## Predicting non-response to ketamine for depression: a symptom-level analysis of real-world data

 $\frac{1}{2}$ Eric A. Miller MD<sup>1,2</sup>, Houtan Totonchi Afshar MD<sup>1,2</sup>, Jyoti Mishra PhD<sup>2</sup>, Dhakshin Ramanathan<br>MD. PhD<sup>1,2,3</sup> , Dhakshin Ramanathan

- MD, PhD<sup>4,2,3</sup><br>1. Department of Mental Health, VA San Diego Medical Center, San Diego, CA 92161, USA<br>2. Deventment of Burkistan, US San Diego, Ja Jalla 64,83993, USA
	- 2. Department of Psychiatry, UC San Diego, La Jolla, CA 92093, USA
	- 3. Center of Excellence for Stress and Mental Health, VA San Diego Medical Center.

# ו<br>ו

## Abstract<br>Background

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. for predicting which patients will, or will not, respond to treatment are lacking.

## **Methods**

This is a retrospective analysis of PHQ-9 item response data from 120 military veterans with TRD This is a retrospective analysis of PHQ-9 is a report of PHQ-9 in the response of PHQ-120 in TRD-0 in<br>who received repeated doses of intravenous racemic ketamine or intranasal eskatamine in a<br>real-world clinic. Regression real-world clinic. Regression models were fit to individual patients' symptom trajectories and<br>model parameters were analyzed to characterize how different symptoms responded to model parameters were analyzed to characterize how different symptoms responded to<br>treatment. Logistic regression classifiers were used to predict treatment response using treatment. Logistic regression classifiers were used to predict treatment response using<br>patients' baseline depression symptoms alone. Finally, by parametrically adjusting the cla patients' baseline depression symptoms alone. Finally, by parametrically adjusting the cl<br>decision thresholds, the full space of models was searched to identify the best models fo decision thresholds, the full space of models was searched to identify the best models for predicting non-response with very high negative predictive value. prediction non-response with very high negative value.<br>Results

|
|
| Results<br>Model slopes indicated progressive improvement on all nine symptoms, but the symptom of depressed mood improved faster than the symptom of low energy. The first principal component (PC) represented a data-driven measurement of overall treatment response, while<br>the second PC divided the symptoms into affective and somatic subdomains. Logistic regression component (PC) represents a data-driven measurement of overall treatment response) and<br>the second PC divided the symptoms into affective and somatic subdomains. Logistic regressior<br>classifiers predicted response better tha the second PC divided the symptoms into an equivale and somatic cheminical segments of the symptoms.<br>Classifiers predicted response better than chance using baseline symptoms, but these models<br>achieved only 60.2% predictiv achieved only 60.2% predictive value. Using threshold tuning, we identified models that can<br>predict non-response with a negative predictive value of 96.4%, while retaining a specificity of achieved only concrete predictive value. Using threshold tuning, we herecan inverse that can<br>predict non-response with a negative predictive value of 96.4%, while retaining a specificity of<br>22.1%, suggesting we could succe predict 2.1%, suggesting we could successfully identify 22% of individuals who would not respond<br>purely based on baseline symptom scores. purely based on baseline symptom scores.<br>Conclusions

# purely based on baseline symptom scores.

 $\frac{1}{\sqrt{2}}$ **Conclusions**<br>We develope<br>respond to k respond to ketamine. This could inform rational treatment recommendations to avoid<br>additional treatment failures. additional treatment failures.

It is made available under a CC-BY-ND 4.0 International license. medRxiv preprint doi: [https://doi.org/10.1101/2023.07.03.23292094;](https://doi.org/10.1101/2023.07.03.23292094) this version posted July 5, 2023. The copyright holder for this preprint<br>(which was not certified by peer review) is the author/funder, who has granted med

Introduction<br>Despite effective treatments for depression, such as psychotherapy [1] and monoamine-based antidepressants [2], many patients do not find remission even after multiple trials of different<br>treatments [3]. With a novel mechanism of action, ketamine offers hope for such patients with treatments [3]. With a novel mechanism of action, ketamine offers hope for such patients with<br>treatment resistant depression (TRD) [4]. However, our ability to predict which patients will treatment resistant depression (TRD) [4]. However, our ability to predict which patients will<br>respond to ketamine remains limited. Although studies have identified potential moderators of treatment resistant depression (TRD) [4]. However, our ability to predict which patients will<br>response, such as obesity [5,6], family history of alcohol use disorder [5,7], and concomitant response, such as obesity [5,6], family history of alcohol use disorder [5,7], and concomitant<br>benzodiazepine use [8], a recent large meta-analysis failed to detect any consistent patient-leve responsion in the such as one of the superior of the sense is to all the number of all the nenzo diazepine use<br>predictors of response to ketamine [9]. In the absence of consistent predictors of response, it predictors of response to ketamine [9]. In the absence of consistent predictors of response, it<br>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) [10], transcranial remains different to strain, and select the optimal treatment for a particular pattent among a<br>growing list of options for TRD, including electroconvulsive therapy (ECT) [10], transcranial<br>magnetic stimulation (TMS) [11], growing list of premise of one manipulations for the computer therapy (CDT) [12], deep brain stimulations (DBS) [13], and psychedelic-assisted psychotherapy [14]. (DBS)  $[13]$ , and psychedelic-assisted psychotherapy  $[14]$ .

Prediction fundamentally rests on a choice of how to measure changes in the outcome of  $\frac{1}{i}$ interest, in this case changes in depression. Most clinical studies utilize questionnaire sum<br>scores to quantify depression before and after treatment, with response defined via a scores to quantify depression before and after treatment, with response defined via a<br>proportional drop in the sum score. Even for validated questionnaires, sum scores can mas proportional drop in the sum score. Even for validated questionnaires, sum scores can<br>individual differences or dynamics at the level of symptoms [15,16] or dimensional fea individual differences or dynamics at the level of symptoms [15,16] or dimensional features of<br>illness [17]. For example, different symptom clusters may have distinct etiological and illness [17]. For example, different symptom clusters may have distinct etiological and<br>physiological underpinnings, which may, in turn, respond differently to treatments [18,19]. physiological underpinnings, which may, in turn, respond differently to treatments [18<br>Although many studies have reported on ketamine's effects specifically on suicidality [ physiological underpinnings, and they, in turn, respond and creative to treatments [18,19].<br>Although many studies have reported on ketamine's effects specifically on suicidality [20], v<br>are aware of only a few studies that Although many studies have reported on heliances showing positions, showinging are aware of only a few studies that have compared how ketamine affects different depressior<br>symptoms or symptom clusters[21–23]. None of those symptoms or symptom clusters[21–23]. None of those studies have modeled trajectories of symptoms over time. symptoms over time.<br>Symptoms over time.

Our first goal, therefore, was to model how individual symptoms of depression change over<br>time for patients undergoing repeated ketamine treatments. This is a secondary analysis of real-(<br>t 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<br>depression. As such, these data reflect how ketamine may be expected to perform for the psychiatry.<br>depression. As such, these data reflect how ketamine may be expected to perform for the<br>treatment of TRD among complex patients, the very patients who could benefit most from depression. As such, these data reflect in the summering to suppressive performance to<br>treatment of TRD among complex patients, the very patients who could benefit most from<br>novel and effective treatments. In smaller studi novel and effective treatments. In smaller studies from the same population, our group has<br>found that ketamine is effective for these patients, but with lower response rates than found that ketamine is effective for these patients, but with lower response rates than<br>commonly reported in clinical trials [24,25]. This further highlights the need for reliable found that he comments is effective for these patients, but with lower response than<br>commonly reported in clinical trials [24,25]. This further highlights the need for reliable<br>predictors of response or, perhaps more impor predictors of response or, perhaps more importantly, predictors of non-response to ketamine<br>among complex patients. predictors of response or, perhaps more important more important more important more to the non-response to ke<br>among complex patients.

As a secondary goal, building on our models of symptom trajectories, we used machine learning  $\frac{1}{2}$ classifiers to predict whether patients would respond to ketamine using their baseline item by<br>item PHQ-9 symptom scores. We also developed a model that can predict non-response, i.e. classifiers to predict whether patients would respond to welching their baseline item by the<br>item PHQ-9 symptom scores. We also developed a model that can predict non-response, i.e.<br>treatment failure, with very high confid treatment failure, with very high confidence for a meaningful subset of the patients. These<br>findings contribute to our understanding of how ketamine affects individual symptoms of treatment failure, with very mga commence for a meaningful classe for the patients. These<br>findings contribute to our understanding of how ketamine affects individual symptoms of<br>depression for complex patients in a real-wo findings contribute to our understanding of how ketamine affects individual symptoms of<br>depression for complex patients in a real-world clinical setting. Our method for identifying depression for complex patients in a real-world clinical setting. Our method for identifying

patients who are unlikely to respond to ketamine could prove useful, if replicated, for guiding<br>treatment recommendations among a growing list of interventions targeting TRD. treatment recommendations and interventions and interventions targeting  $\frac{1}{2}$  of  $\frac{1}{2}$  of  $\frac{1}{2}$ 

### **Methods**

### Patients

Data were obtained from 120 patients who underwent serial ketamine induction treatments for depression at the San Diego Veterans' Administration hospital between January 2020 and June<br>2022. 85 patients were male and 35 were female. Ages ranged from 26 to 75 (mean 45, depression at the San Diego Version 2022. And the San Diego 2022. 85 patients were 2022. 85 patients were male<br>Administration 12) years. Most patients (92%) carried a diagnosis of major depression, but 2022. 86 patients were male and 35 were female. Ages ranged from 26 to 76 (mean 45)<br>standard deviation 12) years. Most patients (92%) carried a diagnosis of major depressic<br>comorbidity was very common. The most frequently standard deviation 12) years interspectents (92%) carried a diagnosis of major depression, and<br>comorbidity was very common. The most frequently co-occurring condition was PTSD in 73% o<br>patients, with less frequent conditio comorbidity, was very common, the most frequently co-occurring common frequently co-occurring<br>patients, with less frequent conditions including generalized anxiety disorder, bipolar disorder,<br>ADHD, borderline personality d patients, with less frequent conditions including generalized annual, accreasing the disorders.<br>ADHD, borderline personality disorder, and various substance use disorders. This study was<br>approved as an institutional review ADHD, borderline personality and various substance use and various various, was<br>approved as an institutional review board (IRB) exemption by the local VA institutional revie<br>board (IRB 1223219). approved as an institutional review board (IRB) exemption by the local VA institution by the local VA institu board (IRB 1223219).<br>*Treatments* 

 $\frac{1}{\sqrt{2}}$ 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<br>formulation of ketamine for all sessions. Intranasal esketamine doses typically started at 56 mg formulation of ketamine for all sessions. Intranasal esketamine doses typically started at 56 mg<br>and were increased to 84 mg after the initial session. Intravenous doses had more variability, as and were increased to 84 mg after the initial session. Intravenous doses had more variability, as<br>they were dosed at 0.5 to 1 mg/kg, with doses adjusted based on tolerability, side effects and they were dosed at 0.5 to 1 mg/kg, with doses adjusted based on tolerability, side effects and<br>efficacy. Formal psychotherapy was not paired with ketamine sessions, though psychological they were doese at 2012 to 2 mg/kg, with doese adjusted at 2012 to 2014 may such 2012 they with<br>efficacy. Formal psychotherapy was not paired with ketamine sessions, though psychological<br>support was available if needed. Mo efficacy. Formal psychodically, the mode pairs with distance sestivity, analogic psychological<br>support was available if needed. Most patients (110 of 120) completed at least eight treatme<br>sessions, with the remaining compl sessions, with the remaining completing between 2 and 7 sessions each.

### Analysis of average PHQ-9 scores

 $\frac{1}{2}$ Patients completed PHQ-9 questionnaires at baseline and prior to ketamine treatment sessions. Patients compressed PHQ-9 questionnaires at baseline and prior to homining treatment sessions.<br>Supplemental Table 1 lists the full text for the nine items of the PHQ-9. Average PHQ-9 sum<br>scores across patients were compute Supplemental Table 2 lists the full text for the ministrict of the PHQ-9. And County Form<br>scores across patients were computed across the first eight sessions. Significant change<br>average PHQ-9 item responses were computed average PHQ-9 item responses were computed across the eight sessions. Significant change in<br>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 sum score was defined as p<br>the first two sessions' data violated normality (p < 0.05, Shapiro-Wilk test). Significant change<br>item response was defined as p < 0.006 (after Bonferroni correction for multiple comparisons) item response was defined as p < 0.006 (after Bonferroni correction for multiple comparisons)<br>on the Friedman test. Only patients who completed a PHQ-9 before all eight sessions (82 of 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.

### Analysis of item trajectories

 $\frac{1}{\sqrt{2}}$ Item responses were analyzed across treatment sessions for each item, i, and for each patient,  $p$ . Linear (Equation 1) and exponential (Equation 2) models were fit to item response<br>trajectories by minimizing the residual sum-squared errors (RSS). In the linear model,  $m$  is the p. Linear (Equation 1) and exponential (Equation 2) models were fit to item response<br>trajectories by minimizing the residual sum-squared errors (RSS). In the linear model,<br>linear slope, t is time in days, and b is the inte trajectories by minimizing the residual sum-squared errors (RSS). In the linear model, *nt* is the<br>linear slope, *t* is time in days, and *b* is the intercept. In the exponential model, *a* is a scaling<br>factor, *m* is a g liftear slope,  $\iota$  is time in days, and *b* is the intercept. In the exponential model,  $\alpha$  is a scaling<br>factor,  $m$  is a growth factor,  $t$  is time in days, and  $b$  is a constant offset. In both models,  $m$ <br>reflects a reflects a rate of change in item response. is time in days, and  $\nu$  is a constant offset. In both models,  $m$  em response. reflects a rate of change in item response.

It is made available under a CC-BY-ND 4.0 International license. medRxiv preprint doi: [https://doi.org/10.1101/2023.07.03.23292094;](https://doi.org/10.1101/2023.07.03.23292094) this version posted July 5, 2023. The copyright holder for this preprint<br>(which was not certified by peer review) is the author/funder, who has granted med



 $y_{i,p}(t) = a_{i,p} e^{t}$  $^{-m_{i,p}t} + b_{i,p}$  (2)<br>ke information criterion (AIC) and Bayesian information criterion<br>urposes of model selection. Specifically, AIC and BIC were com For each model, the Akaike information criterion (AIC) and Bay Journal Controllering (BIC)<br>were computed for the purposes of model selection. Specifically, AIC and BIC were computed<br>using the RSS from model fitting, assumi using the RSS from model fitting, assuming independent and identically distributed (IID)<br>residuals following a normal distribution. For each patient, the winning model was defined as 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<br>model, as the best overall, was used as the basis of all subsequent analyses. Differences in linear given questionnaire was defined as the basis of all subsequent analyses. Differences in line<br>model, as the best overall, was used as the basis of all subsequent analyses. Differences in line<br>slopes between the items was de slopes between the items was defined as p < 0.05 on the Friedman test for repeated measures,<br>with post-hoc evaluation for pairwise differences via Wilcoxon signed-rank test with Bonferroni 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  $[26]$ .

## Principal component analysis for linear slopes

 $\frac{1}{3}$ Using linear models (Equation 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-<br>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 dimensional space of suppose of standarding the set of standarding committed computers in the separation<br>maximal variance across participants. Parallel analysis was used to evaluate the strength of<br>principal components (PC maximal variance across participants. Parallel analysis was used to evaluate the strength or<br>principal components (PCs), in which the actual eigenvalue of each PC was compared with t<br>95<sup>th</sup> percentile of the distribution o 95<sup>th</sup> 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.  $\mathcal{C}$  change. Coefficients for each significant component were analyzed to understand directions of  $\mathcal{C}$ 

## Predicting treatment response

 $\frac{1}{1}$ 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<br>rate of change in symptoms. Due to data normalization, a negative sign implies a greater rate of change in symptoms. Due to data normalization, a negative sign implies a greater<br>improvement along PC1 than the mean across participants. To relate this to classical measures improvement along PC1 than the mean across participants. To relate this to classical mea<br>of response, percent changes in PHQ-9 sum scores were also evaluated. Logistic regressic of response, percent changes in PHQ-9 sum scores were also evaluated. Logistic regression was<br>used to predict treatment response using patients' baseline PHQ-9 item responses. An of response, percent changes in PHQ-9 sum scores in PHQ-9 sum scopenses. An<br>used to predict treatment response using patients' baseline PHQ-9 item responses. An<br>exhaustive feature selection was conducted to evaluate all 51 exhaustive feature selection was conducted to evaluate all 511 possible subsets of PHQ<br>as features. For each model, classification performance was evaluated across 1000 itera as features. For each model, classification performance was evaluated across 1000 iterations of repeated 5-fold cross validation.

## Threshold tuning for high confidence predictions

 $\frac{1}{1}$ 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<br>sensitivity, or between negative predictive value (NPV) and specificity. For all 511 possible sensitivity, or between negative predictive value (NPV) and specificity. For all S11 possibl<br>subsets of features, a parametric search was conducted across 20 different classification subsets of features, a parametric search was conducted across 20 different classification<br>thresholds ranging from 0 to 1. This resulted in over 10,000 distinct classification models, v thresholds ranging  $\overline{\acute{s}}$ rom 0 to 1. This resulted in over 10,000 distinct classification models<br>then underwent cross validation as before. To optimize for potential clinical utility, mode then underwent cross validation as before. To optimize for potential clinical utility, models were  $t$  then underwent cross validation as  $\mathcal{L}_{\mathbf{p}}$  as a before. To optimize  $\mathcal{L}_{\mathbf{p}}$  models were  $\mathcal{L}_{\mathbf{p}}$ 

sensitivity for predicting "response", or (2) at least 90% NPV and at least 10% specificity for<br>predicting "non-response". These criteria were chosen because very high predictive value coul predicting "non-response". These criteria were chosen because very high predictive value contending the sectio<br>provide clinically actionable information, even if only for a subset of the patients. In contras 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<br>have higher sensitivity. Among the models meeting these specifications, performance was models with lower predictive value are unlikely to change clinical management, correlating,<br>have higher sensitivity. Among the models meeting these specifications, performance was<br>ranked by the arithmetic mean of predictiv have higher sensitivity, changing the models meeting meeting performance, performance was<br>ranked by the arithmetic mean of predictive value (PPV or NPV) and coverage of the releva<br>cases (sensitivity or specificity), with t cases (sensitivity or specificity), with the best model defined as the one with highest mean of those two characteristics. cases (sensitivity or specificity), with the best model defined as the one with highest mean of<br>those two characteristics.

## Code and data availability

 $\frac{1}{2}$ 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/ketamine Prediction. Data are available upon request.<br>Results github.com/angevine/industrialsetamine/industrialsetamine<br>Results

 $\frac{1}{1}$ This 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-<br>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 clinical setting. We first confirmed that average PHQ-9 sum scores improved across the<br>ketamine treatment sessions (Figure 1a) (p <  $10^{-18}$ , Friedman test). In addition, average ketamine treatment sessions (Figure 1a) (p <  $10^{-18}$ , Friedman test). In addition, average<br>responses for every individual PHQ-9 item improved over the course of treatment (Figu ketamine treatment sessions (Figure 1a) (p < 10<sup>–10</sup>, Friedman test). In addition, average<br>responses for every individual PHQ-9 item improved over the course of treatment (Figure<br>(ps < 10<sup>–5</sup>, Friedman test). We then proce (ps <  $10^{-5}$ , Friedman test). We then proceeded to analyze symptom trajectories for each individual patient. (ps < 10<sup>-5</sup>, Friedman test). We then proceeded to analyze symptom trajectories for each<br>individual patient. individual patient.<br>Modeling individual symptom trajectories across individuals

 $\frac{1}{\sqrt{2}}$ 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 (Equation 1), we<br>also tested an exponential model (Equation 2), motivated by prior work showing that treatment in the majority of individuals. In addition of a simple linear model (equation 2), i.e.<br>also tested an exponential model (Equation 2), motivated by prior work showing that<br>exponential functions can describe vario exponential functions can describe various interventions for depression [27,28]. We fi<br>and exponential models to item response trajectories for each patient (Figure 1c), and exponential functions can describe various interventions for depression [27,28]. We fit linear<br>and exponential models to item response trajectories for each patient (Figure 1c), and then<br>evaluated the fit of both models us evaluated the fit of both models using AIC and BIC. Linear models were better than exponer<br>models for the majority of patients (104 of 120 for AIC, 105 of 120 for BIC) (Figure 1d), so we models for the majority of patients (104 of 120 for AIC, 105 of 120 for BIC) (Figure 1d), so we<br>chose the linear model as the basis for further analysis to maintain consistency/comparability chose the linear model as the basis for further analysis to maintain consistency/comparability<br>across individuals. chose the linear model as the basis for further analysis to maintain comparability  $\mathcal{L}^{\prime}$ across individuals.<br>The slopes of the linear models describe the rate of change in each symptom of depression for

ך<br>ג 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 .<br>negative for all nine items, confirming that patients generally improved across all symptoms<br>(Figure 1e). We detected a difference in slopes between the items (p = 0.012, Friedman test). negative for all nine trens, communing that patients generally improved across and principle that<br>(Figure 1e). We detected a difference in slopes between the items (p = 0.012, Friedman test)<br>Post-hoc comparisons revealed t (Figure 22, Figure 22, Figure 2011).<br>Post-hoc comparisons revealed that PHQ-9 item #2 (depressed mood) had significantly steepe<br>slopes than item #4 (tiredness) after Bonferroni correction (p = 0.001, Wilcoxon signed-rank Post-hoc comparisons reveals materials remains (approces mosta) managemental, everywisted<br>slopes than item #4 (tiredness) after Bonferroni correction (p = 0.001, Wilcoxon signed-rank<br>test) (Figure 1e). Thus, symptoms of de test) (Figure 1e). Thus, symptoms of depressed mood improved more rapidly than tiredness  $t_{\rm F}$ , (Figure 1e). Thus, symptoms of depressed more rapidly than times,  $\mu$ 

across ketamine treatment, with a difference in median  $\frac{d}{dt}$  of the interpretational range  $\frac{d}{dt}$ (see Methods).<br>Low Dimensional Variables to Capture Trajectory Changes

 $\frac{l}{\mathsf{I}}$ Next, we wondered whether we could parsimoniously capture how patients varied in the highextractional space of cymptom inspected in the material and the computed a proceptious<br>components analysis (PCA) on the trajectory slopes calculated for each item of the PHQ-9 acros<br>subjects. We detected two significant pr subjects. We detected two significant principal components (PCs), which explained 52.2% and<br>16.2% of the variance, respectively (significance calculated using parallel analysis, see Methods) subjects. We detected the significant principal components (PCs), which explained 52.2% of the variance, respectively (significance calculated using parallel analysis, see Method:<br>(Figure 2a). Analysis of coefficients reve (Figure 2a). Analysis of coefficients revealed that the first PC (PC1) described a weighted rate of<br>change for all nine symptoms in the same direction (Figure 2b). Coefficients of the second PC (Figure 2a). Analysis of commension that the first PCC (PC2) described a magnetic revenues<br>change for all nine symptoms in the same direction (Figure 2b). Coefficients of the second PC<br>(PC2) showed opposite signs for somat change for all nine symptoms in the same all substitutes and conduction the second of (PC2) showed opposite signs for somatic symptoms (e.g. appetite, energy, concentration,<br>movement) and affective symptoms (e.g. mood, anh (PC2) showed opposite signs for somatic symptoms (e.g. e.g. e.g. appendix), concentration,<br>movement) and affective symptoms (e.g. mood, anhedonia, thoughts of self-harm or of b<br>failure) (Figure 2c). Projecting participants failure) (Figure 2c). Projecting participants' slopes data onto these two PCs revealed the<br>distribution of patients across these two dimensions of symptom variance (Figure 2d). distribution of patients across these two dimensions of symptom variance (Figure 2d).

Based on the coefficient weights, the sign of the first principal component (PC1) represented  $\frac{1}{\sqrt{2}}$ whether a patient improved more (negative sign) or less (positive sign) than the mean rate of change in symptoms (Figure 2d). This is because each patient's symptom slopes were change in symptoms (Figure 2d). This is because each patient's symptom slopes were<br>standardized by the population mean slopes prior to computing PCA. Therefore, this offers a chandardized by the population mean slopes prior to computing PCA. Therefore, this continuity standard in stand<br>Simple, data-driven way of differentiating treatment responders from non-responders simple, data-driven way of differentiating treatment responders from non-responders. To better<br>illustrate this, we first plotted average PHQ-9 sum scores for patients with negative PC1 simple, data-driven way of anti-comming treatment responders from non-respondent responders.<br>illustrate this, we first plotted average PHQ-9 sum scores for patients with negative PC1<br>(identified as responders, n=62) compar in the this, we made proceed and they are the concerning matrix forming the red (identified<br>identified as responders, n=52) compared to patients with non-negative PC1 (identified<br>non-responders, n=58) (Figure 2e). The sum  $\sim$ non-responders, n=58) (Figure 2e). The sum scores among responders improved significantly<br>between ketamine induction sessions (F(7,336) = 46.7, p <  $10^{-45}$ , partial eta-squared = 0.49, non-responders, n=59, (Figure 20). The sum scores among responsers improved significantly<br>between ketamine induction sessions (F(7,336) = 46.7, p <  $10^{-45}$ , partial eta-squared = 0.49,<br>repeated measures ANOVA). The mean between ketamine induction sessions (F(7,336) = 46.7, p < 10<sup>-45</sup>, partial eta-squared = 0.49,<br>repeated measures ANOVA). The mean change in sum scores between baseline and the final<br>session among responders was -8.7 points session among responders was -8.7 points (95% CI: -9.8 to -7.6). In contrast, we detected no<br>differences in sum scores between sessions among the non-responders (F(7,224) = 2.07, p = 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(7,224) = 2.07, p<br>0.07, Greenhouse-Geisser correction, repeated measures ANOVA). The mean change in sum<br>scores among non-responders was -0.40 (95% CI: 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 l<br>t that the sign of the  $2^{nd}$  PC (PC2) could reflect whether there were relatively greater improvements in the affective symptom clusters (negative sign) or somatic symptom clusters<br>(non-negative sign). To test this, we calculated (for responders only) the trajectories of average improvements in the affective symptom clusters (in gative sign) or somatic symptom clusters<br>(non-negative sign). To test this, we calculated (for responders only) the trajectories of averag<br>scores for the affective and som (non-scores for the affective and somatic subdomains of symptoms, grouped using the PC2 weights<br>noted above. As expected, responders with a non-negative PC2 showed greater improvements noted above. As expected, responders with a non-negative PC2 showed greater improvements<br>in somatic symptoms, compared to responders with a negative PC2 (F(7, 329) = 3.03, p = 0.004 in somatic symptoms, compared to responders with a negative PC2 (F(7, 329) = 3.03, p = 0.004,<br>mixed ANOVA interaction) (Supplemental Figure 1a). In contrast, we were unable to detect a mixed ANOVA interaction) (Supplemental Figure 1a). In contrast, we were unable to detect a<br>significant difference in the rate of improvement in affective symptoms based on PC2 sign (F(7, significant difference in the rate of improvement in affective symptoms based on PC2 sign (F(<br>329) = 1.01, p = 0.42, mixed ANOVA interaction) (Supplemental Figure 1b). Thus, the sign of P  $\overline{329}$ ) = 1.01, p = 0.42, mixed ANOVA interaction) (Supplemental Figure 1b). Thus, the sign of PC2 sign (FC2 sign (FC2 sign (FC2 sign (FC2 sign of PC2 sign (FC2 sign (FC2 sign (FC2 sign (FC2 sign (FC2 sign (FC2 sign ( 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. variance even within those who were categorized as responding overall.

 $\frac{1}{\sqrt{2}}$ Predicting response from baseline symptoms<br>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<br>score. To ensure that our categorization corresponds at least roughly with more standard of treatment response are defined with respect to the percent change in the overall ring open.<br>score. To ensure that our categorization corresponds at least roughly with more standard<br>definitions of treatment response, we score. To ensure that our categorization corresponds at least roughly manifered changes<br>definitions of treatment response, we computed the percent reduction in PHQ-9 sum sco<br>from baseline to the last treatment session for from baseline to the last treatment session for patients categorized by negative PC1<br>(responders) and non-negative PC1 (non-responders). Among responders, the median perce from an enterments in the last treatment session for patients categorized by negative respectively.<br>(responders) and non-negative PC1 (non-responders). Among responders, the media<br>change in PHQ-9 sum scores was a reduction (responders) and non-negative PC1 (non-responders). Among responders, the median percent<br>change in PHQ-9 sum scores was a reduction by 39% (IQR 25%). Among non-responders, the<br>median percent change in sum scores was 0% (IQ change in PHQ-9 sum scores was a reduction by 30% (IQR 22%) (Figure 2f), suggesting that in general<br>median percent change in sum scores was 0% (IQR 22%) (Figure 2f), suggesting that in general<br>this method of classifying su this method of classifying subjects was reasonable.

Thus, using this categorization, we next wanted to see if baseline symptoms might provide some  $\frac{1}{\epsilon}$ ability to classify responders and non-responders. Prior to classification, we first analyzed<br>whether PC1 scores were related to basic demographics (age, gender) and treatment type (IV ability to classify responsible and non-responsibility responsibility, its increasing constants whether PC1 scores were related to basic demographics (age, gender) and treatment type racemic ketamine vs. intranasal esketam racemic ketamine vs. intranasal esketamine). We detected no differences in average PC1 value<br>based on ketamine formulation (esketamine vs IV racemic) (p = 0.59, two-sided t-test), gender based on ketamine formulation (esketamine vs IV racemic) ( $p = 0.59$ , two-sided t-test), gender<br>( $p = 0.19$ , two-sided t-test) or age ( $r = 0.13$ ,  $p = 0.17$ , Pearson's correlation), suggesting these (p = 0.19, two-sided t-test) or age (r = 0.13, p = 0.17, Pearson's correlation), suggesting these<br>simple demographic factors would not be informative for group membership or classification (imple demographic factors would not be informative for group membership or classification<br>(data not shown). We next asked whether treatment response, as defined above using chang (data not shown). We next asked whether treatment response, as defined above using change<br>in PC1, could be predicted from single item PHQ-9 scores alone. Specifically, we used logistic regression to predict the sign of PC1 using baseline PHQ-9 items as features. To identify the best in PC<sub>1</sub>, could be predict the sign of PC1 using baseline PHQ-9 items as features. To identify the b<br>features for prediction, we conducted an exhaustive feature selection by fitting logistic regressions prediction, we conducted an exhaustive feature selection by fitting logistic<br>regressions with all 511 possible subsets of PHQ-9 items as features. For each of these 511 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. l<br>t though all were only moderately better than chance (Figure 2g). The model with best CV accuracy used only two PHQ-9 items as features: item #1 (anhedonia) and item #6 (feeling of<br>failure). Higher baseline responses on either item predicted subsequent treatment response. failure). Higher baseline responses on either item predicted subsequent treatment response.<br>For this model, the average classification accuracy to holdout data was 60.3% (95% confidenc failure). Higher baseline responses on either item predicted subsequent is banken reponse.<br>For this model, the average classification accuracy to holdout data was 60.3% (95% confidence.<br>Interval 59.5 – 61.1%) (Figure 2h). interval 59.5 – 61.1%) (Figure 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 and F1 score was 64.0%. The second-best model used only one item as a feature, item #2<br>(depressed mood), with nearly as good performance as the best model: 60.0% average accuracy and F1 score was 6 recent model used a feature was finite and the second-best model used only depressed mood),<br>(95% CI 59.2% - 60.8%), 60.0% precision, 67.7% recall, and 63.6% F1 score. Patients with h (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.

## Threshold tuning for high confidence predictions

 $\frac{1}{\sqrt{2}}$ 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<br>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 a manumment of around 60% predictive ranks) matrice annually to provide canceling predictive<br>confidence for changing clinical management. We reasoned that a more clinically useful mod<br>would have a very high predictive valu confidence for changing clinical management. We reasoned that a more clinical, a securities is 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<br>predict with over 90% NPV that a patient will not respond to ketamine, then we could be guite 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 ؟<br>ا predictive confidence (Figure 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<br>between 0 to 1 (see Methods). Changing the threshold in this way necessarily results in a between 0 to 1 (see Methods). Changing the threshold in this way necessarily results in a<br>tradeoff between prediction confidence (PPV or NPV) and coverage of the relevant cases between 2 to 2 (see Methods). Changing the threshold in this way necessarily result in a<br>tradeoff between prediction confidence (PPV or NPV) and coverage of the relevant cases<br>(sensitivity or specificity). For model select (sensitivity or specificity). For model selection, we conducted another exhaustive search across<br>all possible subsets of baseline PHQ-9 items, and for each of these feature sets, we applied (sensitivity or specificity). For model selection, we conducted and the summative search across<br>all possible subsets of baseline PHQ-9 items, and for each of these feature sets, we applied<br>threshold tuning to evaluate a ra all possible subsets of all subsetimations of these for the subsets of these feature sets, we applied<br>threshold tuning to evaluate a range of thresholds (Figure 3b). This resulted in over 10,000<br>distinct models that were c distinct models that were cross validated to evaluate model performance on holdout data<br>(Figure 3c). distinct models that were cross values of the cross values of

We then selected only those models with very high predictive value, while at the same time retaining a minimum coverage of the relevant cases (Figure 3b, upper right quadrant). ۱<br>r retaining a minimum coverage of the relevant cases (Figure 3b, upper right quadrant).<br>Specifically, we searched for models with either (1) at least 90% PPV and at least 10% sensiti Specifically, we searched for models with either (1) at least 90% PPV and at least 10% s<br>or (2) at least 90% NPV and at least 10% specificity. We identified hundreds of models or (2) at least 90% NPV and at least 10% specificity. We identified hundreds of models meeting<br>these performance criteria. The best of these models (see Methods) had a NPV of 96.4%, while these performance criteria. The best of these models (see Methods) had a NPV of 96.4%, while<br>retaining a specificity of 22.1%. This model used only three baseline PHQ-9 items as features: retaining a specificity of 22.1%. This model used only three baseline PHQ-9 items as features:<br>item #2 (depressed mood), item #5 (changes in appetite or eating), and item #9 (self-harm or retaining a specificity of 22.21% this model used only three baseline ring of terms as features:<br>item #2 (depressed mood), item #5 (changes in appetite or eating), and item #9 (self-harm or<br>suicidal ideation). Relatively s item an proposed mood), item as permayer moop rate or eating), and item as permanent or suited suited suited s<br>suicidal ideation). Relatively speaking, lower values on item #2 and item #5 and higher values<br>item #9 favored item #9 favored non-response (see Supplemental Methods).

Thus, using threshold tuning, we were able to identify classification models that predict non-ן<br>ו The tuning tuning, we continue to the control to identify computer the substituted of the symptom scores. In contrast, we detected no models for predicting treatment response at the specified performance cutoffs (Figure 3c specified performance cutoffs (Figure 3c, green).

# specified performance cutoffs (Figure 3c, green).

l<br>|
| Discussion<br>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 (Figure 1). We found a range of improved more course the course of actuality and the symptom of suppressed more as<br>improved more rapidly than the symptom of low energy (Figure 1). We found a range of<br>individual differences in these item response trajecto individual differences in these item response trajectories, both in the degree of change in<br>overall depression level, and in specific subdomains of symptoms (Figure 2). We develop overall depression level, and in specific subdomains of symptoms (Figure 2). We develope<br>logistic regression classifiers, which can predict better than chance whether patients will overall depression classifiers, which can predict better than chance whether patients will<br>legistic regression classifiers, which can predict better than chance whether patients will<br>respond to ketamine, using their baseli respond to ketamine, using their baseline symptoms alone (Figure 2). Finally, using threshold<br>tuning, we found classifiers that can identify a subset of patients who are highly *unlikely* to tuning, we found classifiers that can identify a subset of patients who are highly *unlikely* to<br>respond to ketamine with over 96% predictive value (Figure 3). Our findings shed light on hov tuning, we found classifiers that can identify a subset of patients who are highly unlikely to<br>respond to ketamine with over 96% predictive value (Figure 3). Our findings shed light on h<br>ketamine affects specific symptoms ketamine affects specific symptoms and dimensional features of depression, and the method for identifying non-responders may prove useful for informing rational treatment is an  $\epsilon$  information, prove useful for information  $\epsilon$ 

 $\begin{bmatrix} 11-14 \end{bmatrix}$ . [11–14].<br>Few prior studies have analyzed how different depression symptoms respond to ketamine,

i<br>F particularly across repeated doses. Floden and colleagues (2022) [21] conducted a symptom-<br>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, interestingly, not item #9 concerning suicidality [21]. They fovements in all PHQ-9 items except, interestingly, not item #9 concerning suicidality<br>They suggest that this could be due to low levels of baseline suicidality due to exclusion improvements in all PHQ-9 items energy, interestingly, increasing concerning suiteminy,  $\frac{1}{2}$ .<br>They suggest that this could be due to low levels of baseline suicidality due to exclusion of<br>patients at serious risk of s They suggest that the could be due to low levels of the suitable suitable, where suitable to<br>patients at serious risk of suicide from the TRANSFORM 2 trial. In contrast, our data show of<br>reductions in suicidality, highligh patients at serious risk of suite from the TRANSFORM 2 trial. In contrast, our data show it seen<br>reductions in suicidality, highlighting the value of studying outcomes in heterogeneous real-<br>world data with complex patient world data with complex patients.<br>We found that all depression symptoms improved across repeated ketamine sessions, but the

ا<br>s symptom of depressed mood improved faster than the symptom of low energy. This supports<br>results from Chen et al. (2021) [23], who found that a single infusion of low-dose IV ketamine symptom of depressed mood improved manying manying ying that the strengy. This support<br>results from Chen et al. (2021) [23], who found that a single infusion of low-dose IV ketamine<br>resulted in greater reductions in cognit resulted in greater reductions in cognitive and affective symptoms, compared to somatic<br>symptoms [23]. Similarly, Park and colleagues (2020) [22] found that typical/melancholic symptoms [23]. Similarly, Park and colleagues (2020) [22] found that typical/melancholic<br>symptoms improved more rapidly than atypical symptoms of depression after a single dose of IV symptoms improved more rapidly than atypical symptoms of depression after a single do<br>ketamine [22]. Furthermore, using principal component analysis (PCA), we found significa symptoms in provided more rapid more reading to the relation of the set of than at the found significant detai<br>Individual variation in whether patients improved more in affective symptoms (e.g. depressed individual variation in whether patients improved more in affective symptoms (e.g. depressed<br>mood, anhedonia, suicidal ideation) or in somatic symptoms (e.g. changes in appetite, individual variation in the more patients improved more in an extended propertie (e.g. mapper in the mood, anhe<br>doncentration, or energy level) suggesting that symptom response to ketamine is patient-<br>concentration, or ene mood, and the magnetic symptom is suited in the suited of the product of the position.<br>concentration, or energy level) suggesting that symptom response to ketamine is patie<br>specific. specific.<br>Using logistic regression classifiers, we were able to predict which patients would respond

l<br>k better than chance using their baseline symptoms alone. This was likely possible due to a<br>difference in baseline depression severity, such that responders had more severe baseline better than chance using their baseline symptoms alone. This was likely positive alone.<br>difference in baseline depression severity, such that responders had more severe baseline<br>depression than non-responders. These result depression than non-responders. These results may be counterintuitive, as one might expe<br>patients with more severe depression to be more treatment resistant. A study by Jesus-Nu patients with more severe depression to be more treatment resistant. A study by Jesus-Nunes et al. (2022) [29] found that patients with more severe depression were less likely to respond to<br>a single infusion of IV ketamine or esketamine. On the other hand, our data may reflect the et al. (2022) (2021) is an allowing patients with more severe depression were less linely to respond to a single<br>a single infusion of IV ketamine or esketamine. On the other hand, our data may reflect the<br>consistent findin consistent finding that antidepressants separate from placebo most prominently among<br>patients with severe baseline depression [30–32], which has also been observed for ketamir patients with severe baseline depression [30–32], which has also been observed for keta<br>[33]. It is unclear how much of that is due to stronger effects of the active drug or weake patients with severe baseline depression [30–32], which has also been observed for the lation<br>[33]. It is unclear how much of that is due to stronger effects of the active drug or weaker<br>effects of the placebo among patien effects of the placebo among patients with more severe depression. It is also possible that baseline depression severity may signal different underlying psychopathologies, with baseline depression severity may signal different underlying psychopathologies, with<br>consequent differences in response to ketamine, but future studies will be necessary to exp baseling depression severity, may signal different underlying psychopathologies, with<br>consequent differences in response to ketamine, but future studies will be necessary<br>that possibility. consequent differences in response to ketamine, but future studies will be necessary to explore

that possibility.<br>Importantly, the standard classifiers noted above are not likely to be useful in the clinic as they  $\frac{1}{3}$ can only predict response with about 60% positive predictive value. We reasoned that a much<br>higher predictive value would be reguired for a model to guide clinical decision making. We can only predictive value would be required for a model to guide clinical decision making. We<br>therefore optimized the classifiers for either high PPV for identifying responders, or for high herefore optimized the classifiers for either high PPV for identifying responders, or for high therefore optimized the classifiers for either high PPV for identifying responders, or for high  $t$  for either optimized the classifiers for either high  $P$  for  $\alpha$  is  $\beta$  for  $\beta$  for  $\beta$  for  $\beta$  for  $\beta$ 

NPV for identifying non-responders. We did this via threshold tuning, in which we searched over<br>the full space of decision thresholds, and then selected classifiers with optimal cross-validated<br>performance characteristics. performance characteristics. Using this approach, we found models that could predict non-<br>response to ketamine, i.e. treatment failure, with close to perfect negative predictive value. performance characteristics. Using this approach, we found models that could predict *non-*<br>response to ketamine, i.e. treatment failure, with close to perfect negative predictive value.<br>Although this comes at the expense response to ketamme, i.e. treatment failure, with close to perfect negative predictive value.<br>Although this comes at the expense of a reduced specificity, the model could prove useful<br>because it provides directly actionabl Although this comes at the expense of a reduced specificity, the model could prove useful<br>because it provides directly actionable information for those patients that it identifies as n<br>responders. Future pre-registered and these findings replicate and generalize. This simple approach of threshold tuning to find<br>clinically meaningful predictions may be useful for other treatments as well. clinically meaningful predictions may be useful for other treatments as well.<br>Limitations clinically meaningful predictions may be useful for other treatments as well.<br>Limitations

 $\frac{1}{1}$ 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-<br>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 ist control groups, making it impossible to distinguish the role of nonspecific factors including<br>placebo. For the same reason, we cannot exclude the role of regression to the mean, which is a list control groups, making it imposition to distinguite the role of nonspecific factors including<br>placebo. For the same reason, we cannot exclude the role of regression to the mean, which is a<br>possible explanation for obs possible explanation for observing improvement among patients with more severe baseline<br>depression. However, regression to the mean is unlikely to be the only driver of this finding, postum for the ware of the mean is unlikely to be the only driver of this finding,<br>since the average trajectories for responders and non-responders did not converge to a depression. However a serving since the average to a since the average to a since the average to a since the a<br>Common mean value. It is also important to highlight that psychotherapy was not provided in common mean value. It is also important to highlight that psychotherapy was not provic<br>concert with ketamine treatments, so our data do not speak to the potential for ketamir concert with ketamine treatments, so our data do not speak to the potential for ketamine-<br>assisted psychotherapy to prolonging or change its effects [34]. Finally, we lack detailed patient accounts of their treatment, which would be helpful for understanding patient-centered and assisted psychodicity, to prolonging or change in the completion, we have a cancel pattent.<br>accounts of their treatment, which would be helpful for understanding patient-centered and<br>functional outcomes beyond those limite functional outcomes beyond those limited data that are reflected on the PHQ-9 questionnaire<br> $\overline{\phantom{a}}$ functional outcomes beyond those limited data that are reflected on the PHQ-9 questionnaire.

(<br>F **Conclusion**<br>Repeated ketamine and esketamine led to progressive improvements in all symptoms of Repeated Repeated Ketamine and estetamine led to progressive improvements in an eying terms of<br>depression, but depressed mood improved faster than low energy. Using machine learnin<br>classifiers, we could predict non-respons depression, but depressed mood improved many of than low energy. Using machine rearning<br>classifiers, we could predict non-response to ketamine with very high confidence for a subset<br>the patients. If validated in future stu 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.<br>Conflicts of interest unlikely to be served by mightine, who might be better served by other treatments.<br>Conflicts of interest

# $\frac{1}{1}$

The authors declare no conflicts of interest.

## Acknowledgment

 $\frac{1}{2}$ D.R. is supported by funding from the Center of Excellence for Stress and Mental Health. Burroughs Wellcome Fund, Career Award for Medical Scientists. Burroughs Wellcome Fund, Career Award for Medical Scientists.

## ı |
|
|

## Figure legends<br>Figure 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. (b) Mean (SEM) PHQ-9 item scores across patients at each of the first eight ketamine sessions.

(c) One curves are the linear and exponential model fits, respectively. (See Methods for model<br>equations). (d) Bar plots comparing linear and exponential model fits using AIC (left) and BIC equations). (d) Bar plots comparing linear and exponential model fits using AIC (left) and BIC<br>(right), such that the bar heights indicate the number of patient-items where the linear (gre (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<br>each of the PHQ-9 items. P-value shown for pairwise comparison with significant difference or exponential (enerprocential was a better fit. (e) model (e) in exponential exponential patterns for<br>each of the PHQ-9 items. P-value shown for pairwise comparison with significant difference<br>after Bonferroni correction each of the PHQ-9 items. Protocome in the Phanner comparison managemental with significant differences<br>after Bonferroni correction for multiple comparisons. See Supplemental Table 1 for mapping<br>item names to PHQ-9 item tex item names to PHQ-9 item text.

## Figure 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<br>was determined by bootstrapping with parallel analysis (see Methods). (b) Coefficients of the from a PCA of the nine-dimensional data of linear slopes for each patient of guidenties for De<br>was determined by bootstrapping with parallel analysis (see Methods). (b) Coefficients of the<br>first PC of the data for each of first PC of the data for each of the features, corresponding to PHQ-9 item slopes. (c) Coefficier<br>of the second PC of the data for each of the features, corresponding to PHQ-9 item slopes. (d) 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<br>components of the data, colored according to the sign of the first principal component. components of the data, colored according to the sign of the first principal component.<br>Responders (green) are defined as having a negative sign of the first PC, while non-responders Responders (green) are defined as having a negative sign of the first PC, while non-responent (red) are defined as having a positive sign of the first PC. (e) Mean (SEM) PHQ-9 sum sco (red) are defined as having a positive sign of the first PC. (e) Mean (SEM) PHQ-9 sum scores<br>across patients for the first eight ketamine sessions for responders and non-responders. (f)  $\overline{\phantom{a}}$  across patients for the first eight ketamine sessions for responders and non-responders. (f)<br>Violin plots of relative change in PHQ-9 sum score from baseline to last session for responde 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<br>indicates distribution density. Dashed horizontal line indicates 0 change. (g) Histogram of the (green) and non-responders (red). Bars indicate meaning and change (g) Histogram of the<br>indicates distribution density. Dashed horizontal line indicates 0 change. (g) Histogram of the<br>mean cross-validation accuracy for all mean cross-validation accuracy for all models from an exhaustive feature selection over all<br>possible subsets of baseline PHQ-9 items as model features. (h) Histogram of accuracies acros mean cross-call accuracy for all models from an emancing from an experience for all<br>possible subsets of baseline PHQ-9 items as model features. (h) Histogram of accuracies acr<br>all iterations of cross-validation for the bes possible subsets of the subsetimently continue to be accuraciently in the subsets and set all iterations of cross-validation for the best model from the exhaustive feature selection, i.e.<br>the model with overall best mean C the model with overall best mean CV accuracy. Inset lists the other key model performance<br>characteristics for this model. For (b) and (c), see Supplemental Table 1 for mapping of item<br>names to PHQ-9 item text. characteristics for this model. For (b) and (c), see Supplemental Table 1 for mapping of item characteristics for this model. For (a) and (c), see Supplemental Table 1 for mapping of items<br>names to PHQ-9 item text.

## Figure 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<br>line indicates standard decision threshold for logistic regression, which maximizes accuracy. line indicates standard decision threshold for logistic regression, which maximizes accuracy.<br>Green and red dashed lines indicate alternate choices for the decision threshold, which line indicates standard decision in statistic regions regressivity and maximized decision.<br>Green and red dashed lines indicate alternate choices for the decision threshold, which<br>maximize positive predictive value (PPV) fo maximize positive predictive value (PPV) for predicting response or negative predictive value<br>(NPV) for predicting non-response, respectively. (b) Diagram depicting three major steps of (NPV) for predicting non-response, respectively. (b) Diagram depicting three major steps of<br>model selection procedure: feature search, threshold tuning, and cross validation. Grey boxe: (MPT) for predicting non-response, respectively. (b) Diagram depicting three major steps on<br>model selection procedure: feature search, threshold tuning, and cross validation. Grey box<br>represent the different sets of baseli represent the different sets of baseline PHQ-9 items that can be tried as a feature set. Threshold<br>tuning, as in (a), was conducted for each feature sets, resulting in over 10,000 models that then runing, as in (a), was conducted for each feature sets, resulting in over 10,000 models that then<br>underwent cross validation. (c) Scatter plot of cross validation model performance for some of tuning, as in (a), was conducted for each feature sets, resulting in over 20,000 model with the manner.<br>underwent cross validation. (c) Scatter plot of cross validation model performance for some of<br>the best models for pre underwent cross candidation. (c) Scatter plot of cross candidation model performance for some or<br>tradeoff between predictive value (PPV or NPV) and coverage of the relevant cases (sensitivity the best models for predicting response (green) or non-response (red). Plot show the industrial the interentie<br>tradeoff between predictive value (PPV or NPV) and coverage of the relevant cases (sensitivity<br>or specificity). or specificity). Green dots represent the best models, in terms of highest sensitivity (y-axis),  $\alpha$  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<br>models, in terms of highest specificity (y-axis), among all of the models with at least a minimum<br>NPV (x-axis). Dashed grey lines sh NPV (x-axis). Dashed grey lines show cutoffs for our predefined model performance<br>specifications: PPV or NPV > 90%, sensitivity or specificity > 10% (upper right quadrant). specifications: PPV or NPV > 90%, sensitivity or specificity > 10% (upper right quadra specifications: PPV or NPV or NPV or NPV or Specifications: 10% (upper right quadrant).<br>The 10% (upper right quadrant).

# |
|

- |
|
| References<br>
[1] Cuijpers P, Karyotaki E, de Wit L, et al. The effects of fifteen evidence-supported therapies<br>
for adult depression: A meta-analytic review. Psychotherapy Research 2020;30.<br>
[2] Cipriani A, Furukawa TA, Sal For Form Premish Cuite of the Cuivers Pressure Pressure Preparties in Form Cuivers of Fig. 3.<br>
for adult depression: A meta-analytic review. Psychotherapy Research 2020;30.<br>
[2] Cipriani A, Furukawa TA, Salanti G, et al. C
- for adult depression. The analytic review of processing process in 2020;<br>Cipriani A, Furukawa TA, Salanti G, et al. Comparative efficacy and acceptability<br>antidepressant drugs for the acute treatment of adults with major d antidepressant drugs for the acute treatment of adults with major depressive disorder: a<br>systematic review and network meta-analysis. The Lancet 2018;391. and specificate acute treatment of adults with major depressive disorder: a<br>systematic review and network meta-analysis. The Lancet 2018;391.<br>Rush AJ, Trivedi MH, Wisniewski SR, et al. Acute and longer-term outcomes in dep
- sush AJ, Trivedi MH, Wisniewski SR, et al. Acute and longer-term out<br>outpatients requiring one or several treatment steps: A STAR\*D repo [3] Rush AJ, Trivedi MH, Wisniewski SR, et al. Acute and longer-term outcomes in depressed<br>outpatients requiring one or several treatment steps: A STAR\*D report. American Journal<br>of Psychiatry 2006;163. outpatients requiring one or several treatment steps: A State or Sport. American Gammer<br>of Psychiatry 2006;163.<br>McIntyre RS, Rosenblat JD, Nemeroff CB, et al. Synthesizing the evidence for ketamine
- $\overline{\mathsf{M}}$ cintyre RS, Rosenblat.<br>and esketamine in treat. [4] McIntyre RS, Rosenblat JD, Nemeroff CB, et al. Synthesizing the evidence for ketamine<br>and esketamine in treatment-resistant depression: An international expert opinion on the<br>available evidence and implementation. Amer available evidence and implementation. American Journal of Psychiatry 2021;178.<br>Niciu MJ, Luckenbaugh DA, Ionescu DF, et al. Clinical predictors of ketamine response in
- Niciu MJ, Luckenbaugh DA, Ionescu DF, et al. Clinical predictors of ketamine respon<br>treatment-resistant major depression. Journal of Clinical Psychiatry 2014;75. treatment-resistant major depression. Journal of Clinical Psychiatry 2014;75.<br>[6] Freeman MP, Hock RS, Papakostas GI, et al. Body Mass Index as a Moderator of
- treatment-resembent-respects of procession. The commences pattent parameters of Clinical Psychiatry Creatment<br>Treatment Response to Ketamine for Major Depressive Disorder. J Clin Psycho [6] Freeman MP, Hock RS, Papakostas GI, et al. Body Mass Index as a Moderator of<br>Treatment Response to Ketamine for Major Depressive Disorder. J Clin Psychopharmacol<br>2020:40.
- Phelps LE, Brutsche N, Moral JR, et al. Family History of Alcohol Dependence and Initial ----,<br>Phelps LE<br>Antidepre Framily History Le, Brutsche N, Mortal H, Stateman, Morel, J. Franchet Dependence and Initial<br>Antidepressant Response to an N-methyl-D-aspartate Antagonist. Biol Psychiatry<br>2009;65. 2009;65.<br>Andrashko V. Novak T. Brunovsky M. et al. The Antidepressant Effect of Ketamine Is
- ----,---<br>Andrashk<br>Dampene [8] Andrashino V, Novak V, Scand City, NY 2001, The Anti-City, Antiopology Dampened by Concomitant Benzodiazepine Medication. Front Psychiatry 2020;11.<br>[9] Price RB, Kissel N, Baumeister A, et al. International pooled pati
- Price RB, Kissel N, Baumeister A, et al. International pooled patient-level meta-analysis of [9] Price RB, Kissel N, Baumeister A, et al. International pooled patient-level meta-analysis change in the stamine infusion for depression: In search of clinical moderators. Mol Psychiatry 2022;27:5096–112.
- Van Diermen L, Van Den Ameele S, Kamperman AM, et al. Prediction of electroconvulsive van Diermen L, Van<br>therapy response ar therapy response and remission in major depression: Meta-analysis. British Journal of Psychiatry 2018;212. therapy response and remission in major depression: Meta-analysis-ormation: Therapy<br>Psychiatry 2018;212.<br>Berlim MT, Van Den Eynde F, Daskalakis ZJ. A systematic review and meta-analysis on tl
- Psychiatry 2018;212. [11] Berlim MT, Van Den Eynde F, Daskalakis ZJ. A systematic review and meta-analysis on the<br>efficacy and acceptability of bilateral repetitive transcranial magnetic stimulation (rTMS)<br>for treating major depression. Psycho efficacy and acceptability of bilateral repetitive transcranial magnetic summation (rTMS)<br>for treating major depression. Psychol Med 2013;43.<br>Kayser S, Bewernick BH, Matusch A, et al. Magnetic seizure therapy in treatment-
- for treating major depression. Psychological States of New York<br>Kayser S, Bewernick BH, Matusch A, et al. Magnetic se<br>resistant depression: Clinical, neuropsychological and  $\lceil 12 \rceil$ resistant depression: Clinical, neuropsychological and metabolic effects. Psychol Med<br>2015;45.<br>Figee M, Riva-Posse P, Choi KS, et al. Deep Brain Stimulation for Depression.  $2015;45.$
- ----,<br>Figee M,<br>Neurothe  $\frac{1}{2}$  Figure M, Riva-Possel, Stimulation for Dep Brain Stimulation of Deptember. Neurotherapeutics 2022;<br>19.
- $[14]$ Goodwin GM, Aaronson ST, Alvarez O, et al. Single-Dose Psilocybin for a Treatment-2022;387:1637–48.<br>Borsboom D, Cramer AOJ. Network analysis: An integrative approach to th  $2022; 387: 1637 - 48.$
- Borsboom D, Crame<br>psychopathology. Ar [15] Borsboom D, Cramer AOJ. Network analysis: An integrative approach to the structure of<br>psychopathology. Annu Rev Clin Psychol 2013;9.<br>[16] Fried El, Nesse RM. Depression is not a consistent syndrome: An investigation o
- symptom patterns in the STAR $*$ D study. J Affect Disord 2015;172. end the EI, There EI, Nessen Repression is not a consistent symmetric support is an investigation is symptom patterns in the STAR\*D study. J Affect Disord 2015;172.<br>[17] Cuthbert BN, Insel TR. Toward the future of psychiat
- symptom patterns in the STAR<sub>1</sub>, control 2014, 2014, 2014.<br>Cuthbert BN, Insel TR. Toward the future of psychiatric diagnosis:<br>RDoC. BMC Med 2013;11. [17] Cuthbert BN, Insel TR. Toward the future of psychiatric diagnosis: The seven pillar<br>RDoC. BMC Med 2013;11.<br>[18] Chekroud AM, Gueorguieva R, Krumholz HM, et al. Reevaluating the efficacy and
- Med 2014 Med 2022<br>Chekroud AM, Gueorguiev<br>predictability of antidepres predictability of antidepressant treatments: A symptom clustering approach. JAN<br>Psychiatry 2017;74. Psychiatry 2017;74.<br>Fried El, Nesse RM. Depression sum-scores don't add up: why analyzing specific
- depression symptoms is essential. BMC Med 2015;13:72. depression symptoms is essential. BMC Med 2015;13:72.<br>[20] Jollant F, Colle R, Nguyen TML, et al. Ketamine and esketamine in suicidal though
- Jollant F, Colle R, Nguyen TML, et al. Ketamine and esketar<br>behaviors: a systematic review. Ther Adv Psychopharmacc For France, Colle R, Nguyen Time, Colle R, Nguyen Times and Encounter in Suite and September 2004<br>behaviors: a systematic review. Ther Adv Psychopharmacol 2023;13.<br>[21] Floden L, Hudgens S, Jamieson C, et al. Evaluation of
- Floden L, Hudgens S, Jamieson C, et al. Evaluation of Individual Items<br>Health Questionnaire (PHQ-9) and Montgomery-Asberg Depression F [21] Floden L, Hudgens S, Jamieson C, et al. Evaluation of Individual Items of the Patient<br>Health Questionnaire (PHQ-9) and Montgomery-Asberg Depression Rating Scale (MADRS)<br>in Adults with Treatment-Resistant Depression Tr in Adults with Treatment-Resistant Depression Treated with Esketamine Nasal Spray
- Park LT, Luckenbaugh DA, Pennybaker SJ, et al. The effects of ketamine on typical and atypical depressive symptoms. Acta Psychiatr Scand 2020;142. [22] Park LT, Luckenbaugh DA, Pennybaker SJ, et al. The effects of ketamine on typical<br>atypical depressive symptoms. Acta Psychiatr Scand 2020;142.<br>Chen MH, Wu HJ, Li CT, et al. Low-dose ketamine infusion for treating subj
- atypical depressive symptoms. Acta Psychiatr Scand 2020;<br>Chen MH, Wu HJ, Li CT, et al. Low-dose ketamine infusion for tr<br>cognitive, somatic, and affective depression symptoms of treat cognitive, somatic, and affective depression symptoms of treatment-resistant de<br>Asian J Psychiatr 2021;66. Asian J Psychiatr 2021;66.<br>Artin H, Bentley S, Mehaffey E, et al. Effects of intranasal (S)-ketamine on Veterans with
- Artin H, Bentley S, Mehaff<br>Artin H, Bentley S, Mehaff<br>co-morbid treatment-resis co-morbid treatment-resistant depression and PTSD: A retrospective case series.<br>EClinicalMedicine 2022;48. co-mortin treatment-resistant depression and PTSD: Procession and PTSD:<br>EClinicalMedicine 2022;48.<br>Bentley S, Artin H, Mehaffey E, et al. Response to intravenous racemic ketamine a
- Eclinical Medicine 2022;<br>Bentley S, Artin H, Mehaffey<br>switch from intranasal (S)-k [25] Bentley S, Artin H, Mehaffey E, et al. Response to intravenous racemic ketamine after<br>switch from intranasal (S)-ketamine on symptoms of treatment-resistant depression and post-traumatic stress disorder in Veterans: A retrospective case series. Pharmacotherapy:<br>The Journal of Human Pharmacology and Drug Therapy 2022;42:272–9. post-traumatic stress disorder in Veterans: A retrospective case series. A retrospective case in The Journal of<br>The Journal of Human Pharmacology and Drug Therapy 2022;42:272–9.<br>Ricca BP, Blaine BE. Brief Research Report:
- Ricca BP, Blaine BE. Brief Research Report: Notes on a Nonparametric Est<br>Size. The Journal of Experimental Education 2022;90. .<br>[27] Ricca Be. The Journal of Experimental Education 2022;90.<br>[27] Priest RG, Hawley CJ, Kibel D, et al. Recovery From Depressive Illness Does Fit an
- Size. The Journal of Experimental Education 2022;<br>Priest RG, Hawley CJ, Kibel D, et al. Recovery From De<br>Exponential Model. J Clin Psychopharmacol 1996;16. [27] Priest RG, Hawley CJ, Kibel D, et al. Recovery From Depressive Illness Does Fit an<br>Exponential Model. J Clin Psychopharmacol 1996;16.<br>[28] Berlow Y. Zandvakili A. Price L. et al. Modeling Treatment Response to Transcr
- ential Modeling Treat<br>Berlow Y, Zandvakili A, Price L, et al. Modeling Treatm<br>Magnetic Stimulation Using an Exponential Decay Fur Magnetic Stimulation Using an Exponential Decay Function. Biol Psychiatry<br>2021;89:S195–6.
- 2021;89:S195–6.<br>Jesus-Nunes AP, Leal GC, Correia-Melo FS, et al. Clinical predictors of depressive symptom Jesus-Nunes AP, L<br>remission and res Figure Manus Apple AP, 2008 Melon Melon Correia-Melon Frederick Correspondent<br>resistant depression. Human Psychopharmacology: Clinical and Experimental 2022;37:e2836 resistant depression. Human Psychopharmacology: Clinical and Experimental<br>2022;37:e2836. 2022;37:e2836.<br>37:e2836.

It is made available under a CC-BY-ND 4.0 International license. medRxiv preprint doi: [https://doi.org/10.1101/2023.07.03.23292094;](https://doi.org/10.1101/2023.07.03.23292094) this version posted July 5, 2023. The copyright holder for this preprint<br>(which was not certified by peer review) is the author/funder, who has granted med

- $|30|$ Khan A, Leventhal RM, Khan SR, et al. Severity of Depression and Response to<br>Antidepressants and Placebo: An Analysis of the Food and Drug Administration Database.<br>J Clin Psychopharmacol 2002;22.
- Kirsch I, Deacon BJ, Huedo-Medina TB, et al. Initial severity and antidepressant benefits: Kirsch I, Deacon BJ, Huedo-Medir<br>A meta-analysis of data submitte A meta-analysis of data submitted to the food and drug administration. PLoS Med<br>2008;5. A meta-analysis of data submitted to the food and drug and drug administration. PLOMED<br>2008;5.<br>Fournier JC, DeRubeis RJ, Hollon SD, et al. Antidepressant drug effects and depress
- -----,--<br>Fournier<br>severity Endia<br>132] Su TP, Chen MH, Li CT, et al. Dose-related effects of adjunctive ketamine in taiwanese
- severing: A patient-level management and also seed<br>Su TP, Chen MH, Li CT, et al. Dose-related effects of adjui<br>patients with treatment-resistant depression. Neuropsyd patients with treatment-resistant depression. Neuropsychopharmacology 2017;42.<br>[34] Joneborg I, Lee Y, Di Vincenzo JD, et al. Active mechanisms of ketamine-assisted
- patients with treatment constraint depression. Neuropsychopharmacology 2021/2021<br>Joneborg I, Lee Y, Di Vincenzo JD, et al. Active mechanisms of ketamine-assisted<br>psychotherapy: A systematic review. J Affect Disord 2022;315 [34] Joneborg I, Lee Y, Di Vincenzo JD, et al. Active mechanisms of ketapsychotherapy: A systematic review. J Affect Disord 2022;315.  $p$ systematic pychotherapy. A systematic review  $2(2\pi)^2$

## Figure 1



## Figure 2



## Figure 3

medRxiv preprint doi: https://doi.org/10.1101/2023.07.03.23292094; this version posted July 5, 2023. The copyrin<br>
(which was not certified by peer review) is the authority who has granted medRxiv a license to display to Cl medRxiv preprint doi: [https://doi.org/10.1101/2023.07.03.23292094;](https://doi.org/10.1101/2023.07.03.23292094) this version posted July 5, 2023. The copyright holder for this preprint<br>(which was not certified by peer review) is the authoriunder, who has granted medR





