *45*(14), 3019–3031. doi:10.1017/s0033291715000951

LeDell, E., van der Laan, M. J., & Petersen, M. (2016) AUC-Maximizing Ensembles through metalearning. *International Journal of Biostatistics*, *12*(1), 203–218. doi:10.1515/ijb-2015-0035

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-1uwd

Leung, L. B., Rubenstein, L. V., Yoon, J., Post, E. P., Jaske, E., Wells, K. B., & Trivedi, R. B. (2019) Veterans health administration investments In primary care And mental health integration improved care access. *Health Affairs*, *38*(8), 1281–1288. doi:10.1377/hlthaff.2019.00270

Leung, L. B., Ziobrowski, H. N., Puac-Polanco, V., Bossarte, R. M., Bryant, C., Keusch, J., … Kessler, R. C. (2021) Are veterans getting their preferred depression treatment? A national observational study in the veterans health administration. *Journal of General Internal Medicine*, Advance online publication. doi:10.1007/s11606-021-07136-2

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

Naeini, M. P., Cooper, G. F., & Hauskrecht, M. (2015). Obtaining Well Calibrated Probabilities Using Bayesian Binning. Retrieved from https://people.cs.pitt.edu/~milos/research/AAAI_Calibration.pdf.

Naimi, A. I., & Balzer, L. B. (2018) Stacked generalization: An introduction to super learning. *European Journal of Epidemiology*, *33*(5), 459–464. doi:10.1007/s10654-018-0390-z

Park, T., & Casella, G. (2008) The Bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686. doi:10.1198/016214508000000337

Pearson, R., Pisner, D., Meyer, B., Shumake, J., & Beevers, C. G. (2019) A machine learning ensemble to predict treatment outcomes following an internet intervention for depression. *Psychological Medicine*, *49*(14), 2330–2341. doi:10.1017/s003329171800315x

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.

Polley, E. C., Rose, S., & van der Laan, M. J. (2011) Super learning. In *Targeted learning: Casual inference for observational and experimental data* (ed. M. J. van der Laan and S. Rose), pp. 43–66. Springer: New York.

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/.

Roelofs, R., Shankar, V., Recht, B., Fridovich-Keil, S., Hardt, M., Miller, J., & Schmidt, L. (2019). A Meta-Analysis of Overfitting in Machine Learning, Part of Advances in Neural Information Processing Systems 32 (NeurIPS 2019). Retrieved from https://papers.nips.cc/paper/2019/hash/ee39e503b6bedf0c98c388b7e8589aca-Abstract.html.

Ross, E. L., Vijan, S., Miller, E. M., Valenstein, M., & Zivin, K. (2019) The cost-effectiveness of cognitive behavioral therapy versus second-generation antidepressants for initial treatment of major depressive disorder in the United States: A decision-analytic model. *Annals of Internal Medicine*, *171*(11), 785–795. doi:10.7326/m18-1480

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

Rush, A. J., Trivedi, M. H., Wisniewski, S. R., Nierenberg, A. A., Stewart, J. W., Warden, D., … Fava, M. (2006) Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: A STAR*D report. *The American Journal of Psychiatry*, *163*(11), 1905–1917. doi:10.1176/ajp.2006.163.11.1905

SAS Institute Inc (2013) SAS ®Software 9.4 edn. Cary, NC.

Saunders, R., Cohen, Z. D., Ambler, G., DeRubeis, R. J., Wiles, N., Kessler, D., … Buckman, J. E. J. (2021) A patient stratification approach to identifying the likelihood of continued chronic depression and relapse following treatment for depression. *Journal of Personalized Medicine*, *11*(12). doi:10.3390/jpm11121295

Serbanescu, I., Backenstrass, M., Drost, S., Weber, B., Walter, H., Klein, J. P., … Schoepf, D. (2020) Impact of baseline characteristics on the effectiveness of disorder-specific cognitive behavioral analysis system of psychotherapy (CBASP) and supportive psychotherapy in outpatient treatment for persistent depressive disorder. *Frontiers in Psychiatry*, *11*, 607300. doi:10.3389/fpsyt.2020.607300

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

https://doi.org/10.1017/S0033291722000228 Published online by Cambridge University Press

Tymofiyeva, O., Yuan, J. P., Huang, C. Y., Connolly, C. G., Henje Blom, E., Xu, D., & Yang, T. T. (2019) Application of machine learning to structural connectome to predict symptom reduction in depressed adolescents with cognitive behavioral therapy (CBT). *NeuroImage: Clinical*, *23*(101914). doi:10.1016/j.nicl.2019.101914

van Klaveren, D., Balan, T. A., Steyerberg, E. W., & Kent, D. M. (2019) Models with interactions overestimated heterogeneity of treatment effects and were prone to treatment mistargeting. *Journal of Clinical Epidemiology*, *114*, 72–83. doi:10.1016/j.jclinepi.2019.05.029

van Schaik, D. J., Klijn, A. F., van Hout, H. P., van Marwijk, H. W., Beekman, A. T., de Haan, M., & van Dyck, R. (2004) Patients' preferences in the treatment of depressive disorder in primary care. *General Hospital Psychiatry*, *26*(3), 184–189. doi:10.1016/j.genhosppsych.2003.12.001

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.

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/.

https://doi.org/10.1017/S0033291722000228 Published online by Cambridge University Press