Psychiatry Research 335 (2024) 115858

Contents lists available at ScienceDirect

Psychiatry Research

journal homepage: www.elsevier.com/locate/psychres

Predicting non-response to ketamine for depression: An exploratory symptom-level analysis of real-world data among military veterans

Eric A. Miller [a,b,1], Houtan Totonchi Afshar [a,b], Jyoti Mishra [b,c], Roger S. McIntyre [d,e,f], Dhakshin Ramanathan [a,b,c,*]

[a] Department of Mental Health, VA San Diego Medical Center, San Diego, CA 92161, USA
[b] Department of Psychiatry, UC San Diego, La Jolla, CA 92093, USA
[c] Center of Excellence for Stress and Mental Health, VA San Diego Medical Center, USA
[d] Department of Psychiatry, University of Toronto, Toronto, Canada
[e] Department of Pharmacology, University of Toronto, Toronto, Canada
[f] Brain and Cognition Discovery Foundation, Toronto, Canada

ARTICLE INFO

Keywords:
Ketamine
Esketamine
Treatment resistant depression
Predictive modeling
Symptom trajectories

ABSTRACT

Ketamine helps some patients with treatment resistant depression (TRD), but reliable methods for predicting which patients will, or will not, respond to treatment are lacking. Herein, we aim to inform prediction models of non-response to ketamine/esketamine in adults with TRD. This is a retrospective analysis of PHQ-9 item response data from 120 patients with TRD who received repeated doses of intravenous racemic ketamine or intranasal eskatamine in a real-world clinic. Regression models were fit to patients' symptom trajectories, showing that all symptoms improved on average, but depressed mood improved relatively faster than low energy. Principal component analysis revealed a first principal component (PC) representing overall treatment response, and a second PC that reflects variance across affective versus somatic symptom subdomains. We then trained logistic regression classifiers to predict overall response (improvement on PC1) better than chance using patients' baseline symptoms alone. Finally, by parametrically adjusting the classifier decision thresholds, we identified optimal models for predicting non-response with a negative predictive value of over 96 %, while retaining a specificity of 22 %. Thus, we could identify 22 % of patients who would not respond based purely on their baseline symptoms. This approach could inform rational treatment recommendations to avoid additional treatment failures.

1. Introduction

Despite effective treatments for depression, such as psychotherapy (Cuijpers et al., 2020) and monoamine-based antidepressants (Cipriani et al., 2018), many patients do not experience remission even after multiple trials of different treatments (Rush et al., 2006; McIntyre et al., 2023). With a novel mechanism of action, ketamine offers hope for such patients with treatment resistant depression (TRD) (McIntyre et al., 2021). However, our ability to predict which patients will respond to ketamine remains limited.

Prior studies have identified potential moderators of response, such as obesity (Niciu et al., 2014; Freeman et al., 2020), family history of alcohol use disorder (Niciu et al., 2014; Phelps et al., 2009), and concomitant benzodiazepine use (Andrashko et al., 2020). Baseline anhedonia and anxious distress were identified as positive predictors of response to esketamine in one study (Pettorruso et al., 2023), and early symptomatic change has been identified as a potential predictor of remission with IV ketamine (Lipsitz et al., 2021). Despite these promising findings, a recent large meta-analysis failed to detect any consistent patient-level predictors of response to ketamine (Price et al., 2022). In the absence of consistent predictors of response, it remains difficult to stratify and select the optimal treatment for a particular patient among a growing list of options for TRD, including electroconvulsive therapy (ECT) (Van Diermen et al., 2018), transcranial magnetic stimulation (TMS) (Berlim et al., 2013), magnetic seizure therapy (MST) (Kayser et al., 2015), deep brain stimulation (DBS) (Figee et al., 2022), and

* Corresponding author at: Department of Psychiatry, University of California San Diego, 9500 Gilman Drive, La Jolla CA 92093, USA.
E-mail address: dramanathan@ucsd.edu (D. Ramanathan).
[1] Co-Corresponding author: Department of Psychiatry, University of Arizona College of Medicine – Tucson, 1501N Campbell Ave, Tucson, AZ 85724.

https://doi.org/10.1016/j.psychres.2024.115858
Received 3 November 2023; Received in revised form 4 March 2024; Accepted 9 March 2024
Available online 11 March 2024

psychedelic-assisted psychotherapy (Goodwin et al., 2022).

Prediction fundamentally rests on a choice of how to measure changes in the outcome of interest, in this case changes in depression. Most clinical studies utilize questionnaire sum scores to quantify depression before and after treatment, with response defined via a proportional drop in the sum score. With respect to validated questionnaires, sum scores can mask individual differences or dynamics at the level of symptoms (Borsboom and Cramer, 2013; EI Fried and Nesse, 2015) or dimensional features of illness (Cuthbert and Insel, 2013). For example, different symptom clusters may have distinct etiological and physiological underpinnings, which may, in turn, respond differently to treatments (Chekroud et al., 2017; EI Fried and Nesse, 2015). Although many studies have reported ketamine's effects specifically on suicidality (Jollant et al., 2023), we are aware of only a few studies that have compared how ketamine affects different depression symptoms or symptom clusters (Floden et al., 2022; Park et al., 2020; Chen et al., 2021; Rodrigues et al., 2020). None of those studies have modeled trajectories of symptoms over time.

Our first goal, therefore, was to model how individual symptoms of depression change over time for patients undergoing repeated ketamine treatments. This is a secondary analysis of real-world clinical data from a population of military veterans, most of whom have numerous psychiatric comorbidities. Over two-thirds of the patients had a diagnosis of PTSD in addition to major depressive disorder. As such, these data reflect how ketamine may be expected to perform for the treatment of TRD among complex patients, the very patients who could benefit most from innovative treatments. In earlier studies from the same population, our group has found that ketamine is effective for these patients, but with lower response rates than commonly reported in clinical trials (Artin et al., 2022; Bentley et al., 2022). This further highlights the need for reliable predictors of response or, perhaps more importantly, predictors of non-response to ketamine among complex patients.

As a secondary goal, building on our models of symptom trajectories, we used machine learning classifiers to predict whether patients would respond to ketamine using their baseline item by item PHQ-9 symptom scores. We also developed a model that can predict non-response, i.e. treatment failure, with very high confidence for a meaningful subset of the patients. This approach, based on a symptom level analysis, informs personalized treatment assignment considerations. Our method for identifying patients who are unlikely to respond to ketamine could prove useful, if replicated, for guiding treatment recommendations among a growing list of interventions targeting TRD.

2. Methods

2.1. Patients

Data were obtained from 120 patients who underwent serial ketamine or esketamine induction treatments for depression at the San Diego Veterans' Administration hospital between January 2020 and June 2022. 85 patients were male and 35 were female. Ages ranged from 26 to 75 (mean 45, standard deviation 12) years. While only 92 % (110 of 120) had a diagnostic code of MDD in the electronic medical record, all patients were specifically referred to the ketamine clinic for clinical depression. Among the 10 patients without a diagnostic code of MDD, their average baseline PHQ-9 score was 18.2, and all but one had a baseline PHQ-9 score at least in the "moderate depression" range. Comorbidity was also very common. The most frequently co-occurring condition was PTSD in 73 % of patients, with less frequent conditions including generalized anxiety disorder, bipolar disorder, ADHD, borderline personality disorder, and various substance use disorders. This study was approved as an institutional review board (IRB) exempt study by the local VA institutional review board (IRB 1,223,219).

2.2. Treatments

Ketamine was administered via either the intranasal (esketamine, $n = 99$), intravenous (racemic, $n = 20$), or intramuscular (racemic, $n = 1$) routes. Individual patients received the same route and formulation of ketamine for all sessions. Intranasal esketamine doses typically started at 56 mg and were increased to 84 mg after the initial session. Intravenous doses had more variability, as they were dosed at 0.5 to 1 mg/kg, with doses adjusted based on tolerability, side effects and efficacy. Formal psychotherapy was not paired with ketamine sessions, though psychological support was available if needed. Most patients (110 of 120) completed at least eight treatment sessions, with the remaining completing between 2 and 7 sessions each.

2.3. Analysis of average PHQ-9 scores

This is a secondary analysis of PHQ-9 data. Patients completed PHQ-9 questionnaires at baseline and prior to ketamine treatment sessions. Supplemental Table 1 lists the full text for the nine items of the PHQ-9. Average PHQ-9 sum scores across patients were computed across the first eight treatment sessions. Separately, average PHQ-9 item responses were computed across the eight sessions. Significant change in sum score was defined as $p < 0.05$ on the Friedman chi-square test for repeated measures, as the first two sessions' data violated normality ($p < 0.05$, Shapiro-Wilk test). Significant change in item response was defined as $p < 0.006$ (after Bonferroni correction for multiple comparisons) on the Friedman test. Only patients who completed a PHQ-9 before all eight sessions (82 of 120) were included in these repeated measures tests.

2.4. Analysis of item trajectories

Item responses were analyzed across treatment sessions for each item, i, and for each patient, p. Linear (Eq. (1)) and exponential (Eq. (2)) models were fit to item response trajectories by minimizing the residual sum-squared errors (RSS). In the linear model, m is the linear slope, t is time in days, and b is the intercept. In the exponential model, a is a scaling factor, m is a growth factor, t is time in days, and b is a constant offset. In both models, m reflects a rate of change in item response.

$$y_{i,p}(t) = m_{i,p}\ t + b_{i,p} \tag{1}$$

$$y_{i,p}(t) = \ a_{i,p}\ e^{-m_{i,p}\ t} + b_{i,p} \tag{2}$$

For each model, the Akaike information criterion (AIC) and Bayesian information criterion (BIC) were computed for the purposes of model selection. Specifically, AIC and BIC were computed using the RSS from model fitting, assuming independent and identically distributed (IID) residuals following a normal distribution. For each patient, the winning model was defined as the model with lowest AIC and BIC for the majority of items. In turn, the best overall model for a given questionnaire was defined as the winning model for the majority of patients. The linear model, as the best overall, was used as the basis of all subsequent analyses. Differences in linear slopes between the items was defined as $p < 0.05$ on the Friedman test for repeated measures, with post-hoc evaluation for pairwise differences via Wilcoxon signed-rank test with Bonferroni correction for all pairwise comparisons ($p < 0.05 / 36 = 0.0014$). Effect sizes for pairwise differences was defined as the difference in medians divided by the average of the two interquartile range (IQR) values (Ricca and Blaine, 2022).

2.5. Principal component analysis for linear slopes

Using linear models (Eq. (1)), each patient had nine slope parameters describing their change in each of the PHQ-9 items. Principal components analysis (PCA) was computed in the 9-dimensional space of slopes. PCA finds the set of orthogonal linear combinations which capture maximal variance across participants. Parallel analysis was used to

evaluate the strength of principal components (PCs), in which the actual eigenvalue of each PC was compared with the 95th percentile of the distribution of eigenvalues for that PC from 10,000 randomly generated datasets. Coefficients for each significant component were analyzed to understand directions of change.

2.6. Predicting treatment response

Treatment response was defined via the sign for the first principal component (PC1), which by definition captures the majority of variance in the data, and in our case represented a weighted rate of change in symptoms. Due to data normalization, a negative sign implies a greater improvement along PC1 than the mean across participants. To relate this to classical measures of response, percent changes in PHQ-9 sum scores were also evaluated. Logistic regression was used to predict treatment response using patients' baseline PHQ-9 item responses. An exhaustive feature selection was conducted to evaluate all 511 possible subsets of PHQ-9 items as features. For each model, classification performance was evaluated across 1000 iterations of repeated 5-fold cross validation.

2.7. Threshold tuning for high confidence predictions

Logistic regression provides probabilities of test items belonging to each class. The choice of classification threshold results in a tradeoff between positive predictive value (PPV) and sensitivity, or between negative predictive value (NPV) and specificity. For all 511 possible subsets of features, a parametric search was conducted across 20 different classification thresholds ranging from 0 to 1. This resulted in over 10,000 distinct classification models, which then underwent cross validation as before. To optimize for potential clinical utility, models

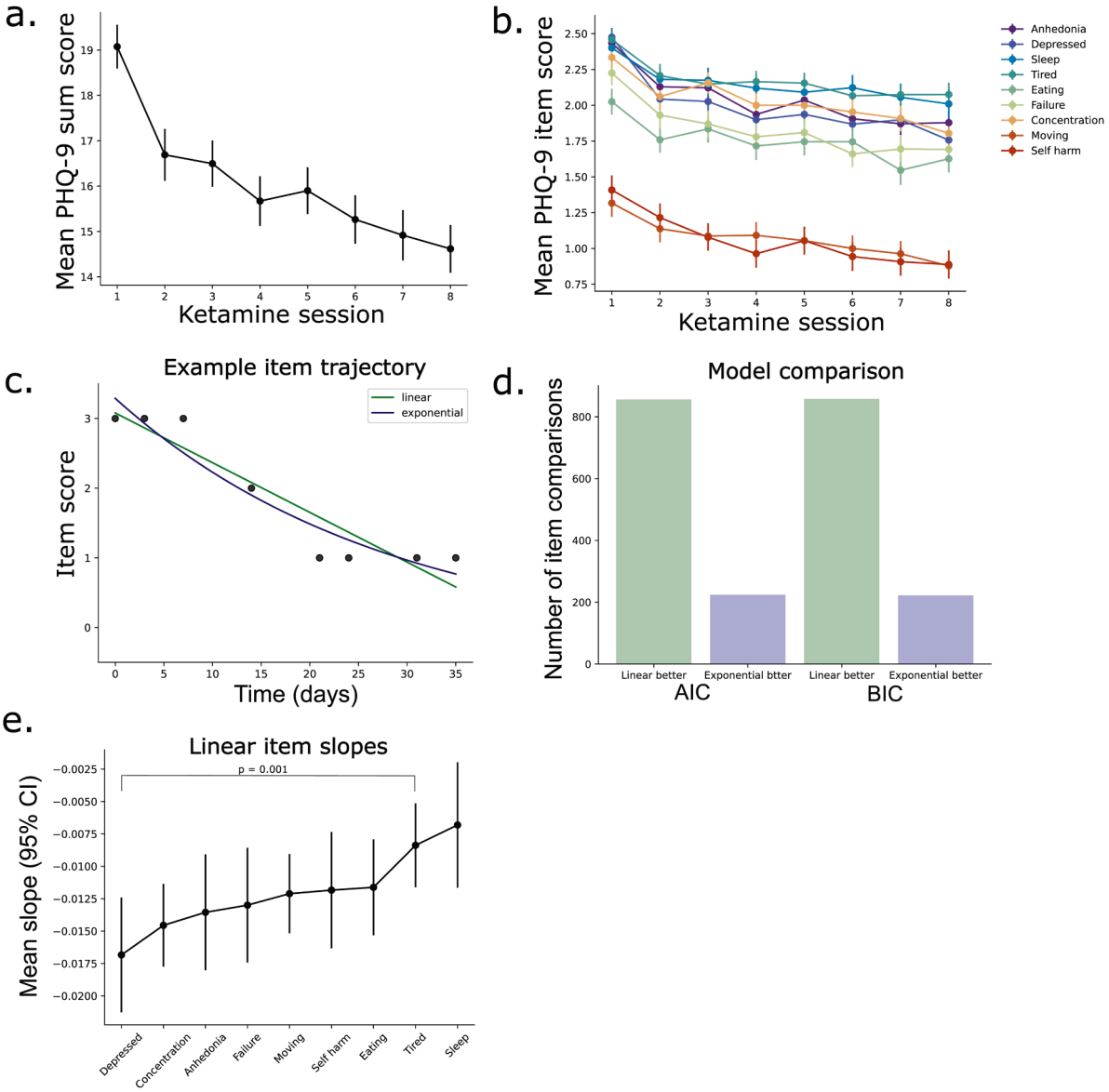


Fig. 1. Effects of ketamine on individual symptoms of depression. (a) Mean (SEM) PHQ-9 sum scores across patients at each of the first eight ketamine sessions. (b) Mean (SEM) PHQ-9 item scores across patients at each of the first eight ketamine sessions, for all nine items. (c) One example patient's item responses for one example item (PHQ-9 item #1). Green and blue curves are the linear and exponential model fits, respectively. (See Methods for model equations). (d) Bar plots comparing linear and exponential model fits using AIC (left) and BIC (right), such that the bar heights indicate the number of patient-items where the linear (green) or exponential (blue) model was a better fit. (e) Mean (95 % CI) linear slopes across patients for each of the PHQ-9 items. P-value shown for pairwise comparison with significant difference after Bonferroni correction for multiple comparisons. See Supplemental Table 1 for mapping of item names to PHQ-9 item text (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

were selected that met a performance specification of either: (1) at least 90 % PPV and at least 10 % sensitivity for predicting "response", or (2) at least 90 % NPV and at least 10 % specificity for predicting "non-response". These criteria were chosen because very high predictive value could provide clinically actionable information, even if only for a subset of the patients. In contrast, models with lower predictive value are unlikely to change clinical management, even if they have higher sensitivity. Among the models meeting these specifications, performance was ranked by the arithmetic mean of predictive value (PPV or NPV) and coverage of the relevant cases (sensitivity or specificity), with the best model defined as the one with highest mean of those two characteristics. To evaluate generalizability of features, the distribution of regression coefficients across all models meeting the specified performance criteria were plotted.

2.8. Code and data availability

All data analysis was conducted with custom Python software utilizing open-source scientific and machine learning packages. The code to reproduce all results and figures is available at github.com/angevine-Miller/ketaminePrediction. Data are available upon request.

3. Results

This exploratory study was motivated by two major goals. First, we sought to understand the dynamics of depression symptoms over the course of repeated ketamine sessions. Second, using a symptom-level modeling approach, we aimed to predict treatment response for patients in a real-world clinical setting. We first confirmed that average PHQ-9 sum scores improved across the ketamine treatment sessions (Fig. 1a) ($p < 10^{-18}$, Friedman test). In addition, average responses for every individual PHQ-9 item improved over the course of treatment (Fig. 1b) (ps $< 10^{-5}$, Friedman test). We then proceeded to analyze symptom trajectories for each individual patient.

3.1. Modeling individual symptom trajectories across individuals

We sought the simplest model that could capture symptom changes over the course of treatment in the majority of individuals. In addition to a simple linear model (Eq. (1)), we also tested an exponential model (Eq. (2)), motivated by prior work showing that exponential functions can describe various interventions for depression (Priest et al., 1996; Berlow et al., 2021). We fit linear and exponential models to item response trajectories for each patient (Fig. 1c), and then evaluated the fit of both models using AIC and BIC. Linear models were better than exponential models for the majority of patients (104 of 120 for AIC, 105 of 120 for BIC) (Fig. 1d), so we chose the linear model as the basis for further analysis to maintain consistency/comparability across individuals.

The slopes of the linear models describe the rate of change in each symptom of depression for patients over the course of their ketamine treatment. The average slopes across patients were negative for all nine items, confirming that patients generally improved across all symptoms (Fig. 1e). We detected a difference in slopes between the items ($p = 0.012$, Friedman test). Post-hoc comparisons revealed that PHQ-9 item #2 (depressed mood) had significantly steeper slopes than item #4 (tiredness) after Bonferroni correction ($p = 0.001$, Wilcoxon signed-rank test) (Fig. 1e). Thus, symptoms of depressed mood improved more rapidly than tiredness across ketamine treatment, with a difference in medians of about 45 % of the interquartile range (see Methods).

3.2. Low dimensional variables to capture trajectory changes

Next, we wondered whether we could parsimoniously capture how patients varied in the high-dimensional space of symptom trajectories. To understand this we computed a principal components analysis (PCA) on the trajectory slopes calculated for each item of the PHQ-9 across subjects. We detected two significant principal components (PCs), which explained 52.2 % and 16.2 % of the variance, respectively (significance calculated using parallel analysis, see Methods) (Fig. 2a). Analysis of coefficients revealed that the first PC (PC1) described a weighted rate of change for all nine symptoms in the same direction (Fig. 2b). Coefficients of the second PC (PC2) showed opposite signs for somatic symptoms (e.g. appetite, energy, concentration, movement) and affective symptoms (e.g. mood, anhedonia, thoughts of self-harm or of being a failure) (Fig. 2c). Projecting participants' slopes data onto these two PCs revealed the distribution of patients across these two dimensions of symptom variance (Fig. 2d).

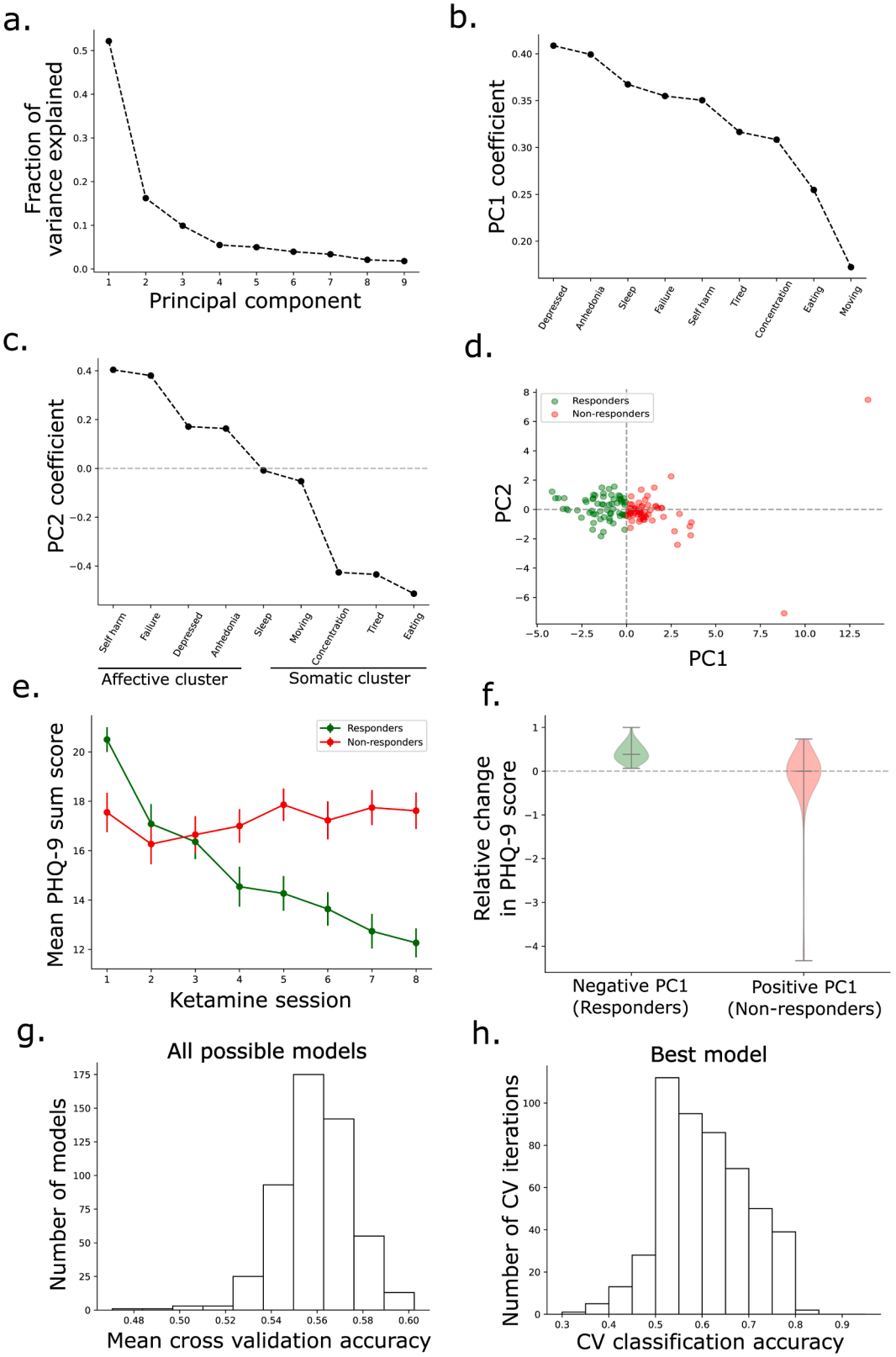
Based on the coefficient weights, the sign of the first principal component (PC1) represented whether a patient improved more (negative sign) or less (positive sign) than the mean rate of change in symptoms (Fig. 2d). This is because each patient's symptom slopes were standardized by the population mean slopes prior to computing PCA. Therefore, this offers a simple, data-driven way of differentiating treatment responders from non-responders. To better illustrate this, we first plotted average PHQ-9 sum scores for patients with negative PC1 (identified as responders, $n = 62$) compared to patients with non-negative PC1 (identified here as non-responders, $n = 58$) (Fig. 2e). The sum scores among responders improved significantly between ketamine induction sessions ($F(7336) = 46.7$, $p < 10^{-45}$, partial eta-squared = 0.49, repeated measures ANOVA). The mean change in sum scores between baseline and the final session among responders was −8.7 points (95 % CI: −9.8 to −7.6). In contrast, we detected no differences in sum scores between sessions among the non-responders ($F(7224) = 2.07$, $p = 0.07$, Greenhouse-Geisser correction, repeated measures ANOVA). The mean change in sum scores among non-responders was −0.40 (95 % CI: −2.0 to 1.2).

Based on the distribution of PC2 coefficients across symptom subdomains, we hypothesized that the sign of the 2nd PC (PC2) could reflect whether there were relatively greater improvements in the affective symptom clusters (negative sign) or somatic symptom clusters (non-negative sign). To test this, we calculated (for responders only) the trajectories of average scores for the affective and somatic subdomains of symptoms, grouped using the PC2 as noted above. Responders with a non-negative PC2 showed greater improvements in somatic symptoms, compared to responders with a negative PC2 ($F(7, 329) = 3.03$, $p = 0.004$, mixed ANOVA interaction) (Supplemental Fig. 1a). In contrast, we were unable to detect a significant difference in the rate of improvement in affective symptoms based on PC2 sign ($F(7, 329) = 1.01$, $p = 0.42$, mixed ANOVA interaction) (Supplemental Fig. 1b). Thus, the sign of PC2 reliably differentiated patients based on their rate of improvement on the somatic subdomain, suggesting improvement in this domain of symptoms reflected another meaningful source of variance even within those who were categorized as responding overall.

3.3. Predicting response from baseline symptoms

We identified above a data-driven approach to characterizing the antidepressant response to ketamine treatments using changes in individual item scores of the PHQ-9. Traditional measures of treatment response are defined with respect to the percent change in the overall PHQ-9 sum score. To ensure that our categorization corresponds at least roughly with more standard definitions of treatment response, we computed the percent reduction in PHQ-9 sum score from baseline to the last treatment session for patients categorized by negative PC1 (responders) and non-negative PC1 (non-responders). Among responders, the median percent change in PHQ-9 sum scores was a reduction by 39 % (IQR 25 %). Among non-responders, the median percent change in sum scores was 0 % (IQR 22 %) (Fig. 2f), suggesting that this method of classifying subjects was a reasonable data-driven measure of categorizing response.

Using this categorization, we first asked whether treatment response



(caption on next page)

Fig. 2. Predicting treatment response from baseline symptoms, (a) Scree plot showing fraction of total variance explained by each principal component (PC) from a PCA over the nine-dimensional data of linear slopes for each patient. Significance of PCs was determined by bootstrapping with parallel analysis (see Methods). (b) Coefficients of the first PC of the data for each of the features, corresponding to PHQ-9 item slopes. (c) Coefficients of the second PC of the data for each of the features, corresponding to PHQ-9 item slopes. (d) Scatter plot of the projections of each patient's set of linear slopes onto the first two principal components of the data, colored according to the sign of the first principal component. Responders (green) are defined as having a negative sign of the first PC, while non-responders (red) are defined as having a positive sign of the first PC. (e) Mean (SEM) PHQ-9 sum scores across patients for the first eight ketamine sessions for responders and non-responders. (f) Violin plots of relative change in PHQ-9 sum score from baseline to last session for responders (green) and non-responders (red). Bars indicate median and extreme values, and width of violin indicates distribution density. Dashed horizontal line indicates 0 change. (g) Histogram of the mean cross-validation accuracy for all models from an exhaustive feature selection over all possible subsets of baseline PHQ-9 items as model features. (h) Histogram of accuracies across all iterations of cross-validation for the best model from the exhaustive feature selection, i.e. the model with overall best mean CV accuracy. See Results text for the other key model performance characteristics for this model. For (b) and (c), see Supplemental Table 1 for mapping of item names to PHQ-9 item text (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

was related to basic factors such as demographics (age, gender) and treatment formulation (IV racemic ketamine vs. intranasal esketamine). We detected no differences in average PC1 value based on ketamine formulation (esketamine vs IV racemic) ($p = 0.59$, two-sided t-test), gender ($p = 0.19$, two-sided t-test) or age ($r = 0.13$, $p = 0.17$, Pearson's correlation), suggesting these simple factors would not be informative

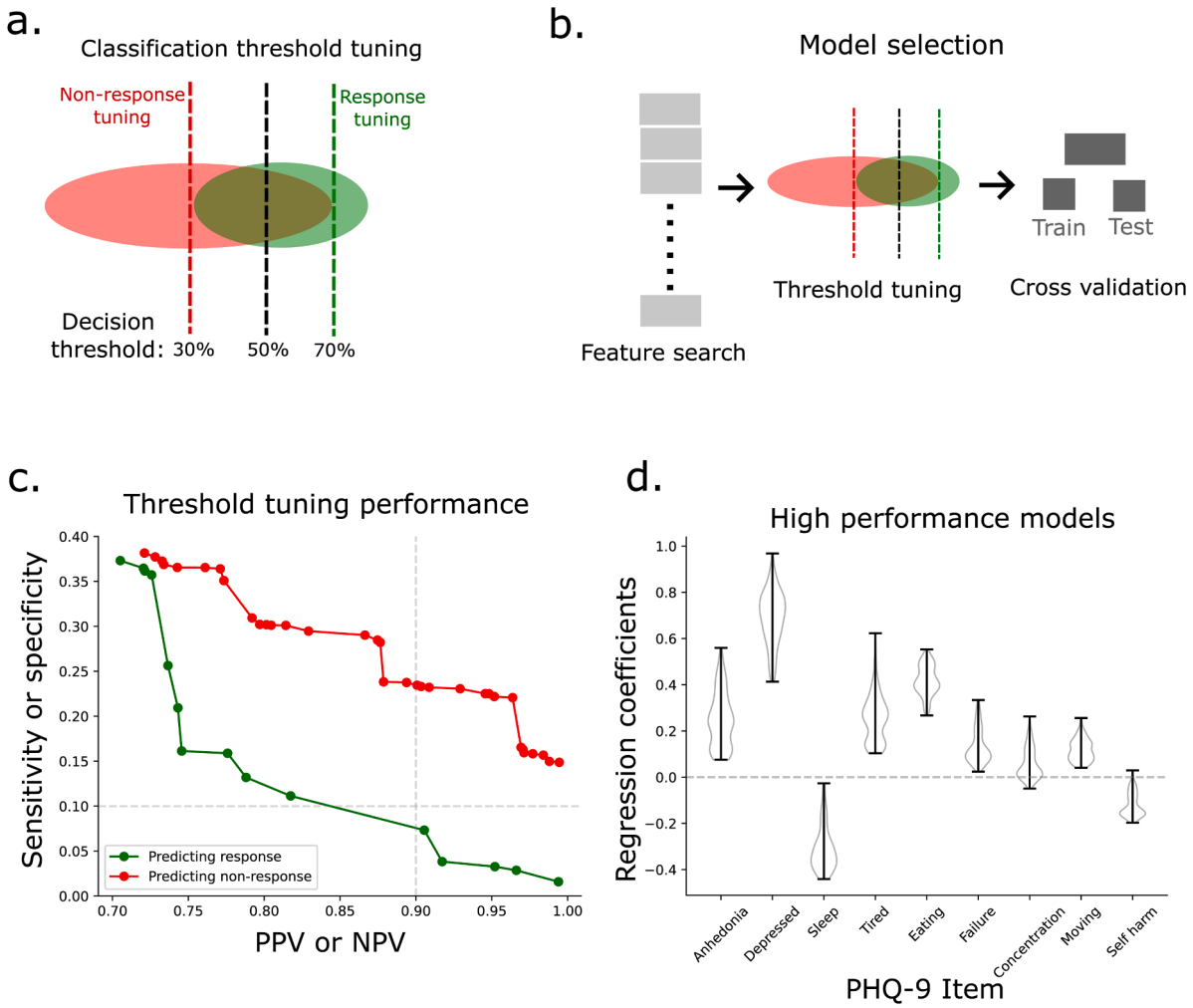


Fig. 3. Threshold tuning to confidently predict non-response, (a) Diagram depicting threshold tuning approach. Red and green ellipses represent a hypothetical projection of data for non-responders and responders, respectively. Black dashed line indicates standard decision threshold for logistic regression, which maximizes accuracy. Green and red dashed lines indicate alternate choices for the decision threshold, which maximize positive predictive value (PPV) for predicting response or negative predictive value (NPV) for predicting non-response, respectively. (b) Diagram depicting three major steps of model selection procedure: feature search, threshold tuning, and cross validation. Grey boxes represent the different sets of baseline PHQ-9 items that can be tried as a feature set. Threshold tuning, as in (a), was conducted for each feature sets, resulting in over 10,000 models that then underwent cross validation. (c) Scatter plot of cross validation model performance for some of the best models for predicting response (green) or non-response (red). Plot shows the inherent tradeoff between predictive value (PPV or NPV) and coverage of the relevant cases (sensitivity or specificity). Green dots represent the best models, in terms of highest sensitivity (y-axis), among all of the models with at least a minimum PPV (x-axis). Red dots represent the best models, in terms of highest specificity (y-axis), among all of the models with at least a minimum NPV (x-axis). Dashed grey lines show cutoffs for our predefined model performance specifications: PPV or NPV > 90 %, sensitivity or specificity > 10 % (upper right quadrant). (d) Violin plots showing the distribution of regression coefficients for each PHQ-9 item, across all models meeting the performance criteria of NPV > 90 % and specificity > 10 % (upper right quadrant in panel c). Positive coefficients favored response, and negative coefficients favored non-response. Dotted horizontal line indicates a coefficient of 0 (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

for predicting treatment response (data not shown).

We next asked whether treatment response could be predicted from patients' baseline symptoms of depression. Specifically, we used logistic regression to predict the sign of PC1 for individual patients, using patients' baseline single-item PHQ-9 item scores as features. To identify the best features for prediction, we conducted an exhaustive feature selection by fitting logistic regressions with all 511 possible subsets of PHQ-9 items as features. For each of these 511 models, performance was evaluated with repeated 5-fold cross-validation (CV). Nearly all of the models (509 of 511, 99.6 %) had better-than-chance average CV accuracy, though all were only moderately better than chance (Fig. 2g).

The model with best CV accuracy used only two PHQ-9 items as features: item #1 (anhedonia) and item #6 (feeling of failure). Higher baseline responses on either item predicted subsequent treatment response. For this model, the average classification accuracy to holdout data was 60.3 % (95 % confidence interval 59.5 – 61.1 %) (Fig. 2h). The precision (PPV) was 60.2 %, recall (sensitivity) was 68.3 %, and F1 score was 64.0 %. The second-best model used only one item as a feature, item #2 (depressed mood), with nearly as good performance as the best model: 60.0 % average accuracy (95 % CI 59.2 % - 60.8 %), 60.0 % precision, 67.7 % recall, and 63.6 % F1 score. Patients with higher baseline depressed mood were more likely to respond.

3.4. Threshold tuning for high confidence predictions

Although the above approach identified many models that can predict treatment response better than chance, none of those models had excellent performance characteristics. They had a maximum of around 60 % predictive value, which is unlikely to provide sufficient prediction confidence for changing clinical management. We reasoned that a more clinically useful model would have a very high predictive value (PPV or NPV). Such a model would provide clinically actionable information, even if only for a subset of the patients. For example, if a model could predict with over 90 % NPV that a patient will not respond to ketamine, then we could be quite confident in recommending an alternative treatment for that patient.

Seeking such a model, we used threshold tuning in order to optimize the classifiers toward high predictive confidence (Fig. 3a). Specifically, instead of using probability 0.5 as the threshold to classify patients as responders and non-responders, we evaluated a range of thresholds between 0 and 1 (see Methods). Changing the threshold in this way necessarily results in a tradeoff between prediction confidence (PPV or NPV) and coverage of the relevant cases (sensitivity or specificity). For model selection, we conducted another exhaustive search across all possible subsets of baseline PHQ-9 items, and for each of these feature sets, we applied threshold tuning to evaluate a range of thresholds (Fig. 3b). This resulted in over 10,000 distinct models that were cross validated to evaluate model performance on holdout data (Fig. 3c).

We then selected only those models with very high predictive value, while at the same time retaining a minimum coverage of the relevant cases (Fig. 3c, upper right quadrant). Specifically, we searched for models with either (1) at least 90 % PPV and at least 10 % sensitivity, or (2) at least 90 % NPV and at least 10 % specificity. We identified hundreds of models that, with the appropriate decision threshold, could predict non-response with those performance criteria. The best of these models (see Methods) had a NPV of 96.4 %, while retaining a specificity of 22.1 %. This model used only three baseline PHQ-9 items as features: item #2 (depressed mood), item #5 (changes in appetite or eating), and item #9 (self-harm or suicidal ideation). Relatively speaking, lower values on item #2 and item #5 and higher values on item #9 favored non-response for this model.

To determine the generality of this observation, we plotted distributions of coefficients across all models meeting the performance criteria of >90 % NPV and >10 % specificity. Of these models, higher values on item #9 (self-harm or suicidal ideation) favored non-response in 98.7 % (224 of 227) models in which it was a feature (Fig. 3d). Item #3 (difficulty sleeping) also consistently favored non-response for 100 % (130 of 130) of the models in which it was a feature (Fig. 3d). In contrast, lower values on the other seven items favored non-response in the majority of models, and lower values on item #1 (anhedonia), item #2 (depressed mood), item #4 (tiredness or low energy), and item #5 (changes in appetite or eating) were particularly consistent predictors of non-response (Fig. 3d). Taken together, these findings suggest that relatively speaking non-responders' symptoms were weighted more towards sleep difficulties and suicidal ideation, as compared to the other symptoms of depression (Fig. 3d).

Thus, using threshold tuning, we were able to identify classification models that predict non-response with very high confidence (over 96 % NPV) for a subset of 22 % of the patients, based purely on patients' baseline PHQ-9 symptom scores. In contrast, we detected no models for predicting treatment response at the specified performance cutoffs (Fig. 3c, green).

4. Discussion

In this study we developed an approach for modeling changes in symptoms of depression over the course of repeated ketamine sessions. Using this approach, we found that all symptoms improved across the course of ketamine treatment, and the symptom of depressed mood improved more rapidly than the symptom of low energy (Fig. 1). We found a range of individual differences in these item response trajectories, both in the degree of change in overall depression level, and in specific subdomains of symptoms (Fig. 2). We developed logistic regression classifiers, which can predict better than chance whether patients will respond to ketamine, using their baseline symptoms alone (Fig. 2). Finally, using threshold tuning, we found classifiers that can identify a subset of patients who are highly *unlikely* to respond to ketamine with over 96 % predictive value (Fig. 3). Our findings shed light on how ketamine affects specific symptoms and dimensional features of depression, and the method for identifying non-responders may prove useful for informing rational treatment recommendations among a growing list of novel therapeutics for treatment resistant depression (Berlim et al., 2013; Kayser et al., 2015; Figee et al., 2022; Goodwin et al., 2022).

Few prior studies have analyzed how different depression symptoms respond to ketamine, particularly across repeated doses. Floden et al. (2022) conducted a symptom-level analysis of data from the TRANSFORM 2 trial of repeated intranasal esketamine for TRD. They found that eight twice-weekly doses of esketamine plus oral antidepressant led to improvements in all PHQ-9 items except item #9 concerning suicidality, which was expected due to exclusion of patients with high suicidality in that trial (Floden et al., 2022). Our data, which did not exclude patients with high suicidality, found that all nine symptoms improved including suicidality. However, there were differences in the rates at which some of the symptoms improved across repeated ketamine sessions.

The symptom of depressed mood improved faster than the symptom of low energy. This supports results from Chen et al. (2021), who found that a single infusion of low-dose IV ketamine resulted in greater reductions in cognitive and affective symptoms, compared to somatic symptoms (Chen et al., 2021). Similarly, Park et al. (2020) found that typical/melancholic symptoms improved more rapidly than atypical symptoms of depression after a single dose of IV ketamine (Park et al., 2020). Furthermore, using principal component analysis (PCA), we found significant individual variation in whether patients improved more in affective symptoms (e.g. depressed mood, anhedonia, suicidal ideation) or in somatic symptoms (e.g. changes in appetite, concentration, or energy level) suggesting that symptom response to ketamine is patient-specific. These patient-specific differences in symptom response, and also baseline symptoms, may inform predictive models. For example, Pettorruso et al. (2023) found that certain baseline symptoms such as anhedonia predicted response to esketamine.

Here, using machine learning classifiers, we also were able to predict

which patients would respond better than chance using their baseline symptoms alone. A key factor was a difference in baseline depression severity, such that responders had more severe baseline depression than non-responders. These results may be counterintuitive, as one might expect patients with more severe depression to be more treatment resistant. A study by Jesus-Nunes et al. (2022) found that patients with more severe depression were less likely to respond to a single infusion of IV ketamine or esketamine. On the other hand, our data may reflect the consistent finding that antidepressants separate from placebo most prominently among patients with severe baseline depression (Khan et al., 2002; Kirsch et al., 2008; Fournier et al., 2010), which has also been observed for ketamine (Su et al., 2017). It is unclear how much of that is due to stronger effects of the active drug or weaker effects of the placebo among patients with more severe depression. It is also possible that baseline depression severity may signal different underlying psychopathologies, with consequent differences in response to ketamine, but future studies will be necessary to explore that possibility.

Importantly, the standard machine learning classifiers noted above are not likely to be useful in the clinic as they can only predict response with about 60 % accuracy and positive predictive value (Fig. 2). Similarly, Pettorruso et al. (2023) were able to predict response to esketamine with an overall accuracy of about 68 %. We reasoned that very high predictive values (over 90 % PPV or NPV) would be required for a model to guide clinical decision making, particularly when deciding not to recommend a potential treatment in individuals with severe or treatment-refractory conditions. We therefore optimized our classifiers for either high PPV for identifying responders, or for high NPV for identifying non-responders. We did this via threshold tuning, in which we searched over the full space of decision thresholds, and then selected classifiers with optimal cross-validated performance characteristics. Using this approach, we found models that could predict *non-response* to ketamine, i.e. treatment failure, with over 96 % NPV.

This excellent NPV necessarily comes at the expense of a higher false-positive rate, so we detect only a subset of the non-responders (22 %). However, the tradeoff is that for those patients who it does identify as non-responders, we will be fairly confident that they will in fact not respond. As such, this model provides clinically actionable information for those patients. Regarding specific predictors, we found that *more* severe suicidal ideation and sleep difficulties were consistent predictors of non-response, whereas *less* severe responses to the other symptoms of depression favored non-response. Less severe anhedonia, depressed mood, tiredness, and appetite symptoms were particularly strong and consistent predictors of non-response. Intriguingly, though, the only symptoms for which higher scores consistently predicted non-response were sleep problems and suicidality.

The latter aligns with Pettorruso et al. (2023), who also found that suicidality predicted non-response to esketamine, whereas anhedonia and hopelessness were positive predictors of response. It is important to note that higher suicidality could predict non-response to ketamine, while at the same time ketamine may reduce suicidality for patients who respond (Jollant et al., 2023). Put differently, ketamine was less likely to be effective for patients with a particular pattern of symptoms, characterized by higher suicidality and sleep problems relative to the other symptoms of depression like anhedonia and depressed mood. Hypothetically, this specific pattern of symptoms may be an indicator of a more difficult to treat depressive illness. It could also suggest the presence of comorbid difficult-to-treat conditions, including personality disorders or PTSD, which were common among these patients. However, for those patients who did respond, all nine symptoms of depression, including suicidality and sleep, were likely to improve. Future pre-registered and controlled studies will be necessary to test whether these findings replicate and generalize. The simple approach of threshold tuning to find clinically meaningful predictions may be useful for other treatments as well.

4.1. Limitations

There are several important limitations of this study. First, this is a non-registered secondary analysis, so the results are fundamentally exploratory and require confirmation with pre-registered and experimental studies. These are real world clinical data without placebo or wait list control groups, making it impossible to distinguish the role of nonspecific factors including placebo. For the same reason, we cannot exclude the role of regression to the mean, which is a possible explanation for observing improvement among patients with more severe baseline depression. However, regression to the mean is unlikely to be the only driver of this finding, since the average trajectories for responders and non-responders did not converge to a common mean value.

An intrinsic limitation of the threshold tuning approach is that excellent negative predictive values (NPV) come at the expense of higher false positive rates, i.e. lower specificity. Our best model has over 96 % NPV at the expense of identifying a subset of 22 % of the actual non-responders. Our rationale is that a very high NPV is required for clinicians and patients to decide *not* to pursue a potentially life-saving intervention for TRD. This allows a clinician interpreting results for an individual patient to confidently recommend alternative treatments, such as ECT or TMS, instead of ketamine. A future research direction would be to improve model specificity without sacrificing NPV. However – again, it is important to note that even in our current model, 22 % of individuals who were referred to the VA ketamine program would not have had to trial the treatment, indicating an immediate clinical benefit to both patients and resource management.

Another potential limitation is the real-world nature of these data, in which patients have numerous psychiatric and medical comorbidities. It is possible that these comorbidities affect the specific ways in which patients respond to ketamine. At the same time, this may improve the generalizability of our results to real-world clinical contexts where comorbidities are common – and is thus also reasonably a strength of this analysis. It is also important to highlight that psychotherapy was not provided in concert with ketamine treatments, so our data do not speak to the potential for ketamine-assisted psychotherapy to prolonging or change its effects (Joneborg et al., 2022). Finally, we lack detailed patient accounts of their treatment, which would be helpful for understanding patient-centered and functional outcomes beyond those limited data that are reflected on the PHQ-9 questionnaire.

5. Conclusion

Repeated ketamine and esketamine led to progressive improvements in all symptoms of depression, but depressed mood improved faster than low energy. Using machine learning classifiers, we could predict non-response to ketamine with very high confidence for a subset of the patients. If validated in future studies, this model could be useful for identifying patients unlikely to benefit from ketamine, who might be better served by other treatments.

Funding

Center of Excellence for Stress and Mental Health, Burroughs Wellcome Fund, Career Award for Medical Scientists

CRediT authorship contribution statement

Eric A. Miller: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Houtan Totonchi Afshar:** Data curation, Writing – review & editing. **Jyoti Mishra:** Formal analysis, Methodology, Writing – review & editing. **Roger S. McIntyre:** Supervision, Writing – review & editing. **Dhakshin Ramanathan:** Data curation, Formal analysis, Funding acquisition,

Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

Dr. Roger S. McIntyre has received research grant support from CIHR/GACD/National Natural Science Foundation of China (NSFC) and the Milken Institute; speaker/consultation fees from Lundbeck, Janssen, Alkermes, Neumora Therapeutics, Boehringer Ingelheim, Sage, Biogen, Mitsubishi Tanabe, Purdue, Pfizer, Otsuka, Takeda, Neurocrine, Neurawell, Sunovion, Bausch Health, Axsome, Novo Nordisk, Kris, Sanofi, Eisai, Intra-Cellular, NewBridge Pharmaceuticals, Viatris, Abbvie and Atai Life Sciences. Dr. S. Roger McIntyre is a CEO of Braxia Scientific Corp. Dr. Miller, Dr. Afshar, Dr. Mishra, and Dr. Ramanathan declare no conflicts of interest.

Acknowledgment

D.R. is supported by funding from the VA Office of Research and Development funded Center of Excellence for Stress and Mental Health, National Institute for Mental Health and the Burroughs Wellcome Fund (Career Award for Medical Scientists).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.psychres.2024.115858.

References

Cuijpers, P., Karyotaki, E., de Wit, L., et al., 2020. The effects of fifteen evidence-supported therapies for adult depression: a meta-analytic review. Psychotherapy Res 30, 279–293.
Cipriani, A., Furukawa, T.A., Salanti, G., et al., 2018. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. Lancet 391, 1357–1366.
Rush, A.J., Trivedi, M.H., Wisniewski, S.R., et al., 2006. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report. Am. J. Psychiatry 163, 1905–1917.
McIntyre, R.S., Alsuwaidan, M., Baune, B.T., et al., 2023. Treatment-resistant depression: definition, prevalence, detection, management, and investigational interventions. World Psychiatry 22, 394–412.
McIntyre, R.S., Rosenblat, J.D., Nemeroff, C.B., et al., 2021. Synthesizing the evidence for ketamine and esketamine in treatment-resistant depression: an international expert opinion on the available evidence and implementation. Am. J. Psychiatry 178, 383–399.
Niciu, M.J., Luckenbaugh, D.A., Ionescu, D.F., et al., 2014. Clinical predictors of ketamine response in treatment-resistant major depression. J. Clin. Psychiatry 75, e417–423.
Freeman, M.P., Hock, R.S., Papakostas, G.I., et al., 2020. Body mass index as a moderator of treatment response to ketamine for major depressive disorder. J. Clin. Psychopharmacol. 40, 287–292.
Phelps, L.E., Brutsche, N., Moral, J.R., et al., 2009. Family history of alcohol dependence and initial antidepressant response to an N-methyl-D-aspartate antagonist. Biol. Psychiatry 65, 181–184.
Andrashko, V., Novak, T., Brunovsky, M., et al., 2020. The Antidepressant effect of ketamine is dampened by concomitant benzodiazepine medication. Front. Psychiatry 11.
Pettorruso, M., Guidotti, R., d'Andrea, G., et al., 2023. Predicting outcome with Intranasal Esketamine treatment: a machine-learning, three-month study in Treatment-Resistant Depression (ESK-LEARNING). Psychiatry Res. 327, 115378.
Lipsitz, O., McIntyre, R.S., Rodrigues, N.B., et al., 2021. Early symptomatic improvements as a predictor of response to repeated-dose intravenous ketamine: results from the Canadian rapid treatment center of excellence. Prog. Neuropsychopharmacol. Biol. Psychiatry 105, 110126.
Price, R.B., Kissel, N., Baumeister, A., et al., 2022. International pooled patient-level meta-analysis of ketamine infusion for depression: in search of clinical moderators. Mol. Psychiatry 27, 5096–5112.
Van Diermen, L., Van Den Ameele, S., Kamperman, A.M., et al., 2018. Prediction of electroconvulsive therapy response and remission in major depression: meta-analysis. Br. J. Psychiatry 212, 71 - 80.
Berlim, M.T., Van Den Eynde, F., Daskalakis, Z.J., 2013. A systematic review and meta-analysis on the efficacy and acceptability of bilateral repetitive transcranial magnetic stimulation (rTMS) for treating major depression. Psychol. Med. 43, 2245 - 2254.
Kayser, S., Bewernick, B.H., Matusch, A., et al., 2015. Magnetic seizure therapy in treatment-resistant depression: clinical, neuropsychological and metabolic effects. Psychol. Med. 45, 1073 - 1092.
Figee, M., Riva-Posse, P., Choi, K.S., et al., 2022. Deep Brain Stimulation for Depression. Neurotherapeutics 19, 1229–1245.
Goodwin, G.M., Aaronson, S.T., Alvarez, O., et al., 2022. Single-dose psilocybin for a treatment-resistant episode of major depression. New England Journal of Medicine 387, 1637–1648.
Borsboom, D., Cramer, A.O.J., 2013. Network analysis: an integrative approach to the structure of psychopathology. Annu. Rev. Clin. Psychol. 9, 91–121.
Fried, E.I., Nesse, R.M., 2015a. Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR*D study. J. Affect. Disord. 172, 96–102.
Cuthbert, B.N., Insel, T.R., 2013. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. BMC. Med. 11.
Chekroud, A.M., Gueorguieva, R., Krumholz, H.M., et al., 2017. Reevaluating the efficacy and predictability of antidepressant treatments: a symptom clustering approach. JAMA Psychiatry 74, 370–378.
Fried, E.I., Nesse, R.M., 2015b. Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. BMC. Med. 13, 72.
Jollant, F., Colle, R., Nguyen, T.M.L., et al., 2023. Ketamine and esketamine in suicidal thoughts and behaviors: a systematic review. Ther. Adv. Psychopharmacol. 13, 20451253231151327.
Floden, L., Hudgens, S., Jamieson, C., et al., 2022. Evaluation of individual items of the Patient Health Questionnaire (PHQ-9) and Montgomery-Asberg Depression Rating Scale (MADRS) in adults with treatment-resistant depression treated with esketamine nasal spray combined with a new oral antidepressant. CNS. Drugs 36, 649–658.
Park, L.T., Luckenbaugh, D.A., Pennybaker, S.J., et al., 2020. The effects of ketamine on typical and atypical depressive symptoms. Acta Psychiatr. Scand. 142, 394–401.
Chen, M.H., Wu, H.J., Li, C.T., et al., 2021. Low-dose ketamine infusion for treating subjective cognitive, somatic, and affective depression symptoms of treatment-resistant depression. Asian J. Psychiatr. 66.
Rodrigues, N.B., McIntyre, R.S., Lipsitz, O., et al., 2020. Changes in symptoms of anhedonia in adults with major depressive or bipolar disorder receiving IV ketamine: results from the Canadian rapid treatment center of excellence. J. Affect. Disord. 276, 570–575.
Artin, H., Bentley, S., Mehaffey, E., et al., 2022. Effects of intranasal (S)-ketamine on Veterans with co-morbid treatment-resistant depression and PTSD: a retrospective case series. EClinicalMedicine 48, 101439.
Bentley, S., Artin, H., Mehaffey, E., et al., 2022. Response to intravenous racemic ketamine after switch from intranasal (S)-ketamine on symptoms of treatment-resistant depression and post-traumatic stress disorder in Veterans: a retrospective case series. Pharmacotherapy J. Hum. Pharmacol. Drug Therapy 42, 272–279.
Ricca, B.P., Blaine, B.E., 2022. Brief research report: notes on a nonparametric estimate of effect size. J. Exp. Educ. 90, 249–258.
Priest, R.G., Hawley, C.J., Kibel, D., et al., 1996. Recovery from depressive illness does fit an exponential model. J. Clin. Psychopharmacol. 16, 420–424.
Berlow, Y., Zandvakili, A., Price, L., et al., 2021. Modeling treatment response to transcranial magnetic stimulation using an exponential decay function. Biol. Psychiatry 89, S195–S196.
Jesus-Nunes, A.P., Leal, G.C., Correia-Melo, F.S., et al., 2022. Clinical predictors of depressive symptom remission and response after racemic ketamine and esketamine infusion in treatment-resistant depression. Hum. Psychopharmacol. Clin. Exp. 37, e2836.
Khan, A., Leventhal, R.M., Khan, S.R., et al., 2002. Severity of depression and response to antidepressants and placebo: an analysis of the food and drug administration database. J. Clin. Psychopharmacol. 22, 40–45.
Kirsch, I., Deacon, B.J., Huedo-Medina, T.B., et al., 2008. Initial severity and antidepressant benefits: a meta-analysis of data submitted to the food and drug administration. PLoS Med 5, e45.
Fournier, J.C., DeRubeis, R.J., Hollon, S.D., et al., 2010. Antidepressant drug effects and depression severity: a patient-level meta-analysis. JAMA 303, 47–53.
Su, T.P., Chen, M.H., Li, C.T., et al., 2017. Dose-related effects of adjunctive ketamine in taiwanese patients with treatment-resistant depression. Neuropsychopharmacology 42, 2482–2492.
Joneborg, I., Lee, Y., Di Vincenzo, J.D., et al., 2022. Active mechanisms of ketamine-assisted psychotherapy: a systematic review. J. Affect. Disord. 315, 105–112.