Psychological Medicine

cambridge.org/psm

Optimizing precision medicine for second-step depression treatment: a machine learning approach

Joshua Curtiss[1], Jordan W. Smoller[2] and Paola Pedrelli[1]

[1]Depression Clinical and Research Program, Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA and [2]Center for Precision Psychiatry, Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA

Original Article

Cite this article: Curtiss J, Smoller JW, Pedrelli P (2024). Optimizing precision medicine for second-step depression treatment: a machine learning approach. *Psychological Medicine* **54**, 2361–2368. https://doi.org/10.1017/S0033291724000497

Received: 18 July 2023
Revised: 22 January 2024
Accepted: 20 February 2024
First published online: 27 March 2024

Keywords:
depression; machine learning; precision medicine

Corresponding author:
Joshua Curtiss;
Email: jcurtiss@mgh.harvard.edu

Abstract

Background. Less than a third of patients with depression achieve successful remission with standard first-step antidepressant monotherapy. The process for determining appropriate second-step care is often based on clinical intuition and involves a protracted course of trial and error, resulting in substantial patient burden and unnecessary delay in the provision of optimal treatment. To address this problem, we adopt an ensemble machine learning approach to improve prediction accuracy of remission in response to second-step treatments.
Method. Data were derived from the Level 2 stage of the STAR*D dataset, which included 1439 patients who were randomized into one of seven different second-step treatment strategies after failing to achieve remission during first-step antidepressant treatment. Ensemble machine learning models, comprising several individual algorithms, were evaluated using nested cross-validation on 155 predictor variables including clinical and demographic measures.
Results. The ensemble machine learning algorithms exhibited differential classification performance in predicting remission status across the seven second-step treatments. For the full set of predictors, AUC values ranged from 0.51 to 0.82 depending on the second-step treatment type. Predicting remission was most successful for cognitive therapy (AUC = 0.82) and least successful for other medication and combined treatment options (AUCs = 0.51–0.66).
Conclusion. Ensemble machine learning has potential to predict second-step treatment. In this study, predictive performance varied by type of treatment, with greater accuracy in predicting remission in response to behavioral treatments than to pharmacotherapy interventions. Future directions include considering more informative predictor modalities to enhance prediction of second-step treatment response.

Introduction

Antidepressant monotherapy, such as selective serotonin reuptake inhibitors (SSRIs), represents the most common first-step treatment for depression, accounting for 87% of delivered treatment modalities (Olfson, Blanco, & Marcus, 2016). Unfortunately, only 28–33% of individuals receiving SSRI monotherapy achieve successful remission (Trivedi et al., 2006), indicating that a majority of patients are still symptomatic and require further treatment.

There are limited guidelines to determine which treatment strategy would be optimal for a given individual who does not successfully respond to an initial course of SSRI. Unfortunately, available guidelines are derived from nomothetic evidence collected across group averages, which obfuscates individual differences that are better addressed by more idiographic approaches. Common principles of these nomothetic guidelines, as embodied in Bennabi et al. (2019) and Gelenberg et al. (2010), have included switching treatment after no response to initial treatment and augmenting treatment after partial response to initial treatment. Such recommendations are often general in nature and are applied in combination with clinical intuition. In practice, the sequencing of treatment selection frequently results in a protracted process of trial and error in identifying the most appropriate treatment for an individual patient. As a consequence, burden is exacerbated both in terms of *duration of illness* and *number of provider visits* (Cusin & Peyda, 2019; Kazdin & Blase, 2011).

Recently, mental health research has increasingly embraced precision medicine as a framework for personalizing treatment using more computationally rigorous procedures (Bernardini et al., 2017). A precision medicine approach leveraging machine learning may facilitate the effort to reduce the amount of time required to identify the optimal second-step treatment strategy. Commonly collected information from a routine psychiatric assessment visit (e.g. demographics, questionnaires, diagnostic information, prior treatment history, etc.) could be provided to machine learning algorithms to predict whether a specific second-step treatment

© The Author(s), 2024. Published by Cambridge University Press

strategy (i.e. augmentation or switching) would improve the probability of remission. Such a machine learning framework fosters a more idiographic approach to second-step treatment selection relative to standard nomothetic guidelines.

Prior efforts to apply machine learning as a precision medicine approach to depression treatment have illustrated how disease burden can be mitigated. One study validated a machine learning algorithm for initial treatment remission for depression that outperformed the baseline remission rate by 11% (Mehltretter et al., 2020). Assuming five visits per treatment for each patient, it was estimated that this improvement in prediction could result in approximately 1772 fewer doctor's visits and decreased duration of disease burden.

The current study introduces a computationally principled framework for improving predictions about patient treatment remission using ensemble machine learning with the super learner algorithm. A principal advantage of ensemble learners is that they leverage several individual algorithms to obtain better performance than any of the constituent learners alone. Specifically, ensemble machine learning using a super learner approach involves consolidating several individual machine learning algorithms into an ensemble algorithm that represents a weighted average of all individual learners (e.g. random forest, elastic net regression, neural network, etc.). Thus, rather than performing predictive modeling with several individual machine learning algorithms and selecting the best performing algorithm, as is common practice, the ensemble approach capitalizes on the individual strengths of several disparate learning algorithms by including them all in the final model. The ensemble machine learning algorithms will be applied to the largest, multi-center clinical trial ever conducted on treatments for depression: the STAR*D study. To date, machine learning approaches to depression treatment have been primarily used for first-step interventions of standard care (e.g. SSRI monotherapy) (Chekroud et al., 2016; Mehltretter et al., 2020). This project represents the first study to employ an ensemble machine learning approach to improve predictions of treatment response to one of seven second-step interventions for depression. Precision medicine may facilitate more expeditious provision of care that is personalized to specific depressed individuals. Specifically, this study has the potential to inform efforts to develop clinical decision-support tools that could improve patient outcomes.

Methods

The current study is a secondary data analysis of the STAR*D study, which was a multicenter longitudinal NIMH-sponsored study. STAR*D was designed to determine the short- and long-term effects of different sequences of medication and/or psychotherapy for the treatment of unipolar depression that has not responded adequately to an initial standard antidepressant trial. In particular, STAR*D compared the effectiveness of different sets of treatment options: (1) augmenting the first antidepressant with another medication or psychotherapy, (2) switching to another antidepressant or psychotherapy.

Trial design

At Level 1, patients initially received citalopram (CIT) for a minimum of 8 weeks. Patients who did not remit during Level 1 were offered Level 2 treatment, which represented two overall treatment strategies: (1) Medication or Psychotherapy Switch – switching from CIT to another antidepressant medication or Cognitive Psychotherapy (CT), and (2) Medication or Psychotherapy Augmentation – augmenting CIT with a second medication or CT. For those who switch treatments at Level 2, sertraline (SER) (a second SSRI), venlafaxine (VEN) (an antidepressant with both noradrenergic and serotonergic effects), bupropion (BUP) (an antidepressant with both noradrenergic and dopaminergic effects), or CT were available. Likewise, within the Medication or Psychotherapy Augmentation strategy, the three treatments for augmenting CIT were BUP, buspirone (BUS) (an antianxiety medication), or CT.

Study inclusion and exclusion criteria

Inclusion criteria permitted outpatients who were 18–75 years of age and had a nonpsychotic major depressive episode determined by a baseline HAM-D score $\geqslant$14. They were eligible if their clinicians determined that outpatient treatment with an antidepressant medication was both safe and indicated. Patients who were pregnant or breast-feeding and those with a primary diagnosis of bipolar, psychotic, obsessive-compulsive, or eating disorders were excluded from the study, substance dependence (only if it required inpatient detoxification) or a clear history of nonresponse or intolerance to any protocol antidepressant in the first two treatment steps were also exclusionary. To ensure broad and inclusive eligibility criteria, co-morbidity with other Axis I anxiety and emotional disorders was permitted. Diagnostic criteria were assessed with a clinician administered checklist measuring Axis I symptoms of the *Diagnostic and Statistics Manual*, fourth edition revised (Trivedi et al., 2006).

Participants

In Level 1 of STAR*D, 4041 depressed patients enrolled, and, after exclusion criteria were applied and study discontinuation was considered, 2876 patients were included in the Level 1 results. In Level 2, 1439 patients were randomized to one of the seven treatment options deemed acceptable to each participant. Patients were 63.7% female, and the mean age was 40.8 (S.D. = 13). Regarding self-reported ethnicity, 75.8% of patients were white, 17.6% were African American, and 6.6% identified as other.

Measures

In the current analysis, the predictor variables were all assessed prior to initiation of Level 2 treatments. Specifically, demographic predictors were collected at the very onset of the STAR*D study design, and clinical assessment variables were collected subsequent to Level 1 treatment, but prior to Level 2 treatment. The outcome variable (i.e. treatment remission) was estimated post-treatment for each Level 2 intervention.

Predictor variables

Demographic and psychosocial variables

Includes age, sex, race, marital status, education in years, occupational employment status, comorbid disorders, age of depression onset, and length of illness.

https://doi.org/10.1017/S0033291724000497 Published online by Cambridge University Press

Clinical assessment variables

Includes the Quick Inventory of Depressive Symptomatology (QIDS; Rush et al., 2003), Hamilton Rating Scale for Depression (HAMD; Hamilton, 1960), Short-Form Health Survey (SF-12; Ware, Kosinski, and Keller, 1996), Work and Social Adjustment Scale (WSAS; Mundt, Marks, Shear, and Greist, 2002), Work Productivity and Activity Impairment Questionnaire (WPAI; Reilly, Zbrozek, and Dukes, 1993), Quality of Life Enjoyment and Satisfaction Questionnaire (Q-LES-Q; Endicott, Nee, Harrison, and Blumenthal, 1993), and the first 100 items of the Psychiatric Diagnostic Symptom Questionnaire (PDSQ; Zimmerman and Mattia, 2001).

Outcome variables

Treatment remission

Consistent with the original study (Trivedi et al., 2006), treatment remission is defined as a post-treatment score of $\leqslant 7$ on the total HAM-D scale, treated as a binary, dummy-coded outcome variable (i.e. remitted = 1, non-remitted = 0).

Data analysis

To predict treatment remission, machine learning algorithms were evaluated using all the available predictor variables listed above (total = 115). Listwise deletion was applied prior to final analyses, as the machine learning strategy currently used requires no missing data. Specifically, a super learner approach was adopted to optimize prediction performance. Super learning involves consolidating several individual machine learning algorithms into a stacked ensemble algorithm, representing a weighted average of all individual learners (Polley, LeDell, Kennedy, & Laan, 2021; Van der Laan, Polley, & Hubbard, 2007; Van der Laan & Rose, 2011). The final aggregate model is at least as accurate as the best individual algorithm included in the ensemble (Boehmke & Greenwell, 2019). The ensemble model combined several well-established algorithms designed for categorical outcomes that cover a variety of analytical assumptions: elastic net binomial regression, ranger random forest, support vector machine, neural network, extreme gradient boosting, Bayesian generalized linear regression, and the overall mean (Boehmke & Greenwell, 2019). For a brief description of each procedure please refer to online Supplementary Appendix I in the Supplementary Materials. Furthermore, a *t* test variable filtering procedure was employed for only the training sample folds to identify the top ten predictors for each treatment condition. In brief, *t* test variable filtering examines the relationship between the binary outcome variable and the predictor variables in the training folds by preforming *t* tests with the binary outcome variable of the machine learning model now being used as the independent variable. The rank order of the *t* test *p* values are determined across all of the machine learning predictor variables, and the top ten most significant predictor variables were retained. The reasons for variable filtering were twofold. First, variable filtering can mitigate the influence of excessive irrelevant and non-informative predictors, which can undermine model performance and increase error for several algorithms (Kuhn & Johnson, 2019). Second, deriving models with fewer predictors can facilitate future implementation.

When predicting categorical classification outcomes, it is important to consider the presence of potential class imbalances in the outcome variable. Large class imbalances can bias machine learning results toward favoring prediction of the more frequent outcome. Class imbalances (i.e. minority class representing less than 25%) were observed for the CITCT and CT conditions. To address this, we implemented the Synthetic Minority Over-sampling Technique (SMOTE), which combines over-sampling of the minority class and under-sampling of the majority class to promote a more balanced class structure for the outcome variable (Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

Machine learning models were examined using nested cross-validation, which partitions the sample into outer loop and an inner loop subsets. For the outer loop, ten folds were specified, of which nine are used in the training process and final predictions are made in the remaining subset. This process is repeated for each of the remaining subsets, and results are averaged to produce a single estimate. The inner loop also contains ten folds. The inner loop is used to build the ensemble estimator and the weighting of the constituent individual learners, whereas the outer loop test set is used to provide a final estimation of the performance of the ensemble model. Nested cross-validation ensures that the final testing folds are not contaminated as a consequence of data leakage. Variable filtering and SMOTE procedures were applied to only the training folds. Separate ensemble machine learning models were estimated for each of the seven treatment conditions.

Performance metrics included area under the receiver operator curve (AUC, a measure of model discrimination), sensitivity, specificity, positive predictive value (PPV, i.e. the probability that a patient predicted to achieve remission actually does so), negative predictive value (NPV i.e. the probability that a patient predicted to not achieve remission actually does not), and accuracy (the proportion of total outcomes – remitted *v.* not remitted--classified correctly). PPV and NPV were determined using the sensitivity and specificity values empirically estimated from the final model performance for each model. The class probability threshold was 0.50 or greater for determining predicted remission status. Analyses were performed using the following R packages: nestedcv (Lewis, Spiliopoulou, & Goldmann, 2022), SuperLearner (Polley et al., 2021), and Caret (Kuhn, 2022).

Results

Model performance with full predictor Set

Results of the final ensemble models with nested cross-validation revealed differential classification performance across treatment type (Table 1). These results are visually depicted in Fig. 1. Discrimination was significantly better-than-chance for CT (AUC = 0.82, 95% CI 0.64–0.99) and BUP (AUC = 0.66, 95% CI 0.58–0.74). However, sensitivity was poor for the BUP model (5%) and moderate (33%) for the CT model. The remaining super learner models yielded AUCs that did not outperform chance: CITBUP (AUC = 0.52, 95% CI 0.45–0.59), CITBUS (AUC = 0.55, 95% CI 0.44–0.63), CITCT (AUC = 0.51, 95% CI 0.36–0.66), SER (AUC = 0.57, 95% CI 0.48–0.66), and VEN (AUC = 0.57, 95% CI 0.49–0.65).

Model performance with filtered predictor set

As a consequence of the *t* test filter, each super learner model was trained on the top ten predictors most associated with remission. In Table 2, the top ten predictors are displayed for each intervention. Of note, the variables that were most commonly predictive across the treatments include (1) work and social related

https://doi.org/10.1017/S0033291724000497 Published online by Cambridge University Press

Table 1. Predictive performance with full predictor set

Treatment	AUC	AUC 95% CI	Sensitivity	Specificity	Positive predictive value	Negative predictive value	Accuracy	Accuracy 95% CI
BUP	0.66***	0.58–0.74	0.05	0.98	0.50	0.75	0.74	0.68–0.80
CITBUP	0.52	0.45–0.59	0.08	0.93	0.42	0.62	0.60	0.54–0.66
CITBUS	0.55	0.44–0.63	0.04	0.98	0.43	0.67	0.67	0.61–0.73
CITCT	0.51	0.36–0.66	0.10	0.85	0.18	0.74	0.66	0.55–0.76
CT	0.82***	0.64–0.99	0.33	0.98	0.75	0.88	0.88	0.76–0.95
SER	0.57	0.48–0.66	0.00	0.98	0.00	0.73	0.72	0.65–0.77
VEN	0.57	0.49–0.65	0.00	0.99	0.00	0.74	0.73	0.67–0.79

Note: *** = $p < 0.001$; AUC, area under the curve; CI, confidence interval; BUP, Bupropion; CITBUP, Citalopram + Bupropion; CITBUS, Citalopram + Buspirone; CITCT, Citalopram + Cognitive Therapy; CT, Cognitive Therapy; SER, Sertraline; VEN, Venlafaxine.

impairment (WSAS), which was predictive for four treatments (i.e. BUP, CITBUS, SER, and VEN); (2) quality of life (QLESQ), which was predictive for four treatments (i.e. BUP, CITBUS, CITCT, and VEN); depression symptomatology as measured by the QIDS (QVTOT), which was predictive for four treatments (i.e. BUP, CITBUS, CT, and SER); and overall depression severity as measured by the HDRS (HDTOT), which was predictive for three treatments (i.e. CITBUP, SER, and VEN).

For models with the top ten filtered variables, results revealed classification performance as significantly better than chance for five of the second-step treatments (Table 3). These results are visually depicted in Fig. 2. The best predictive performance was still associated with the CT (AUC = 0.72, 95% CI 0.504–0.94) and BUP (AUC = 0.70, 95% CI 0.62–0.79) treatments. Furthermore, statistically significant prediction was obtained for CITCT (AUC = 0.65, 95% CI 0.501–0.81), VEN (AUC = 0.60, 95% CI 0.51–0.68), and SER (AUC = 0.59, 95% CI 0.51–0.67). The remaining two super learner models were associated with AUC performance that did not outperform chance: CITBUP (AUC = 0.55, 95% CI 0.47–0.62) and CITBUS (AUC = 0.53, 95% CI 0.46–0.61).

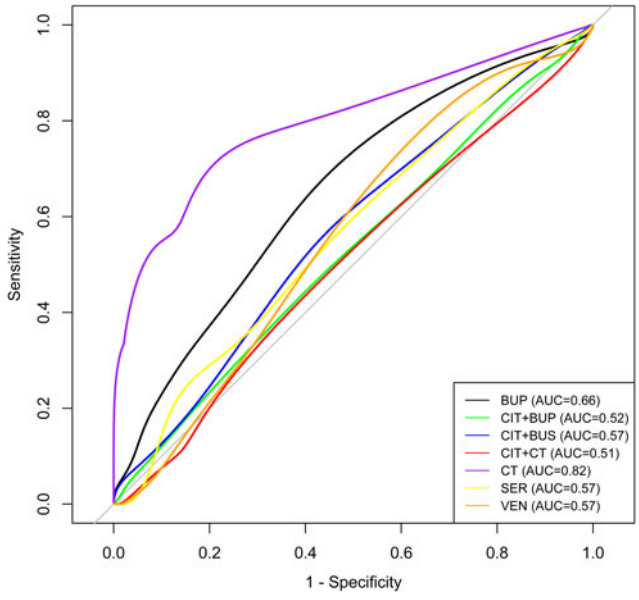


Figure 1. ROC curve for models with full predictor set.

Discussion

The current study constitutes the first comprehensive effort to predict second-step treatment outcomes for MDD using ensemble machine learning algorithms. There exists relatively little prognostic information about what factors are predictive of successful pharmacotherapy and cognitive therapy for patients who do not achieve remission after an initial trial of SSRI treatment. Consistent with a precision psychiatry framework (Bernardini et al., 2017), efforts such as ours can be useful for developing clinical decision-support tools to guide treatment of depression. In the current study, prediction of treatment remission was evaluated with ensemble super learner algorithms, including models trained on the full set of predictor variables and on a more parsimonious subset of the top ten predictors.

Results of the final super learner models demonstrated differential predictive performance dependent on the particular intervention. AUC values ranged from 0.51 to 0.82 for the models trained on the full predictor set. The best predictive performance was associated with the cognitive therapy (AUC = 0.82) and bupropion (AUC = 0.66) second-step treatments, whereas prediction of remission status failed to significantly outperform the chance for the remaining five treatment strategies. For cognitive therapy, both the positive predictive value (0.75) and negative predictive value (0.88) were good, indicating that a high proportion of the machine learner's predictions of remitters and non-remitters were accurate. For bupropion, the positive predictive value (0.50) indicated that roughly half of the learner's predictions of being a remitter were accurate, whereas the negative predictive value (0.75) indicted that a majority of the learner's predictions of being a non-remitter were accurate.

Of note, training super learner models on a more parsimonious subset of features, identified by a univariate filtering feature selection process, resulted in statistically significant prediction for more treatments. Again, predictive performance was best for cognitive therapy (AUC = 0.72) and bupropion (AUC = 0.70), and statistically significant prediction was also obtained for combined citalopram and cognitive therapy (AUC = 0.65), venlafaxine (AUC = 0.60), and sertraline (AUC = 0.59). For every treatment except the combined citalopram and bupropion intervention, the positive predictive values were regularly lower than the negative predictive values, indicating predictions of being a non-remitter may be more accurate than predictions of being a remitter. Submitting the models to such feature selection processes may have augmented successful prediction for more treatments, as

https://doi.org/10.1017/S0033291724000497 Published online by Cambridge University Press

https://doi.org/10.1017/S0033291724000497 Published online by Cambridge University Press

Table 2. Top ten filtered variables

	BUP		CITBUP		CITBUS		CITCT		CT		SER		VEN	
1	Work and social impairment (WSAS)	0.93*	Overall mental health (SF-12)	1.02*	Work and social impairment (WSAS)	0.98	Believe better off dead	12.07***	Depression severity (QIDS)	0.79	Total depression (HAMD)	0.97	Total depression (HAMD)	0.93*
2	Anxiety travelling	0.29*	Obsessions about acting violently	0.67	Quality of life (QLESQ)	1.01	Quality of life (QLESQ)	1.10***	Inattention	0.08*	Paranoia about being in danger	0.16	Belief of being controlled by some force	0.14
3	Anxiety in open spaces	0.20	Total depression (HAMD)	0.97	Overall physical health (SF-12)	1.01	Sadness	0.29	Restlessness	0.07	Washing compulsions	0.36	Anxiety in open spaces	0.20
4	Depression severity (QIDS)	0.97	Suicidal thoughts	0.66	Anxiety leaving home	0.57	Tiredness	0.34	Indecisiveness	1.07	Inattention and indecision	0.56	Work and social impairment (WSAS)	0.97
5	Anxiety being home alone	0.41	Obsessions about forgetting	0.72	Depression severity (QIDS)	0.97	Insomnia	1.30	Reduced joy	0.23	Specific suicide plan	0.43	Avoid situation to prevent panic attack	0.76
6	See/hear things other people didn't	0.34	Anxiety of shortness of breath	0.64	Avoidance of trauma	0.74	Disgust after overeating	0.30	Panic attacks	0.10	Depression severity (QIDS)	1.00	Quality of life (QLESQ)	1.00
7	Quality of life (QLESQ)	0.99	Sex	1.87*	Aches/pains	0.68	Drinking causing marriage problems	14.82	Social anxiety at parties	0.26	Guilt	0.68	Cutting down on drinking	0.42
8	Activity Impairment (WPAI)	1.00	Suicidal plans	0.78	Trauma reminders that cause shaking	0.85	Upset after eating binge	0.21	Avoidance of social situations	0.16	Low self-esteem	0.65	Panic attacks leading to avoidance	0.75
9	Strict diet	0.23	Concern about drug problem	0.46	Nervousness	0.92	Suicidal thoughts	1.12	Reduced interest	0.80	Work and social impairment (WSAS)	0.98	Years of schooling	1.13*
10	Obsessions about acting violently	0.62	Obsessions about checking	0.75	Trauma flashbacks	0.87	Pessimistic thoughts	6.22	Years of schooling	1.46	Seriously considering suicide	0.60	Anxiety in crowded places	0.90

BUP, bupropion; CITBUP, citalopram + Bupropion; CITBUS, citalopram + cuspirone; CITCT, citalopram + cognitive therapy; CT, cognitive therapy; SER, sertraline; VEN, venlafaxine.
Variables are filtered based on the ranking of the smallest *t* test *p* values. Unless otherwise specified (i.e. HAMD, QIDS, WSAS, QLESQ, QPAI, SF-12, sex, and years of schooling), all variables denote individual items from the PDSQ.
Note: Exponentiated coefficients (i.e. odds ratios) are presented from the Bayesian generalized linear model algorithm to facilitate interpretation. * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$.

Table 3. Predictive performance with top ten predictors

Treatment	AUC	AUC 95% CI	Sensitivity	Specificity	Positive predictive value	Negative predictive value	Accuracy	Accuracy 95% CI
BUP	0.70***	0.62–0.79	0.21	0.94	0.57	0.78	0.76	0.70–0.81
CITBUP	0.55	0.47–0.62	0.22	0.82	0.43	0.63	0.59	0.52–0.65
CITBUS	0.53	0.46–0.61	0.12	0.92	0.42	0.68	0.66	0.59–0.71
CITCT	0.65*	0.501–0.81	0.30	0.85	0.78	0.25	0.71	0.60–0.81
CT	0.72*	0.504–0.94	0.33	0.89	0.38	0.88	0.80	0.68–0.89
SER	0.59*	0.51–0.67	0.03	0.94	0.18	0.72	0.70	0.63–0.76
VEN	0.60*	0.51–0.68	0.03	0.95	0.18	0.74	0.71	0.65–0.77

Note: * = $p < 0.05$; *** = $p < 0.001$; AUC, area under the curve; CI, confidence interval; BUP, bupropion; CITBUP, citalopram + bupropion; CITBUS, citalopram + buspirone; CITCT, citalopram + cognitive therapy; CT, cognitive therapy; SER, sertraline; VEN, venlafaxine.

feature selection filtering can mitigate potential overfitting and enhance model accuracy by reducing noise or redundancy in the predictor set (Kuhn & Johnson, 2013). Simulations have illustrated how including non-informative predictors can undermine model performance for various algorithms (Kuhn & Johnson, 2019). Thus, variable filtering procedures reducing non-informative predictors may have enhanced model performance in the current study.

Among all the second-step treatment options, the super learner models yielded the best performance for cognitive therapy using both the full and parsimonious predictor set. These results are broadly consistent with extant literature on moderators and predictors of psychotherapy and pharmacotherapy for depression, suggesting that traditional demographic and self-report clinical variables have attained variable success in predicting treatment outcomes in general (Kessler, 2018; Papakostas & Fava, 2008; Perlman et al., 2019; Spielmans & Flückiger, 2018). With the exception of bupropion, it appears that machine learning algorithms afforded worse prediction for interventions involving pharmacotherapy. There is no obvious reason why prediction of remission status of cognitive therapy was more successful than that of pharmacotherapy. Overall, these results underscore the need to improve precision medicine approaches for pharmacotherapy treatments. Future research may consider adopting predictor sets that extend beyond self-report psychological constructs such as genetic variables, inflammatory markers, neuroimaging substrates, blood lab markers, and electronic health record data. Comprehensive reviews have identified such features as having potential prognostic value for antidepressant outcomes (Perlman et al., 2019), and it may be profitable for machine learning models to complement traditional psychosocial variables with such predictors. Furthermore, electronic health record data has demonstrated clinical utility in predicting important clinical outcomes (e.g. first-line antidepressant response, treatment resistance, treatment dropout, etc.; Hughes et al., 2020; Lage, McCoy, Perlis, & Doshi-Velez, 2022; Pradier, McCoy, Hughes, Perlis, & Doshi-Velez, 2020; Sheu et al., 2023).

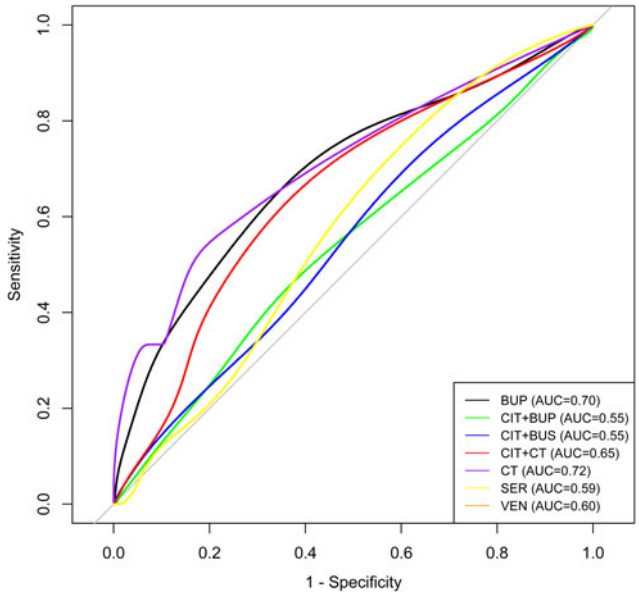


Figure 2. ROC curve for models with top ten predictors.

Of the top predictors identified as part of the feature selection filters, work and social impairment, quality of life, and depression severity were the most common predictors shared across all treatment strategies. This pattern of results is consistent with treatment moderators identified in prior reviews (Papakostas & Fava, 2008; Perlman et al., 2019). Indeed, the original evaluation of first-step citalopram treatment revealed that these same constructs differentiated remission status in the STAR*D trial (Trivedi et al., 2006). It is noteworthy that these factors are also broadly predictive of second-step treatment outcomes across several different intervention options, attesting to their clinical utility and value.

One other striking feature of the results is that all the models were associated with high specificity and relatively modest sensitivity, irrespective of whether the models were trained on the full predictor set or filtered predictor set. Moreover, the models demonstrated better negative predictive value rates relative to positive predictive value rates. In terms of interpretability, this pattern would indicate that the predictions classified as non-remitters have greater accuracy, which fosters more confidence in instances when the algorithm predicts someone to be a non-remitter. This balance of performance metrics warrants a nuanced perspective for potential clinical implementation.

Embracing a precision medicine perspective has the potential to transform our approach to second-step care for depression. Within the overall context of machine learning in psychiatry, model performance in terms of AUC for several of the treatments in the current study was similar to that of other machine learning studies predicting treatment outcomes for depression (Chekroud

https://doi.org/10.1017/S0033291724000497 Published online by Cambridge University Press

et al., 2016; Nie, Vairavan, Narayan, Ye, & Li, 2018). Notably, the current study is the first to comprehensively evaluate the utility of super learner algorithms across a wide array of second-step treatment options. Future efforts to develop accurate machine learning models predicting treatment outcomes can facilitate clinical decision-making. Machine learning algorithms can be incorporated into clinical support tools that provide clinicians with the predicted probability of remission from a given treatment, which can better inform treatment planning and discussions with patients about the relative benefits of an intervention.

A principal strength of the current study is that it examines prediction of outcomes for *multiple* treatments rather than prediction of a single intervention, as is commonplace for most research (Perlman et al., 2019). By developing predicative algorithms for multiple treatments, as opposed to a single treatment, patient data can be submitted as predictive features to several different treatment algorithms simultaneously to determine the probability of remission in response to each intervention. This approach may promote a more informed decision-making process between patients and providers about what treatment protocol may be preferable, after considering a confluence of factors such as predictive modeling results for each treatment, patient preference, potential side-effects, as well as other relevant information. Obtaining the predictive probability of success for each treatment strategy better facilitates these types of more in-depth patient-provider conversations and decision-making, which has advantages over a winner-take-all approach in which machine learning is used to recommend only one treatment strategy based on the highest predictive probability value. This latter approach can obscure important information and deter thoughtful consideration of the value of these other factors that influence treatment preference.

Given the relatively higher negative predictive value performance of the models, the machine learning algorithms can afford insight into which specific second-step treatments may not confer clinical benefit. Furthermore, the results the present study are noteworthy in light of the fact that models were trained only on self-reported demographic and clinical symptom data, which represents a more feasible approach in comparison to using more expensive and burdensome data sources for machine learning models such as neuroimaging or biomarker data (Lee et al., 2018).

Notwithstanding the strengths of the current study, certain limitations warrant mention. First, although some models evinced good predicative performance, AUC values were not statistically better than chance in predicting remission status for several treatments. Thus, it would be contraindicated to recommend the current models for clinical use in their current form, as the predictions from certain models suffer from poor accuracy and may incur a potential risk of misclassification. Given that the current study leveraged the largest clinical trial of second-step depression treatments, the current results provide an important evidence base about the strengths and limitations of machine learning prediction models in this setting. Future research may be able to improve model accuracy by considering other potential predictors that were not available in the STAR*D study (e.g. genetic factors, neuroimaging substrates, electronic health record data, etc.). Second, the sample size for some treatment conditions was relatively modest in the context of machine learning research. That notwithstanding, the STAR*D trial represents the most extensive and well-powered study examining second-step treatments for depression, and machine learning studies have resulted in success using smaller sample sizes (e.g. Flygare et al., 2020). To mitigate this, we applied nested cross-validation, which can reduce bias in models with limited sample sizes. Third, for three of the treatment conditions, the remission status outcome was imbalanced and adversely influenced model performance. To address this, we used the SMOTE procedure to produce a more balanced class profile for model training. Fourth, results might be influenced by the nature of randomization for the STAR*D trial, as patients willing to accept randomization to different treatment strategies may exhibit differences that could influence the difference across the treatment groups.

The current study constitutes the most comprehensive examination to date of super learning ensemble models in predicting remission across an extensive variety of different second-step treatments for depression. Before translating predictive modelling to clinical practice, improvements in model performance will be needed and might be achieved through using a more diverse array of multimodal predictors (e.g. self-report, biomarker data, etc.). In conclusion, the current study illustrates a proof-of-concept approach using machine learning as a tool for improving personalized second-step treatment for depression and provides insight into important next steps for future precision medicine research on depression.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0033291724000497

Acknowledgements. The authors would like to commend Dr Myles Lewis for his valuable assistance with the R package 'nestedcv', and to thank Dr Christopher Kennedy for his conversations about machine learning.

Funding statement. The study received financial support from the Kaplen Fellowship on Depression.

Competing interests. The authors declare no competing interests.

References

Bennabi, D., Charpeaud, T., Yrondi, A., Genty, J. B., Destouches, S., Lancrenon, S., … Dorey, J. M. (2019). Clinical guidelines for the management of treatment-resistant depression: French recommendations from experts, the French Association for Biological Psychiatry and Neuropsychopharmacology and the foundation FondaMental. *BMC Psychiatry*, *19*(1), 1–12.

Bernardini, F., Attademo, L., Cleary, S. D., Luther, C., Shim, R., Quartesan, R., & Compton, M. T. (2017). Risk prediction models in psychiatry: Toward a new frontier for the prevention of mental illnesses. *Journal of Clinical Psychiatry*, *78*, 572–583.

Boehmke, B., & Greenwell, B. M. (2019). *Hands-on machine learning with R*. New York: CRC Press.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., … Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. *The Lancet Psychiatry*, *3*(3), 243–250.

Cusin, C., & Peyda, S. (2019). Treatment-Resistant depression. In B. G. Shapero, D. Mischoulon, & C. Cusin (Eds.), *The Massachusetts general hospital guide to depression: New treatment insights and options* (pp. 3–19). Cham: Springer International Publishing.

Endicott, J., Nee, J., Harrison, W., & Blumenthal, R. (1993). Quality of life enjoyment and satisfaction questionnaire: A new measure. *Psychopharmacology Bulletin*, *29*(2), 321–326.

Flygare, O., Enander, J., Andersson, E., Ljótsson, B., Ivanov, V. Z., Mataix-Cols, D., & Rück, C. (2020). Predictors of remission from body dysmorphic

disorder after internet-delivered cognitive behavior therapy: A machine learning approach. *BMC Psychiatry*, *20*, 1–9.
Gelenberg, A. J., Freeman, M. P., Markowitz, J. C., Rosenbaum, J. F., Thase, M. E., Trivedi, M. H., & Van Rhoads, R. S. (2010). American Psychiatric association practice guidelines for the treatment of patients with major depressive disorder. *American Journal of Psychiatry*, *167*(Suppl 10), 9–118.
Hamilton, M. (1960). The hamilton depression scale – accelerator or break on antidepressant drug discovery. *Psychiatry*, *23*, 56–62.
Hughes, M. C., Pradier, M. F., Ross, A. S., McCoy, T. H., Perlis, R. H., & Doshi-Velez, F. (2020). Assessment of a prediction model for antidepressant treatment stability using supervised topic models. *JAMA Network Open*, *3*(5), e205308–e205308.
Kazdin, A. E., & Blase, S. L. (2011). Rebooting psychotherapy research and practice to reduce the burden of mental illness. *Perspectives on Psychological Science*, *6*(1), 21–37.
Kessler, R. C. (2018). The potential of predictive analytics to provide clinical decision support in depression treatment planning. *Current Opinion in Psychiatry*, *31*(1), 32–39.
Kuhn, M. (2022). caret: Classification and regression training (R package version 6.0-93) [Computer software]. The Comprehensive R Archive Network. Available from https://CRAN.Rproject.org/package=caret
Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York: Springer.
Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. New York: CRC Press.
Lage, I., McCoy Jr, T. H., Perlis, R. H., & Doshi-Velez, F. (2022). Efficiently identifying individuals at high risk for treatment resistance in major depressive disorder using electronic health records. *Journal of Affective Disorders*, *306*, 254–259.
Lee, Y., Ragguett, 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.
Lewis, M., Spiliopoulou, A., & Goldmann, K. (2022). nestedcv: Nested cross-validation with 'glmnet' and 'caret' (R package version 0.3.6) [Computer software]. The Comprehensive R Archive Network. Available from https://github.com/myles-lewis/nestedcv
Mehltretter, J., Fratila, R., Benrimoh, D. A., Kapelner, A., Perlman, K., Snook, E., … Turecki, G. (2020). Differential treatment benefit prediction for treatment selection in depression: A deep learning analysis of STAR*D and CO-MED data. *Computational Psychiatry*, *4*, 61–75.
Mundt, J. C., Marks, I. M., Shear, M. K., & Greist, J. M. (2002). The work and social adjustment scale: A simple measure of impairment in functioning. *The British Journal of Psychiatry*, *180*(5), 461–464.
Nie, Z., Vairavan, S., Narayan, V. A., Ye, J., & Li, Q. S. (2018). Predictive modeling of treatment resistant depression using data from STAR* D and an independent clinical study. *PloS One*, *13*, e0197268.
Olfson, M., Blanco, C., & Marcus, S. C. (2016). Treatment of adult depression in the United States. *JAMA Internal Medicine*, *176*(10), 1482–1491.
Papakostas, G. I., & Fava, M. (2008). Predictors, moderators, and mediators (correlates) of treatment outcome in major depressive disorder. *Dialogues in Clinical Neuroscience*, *10*(4), 439–451.
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.
Polley, E., LeDell, E., Kennedy, C., & Laan, M. V. D. (2021). SuperLearner: Super learner prediction (R package version 2.0-28) [Computer software]. The Comprehensive R Archive Network. Available from https://CRAN.R-project.org/package=SuperLearner
Pradier, M. F., McCoy Jr, T. H., Hughes, M., Perlis, R. H., & Doshi-Velez, F. (2020). Predicting treatment dropout after antidepressant initiation. *Translational Psychiatry*, *10*(1), 60.
Reilly, M. C., Zbrozek, A. S., & Dukes, E. M. (1993). The validity and reproducibility of a work productivity and activity impairment instrument. *Pharmacoeconomics*, *4*(5), 353–365.
Rush, A. J., Trivedi, M. H., Ibrahim, H. M., Carmody, T. J., Arnow, B., Klein, D. N., … Thase, M. E. (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.
Sheu, Y. H., Magdamo, C., Miller, M., Das, S., Blacker, D., & Smoller, J. W. (2023). AI-assisted prediction of differential response to antidepressant classes using electronic health records. *NPJ Digital Medicine*, *6*(1), 73.
Spielmans, G. I., & Flückiger, C. (2018). Moderators in psychotherapy meta-analysis. *Psychotherapy Research: Journal of the Society for Psychotherapy Research*, *28*(3), 333–346.
Trivedi, M. H., Rush, A. J., Wisniewski, S. R., Nierenberg, A. A., Warden, D., Ritz, L., … Star* D Study Team. (2006). Evaluation of outcomes with citalopram for depression using measurement-based care in STAR* D: Implications for clinical practice. *American Journal of Psychiatry*, *163*(1), 28–40.
Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, *6*(1), 25.
Van der Laan, M. J., & Rose, S. (2011). *Targeted learning: Causal inference for observational and experimental data* (Vol. *10*, pp. 978–971). New York: Springer.
Ware, J. E., Kosinski, M., & Keller, S. D. (1996). A 12-item short-form health survey: Construction of scales and preliminary tests of reliability and validity. *Medical Care*, *34*(3), 220–233.
Zimmerman, M., & Mattia, J. I. (2001). A self-report scale to help make psychiatric diagnoses: The psychiatric diagnostic screening questionnaire. *Archives of General Psychiatry*, *58*(8), 787–794.

https://doi.org/10.1017/S0033291724000497 Published online by Cambridge University Press