STATISTICAL ANALYSIS PLAN

Datathon 3

Authors: Han Chang Chiam, Giulia Varvara, Minoo Matbouriahi, Yazid Zalai, Ulrich Mansmann, Martin Plöderl

Statistical Analysis Plan Status: Version 1.0
Statistical Analysis Plan Date: 29/06/2026

Team 2

Contents

List of Abbreviations and Definitions

TERM	Abbreviation / Definition
IPD	Individual Participant Data
MDD	Major Depressive Disorder
PHQ-9	Patient Health Questionnaire-9
MAP	meta-analytic predictive
ICC	Intraclass Correlation Coefficient
VR	Variability ratio
ANCOVA	Analysis of Covariance
PITE	Predicted individual treatment effect
LASSO	Least Absolute Shrinkage and Selection Operator
HTE	Heterogeneity of treatment effect
PATH	Statement on Predictive Approaches to Treatment effect Heterogeneity

1 Medical background

1.1 Super-Responders

Esketamine is statistically significantly superior to placebo regarding the reduction of depression symptoms as assessed with the Montgomery–Åsberg Depression Rating Scale (MADRS) [1]. However, the drug placebo difference is only about 3 MADRS points, corresponding with a standardized mean difference (SMD) of about 0.3. This efficacy estimate is substantially smaller than those reported in early ketamine trials [2]. Moreover, an efficacy of this size is likely not clinically meaningful according to conventional thresholds of clinical significance [3] and hardly superior to conventional antidepressants [4].

However, (small) average drug-placebo difference may be misleading because some patients may benefit especially well due to heterogeneity of treatment effect (HTE), opening up the quest for "precision medicine", that is, finding predictors who may benefit especially well from a treatment. Successful prediction of "super-responders" could make it possible to only expose those patients to the side effects and costs of a drug who actually benefit, thus guaranteeing a positive harm/benefit ratio. Consequently, precision psychiatry has become a hot topic, especially with big data and the expanding biomarker research with its emerging potential predictors of treatment success [5, 6]. However, despite massive research efforts to predict treatment success with conventional antidepressants, no robust predictors useful for clinical practice have been found [7]. The quest is ongoing [8] and there are some promising first exploratory results based on meta-research using individual patient level data (IPD) [9].

Approaches to explore HTE can be twofold. First, potential predictors of HTE can be detected via interaction effects between the predictor and the treatment (drug/placebo) in IPD data of clinical trials. For esketamine, one IPD meta-analysis on gender as treatment effect moderator found one statistically significant effect out [10], though there was no adjustment for multiple testing and this finding may be a false positive.

Second, it can be explored if HTE can be assumed at all by an investigation of the distributions of the outcome in the drug and placebo arms. If there are "super responders" in the drug arms, then this should alter the outcome distribution in the drug arm but not the placebo arm. Indeed, a mixture model analysis of individual patient data from conventional antidepressants showed that the non-normal outcome distributions could be decomposed into three different subdistributions with finite mixture modelling [11]. This included a subdistribution of "large" responses, and in the antidepressant arms, 15% more were categorized in this group, compared to placebo. However, the trimodal distribution could not be replicated for self-reported depression and this raises the question if issues such as rating biases may explain non-normality in response distributions [12]. Another approach is to examine variances in treatment and placebo arms. If there is a subgroup of patients responding especially well to the treatment, there should be a larger variance in the drug-arms. This has been investigated for conventional antidepressants via variance-ratio meta-analyses, but without evidence for HTE [13, 14].

We use both approaches to explore HTE in esketamine trials. We are aware of the inherent power issues for both mixture models and variance ratio meta-analyses. That is, we can only detect larger HTE. However, we have the largest IPD data available and at least we can estimate boundaries of HTE.

1.2 Rating biases in clinician versus patient ratings

Most clinical trials use clinician rating scales such as the MADRS for the primary outcome. For conventional antidepressants and psychotherapy, outcomes based on self-reports produced smaller drug-placebo differences [15, 16, 17], raising the issue of rating biases and psychometric differences and the need to go beyond relying on one type of assessment [16, 18]. Rating biases can occur due to a combination of unblinding and expectancy effects. If raters can guess correctly the treatment arms, their rating may be influenced by existing beliefs about the effects of the treatments and this usually leads to more favorable efficacy ratings, for example in open label trials versus trials with blinded raters [19]. For ketamine and esketamine, due to the pronounced psychotropic effects, blinding is hardly possible. The few studies where blinding success was assessed showed that clinicians, patients, but also independent raters can guess better than chance if patients received (es)ketamine or placebo [20]. Consequently, raters may, unconsciously or consciously, tend to choose a different response in line with their expectations, especially when there is ambiguity which response level to choose from. For example, clinicians who believe that esketamine is clearly superior to placebo may choose the response option indicating lower depression symptoms. Moreover, given the frequent ambiguity about which response option is best fitting patients description of the depression symptom and time pressure, clinicians may tend to harmonize the response options, meaning that there is less variation in the response options. This may be less an issue for patient self reports, because patients likely have less time pressure and maybe different expectations about treatments. Therefore, it has been suggested that a comparison between self- and patient reports may be used as an estimation of rating bias [21].

To investigate this for esketamine, we will compare efficacy estimates from self-report and clinician rating scales and also compare if ratings tend to be harmonized differentially in clinician-rating scales than in patients

rating scales. For esketamine, patient ratings were already used anchoring to explore how this corresponds with improvement on clinician rating scales [22], but the correlations and comparisons of variances were not investigated yet. Clinician ratings and patient ratings were compared for the TRANSFORM II trial but not from TRANSFORM I and III [22].

Unfortunately, blinding success was not directly assessed in Janssen's esketamine trials. The occurrence of adverse events may be used as a proxy for functional unblinding. This was already investigated using dissociation, a common effect of esketamine. One study used the TRANSFORM II study and the SUSTAIN study data and found no effects for dissociative experience on efficacy [23]. Another study pooled the TRANSFORM I and II studies and again did not find effects of dissociation on efficacy outcomes in linear mixed effects model. However, dissociation on day 1 was significantly associated with a significant reduction in the MADRS score at day 2 [23].

2 Study Overview

This statistical analysis plan (SAP) describes secondary analyses of individual participant data (IPD) from randomized, placebo-controlled esketamine trials in adults with major depressive disorder (MDD). The analyses will be performed using data made available through the Yale Open Data Access (YODA) Project Platform [24].

The SAP focuses on two predefined clinical topics:

1. **Topic 1: Heterogeneity of treatment effect in MADRS.** We will first describe MADRS outcome distributions and baseline-adjusted residual distributions by treatment arm. We will then assess evidence compatible with treatment-effect heterogeneity using variance-ratio analyses and finite mixture models. If these analyses suggest relevant heterogeneity, we will explore whether baseline covariates help explain this heterogeneity using predicted individual treatment effect models and flexible exploratory modelling. As an optional methodological extension, if relevant baseline variables are identified in the covariate-based HTE analyses, we may use simulation to illustrate how individual-level borrowing from historical trials, while accounting for baseline covariates, could affect the estimation of subgroup-specific treatment effects.
2. **Topic 2: Comparison with PHQ-9.** Where PHQ-9 data are available, we will repeat the main distributional analyses, including descriptive plots, variance-ratio analyses, and finite mixture models, and compare the findings descriptively with the MADRS results. Optional analyses may consider the joint behaviour of MADRS and PHQ-9, including their correlation.

2.1 Overall Study Objective

The overall objective is to use shared IPD from randomized esketamine trials to investigate whether the average treatment effect conceals clinically relevant variation between patients, whether treatment effects differ between clinician-rated and patient-reported outcomes, whether previous depression trial data can be used responsibly for historical borrowing.

The analyses are exploratory and hypothesis-generating. They are intended to clarify methodological questions, identify signals worthy of future confirmation, and support transparent analysis strategies for secondary use of clinical trial data.

3 Study Methods

3.1 Trial Eligibility Criteria

Eligible trials must meet all of the following criteria:

1. randomized, double-blind, placebo-controlled clinical trial;
2. evaluation of esketamine as an intervention;
3. adult participants aged 18 years or older;
4. diagnosis of major depressive disorder;
5. availability of individual participant data;
6. availability of baseline depression severity and randomized treatment assignment;
7. availability of at least one post-baseline depression severity assessment.

Trials will be classified according to clinical population and design characteristics. In particular, trials in treatment-resistant depression and trials in patients with acute suicidality will be analysed separately unless pooling is clinically and methodologically justified.

3.2 Participant Eligibility Criteria

Eligible participants are randomized adults with known treatment assignment and available baseline depression severity assessment. Endpoint-specific analyses will include participants with the relevant outcome observed or derivable within the predefined visit window. Participants without endpoint data will be included in descriptive summaries of missing data and in longitudinal or joint modelling analyses where appropriate.

3.3 Intervention and Comparator

The intervention is esketamine administered according to the trial-specific protocol. Esketamine may have been administered as augmentation to standard-of-care antidepressant therapy or, where applicable, as monotherapy.

The comparator is placebo administered according to the trial-specific protocol, either as augmentation to standard of care or as monotherapy.

3.4 Outcomes

3.4.1 Primary Clinician-Rated Outcome

The primary clinician-rated outcome is the Montgomery–Åsberg Depression Rating Scale (MADRS) total score at the main endpoint.

For treatment-resistant depression trials, the main endpoint is Day 28, defined as the assessment closest to Day 28 within a Day 25–28 window. If multiple assessments are available within this window, the latest assessment will be used.

For trials in patients with acute suicidality, the main endpoint is 24 hours after the first dose. If an exact 24-hour assessment is not available, the assessment closest to 24 hours after first dose will be used.

Because these populations and endpoint timings address different clinical questions, treatment-resistant depression trials and acute suicidality trials will not be pooled in the primary analysis unless there is a clear justification.

3.4.2 Patient-Reported Outcome

The main patient-reported outcome is the Patient Health Questionnaire-9 (PHQ-9) total score at the corresponding endpoint, where available. Other patient-reported depression outcomes may be included if they can be harmonized across trials and measured at comparable time points.

4 Analysis Populations

4.1 Main Analysis Population

The main analysis population will include all randomized participants with known treatment assignment and available baseline MADRS total score.

4.2 Endpoint Analysis Population

Endpoint analyses will include participants from the main analysis population with an observed endpoint value within the predefined visit window.

4.3 Clinician-Rated Versus Patient-Reported Outcome Population

Analyses comparing clinician-rated and patient-reported outcomes will be restricted to participants and trials with both outcome types measured at comparable time points.

5 General Statistical Principles

All analyses will account for the multi-trial structure of the data. Trial will be included as a fixed effect in simple endpoint models and as a random effect or hierarchical level in meta-analytic and Bayesian models where appropriate.

Continuous outcomes will be summarized by treatment arm and trial using means, standard deviations, medians, interquartile ranges, minima, maxima, and graphical displays of the full distribution. Binary outcomes will be summarized using counts and percentages.

Treatment effects and relevant measures of interest (Correlations, regression coefficients, ICCs, etc.) will be reported with confidence intervals or credible intervals, as appropriate. Because the analyses are exploratory, p-values will be interpreted descriptively. Emphasis will be placed on effect sizes, uncertainty, consistency across trials, and sensitivity to modelling assumptions.

No formal adjustment for multiplicity is planned. The interpretation of findings will distinguish between confirmatory evidence, exploratory evidence, and hypothesis-generating signals.

6 Statistical Analysis

6.1 Module 1: Heterogeneity of Treatment Effect and Possible Super-Responders

6.1.1 Analysis of the Primary Endpoint

Let i index participants and $r = 1, \ldots, R$ index trials. Let Y_{ir} denote the endpoint MADRS total score, B_{ir} the baseline MADRS total score, and $T_{ir} \in \{0, 1\}$ the randomized treatment indicator, where $T_{ir} = 1$ denotes esketamine and $T_{ir} = 0$ denotes placebo. Trial fixed effects are denoted by η_r. The primary endpoint analysis will use an ANCOVA model adjusting for baseline MADRS total score, treatment arm, and trial:

$$Y_{ir} = \alpha + \beta T_{ir} + \gamma B_{ir} + \eta_r + \epsilon_{ir}.$$

The coefficient β estimates the baseline- and trial-adjusted mean difference in endpoint MADRS between esketamine and placebo. Since lower MADRS scores indicate lower depression severity, negative values of β favour esketamine.

MADRS at day 28 (treatment resistant depression) or day 2 (suicidal depression) and baseline MADRS by trial and arm will be summarized descriptively and used for graphical visualization of distributions.

Variability-ratio analysis. Variability-ratio analyses will be used to compare the variability of baseline-adjusted MADRS outcomes between the esketamine and placebo groups. Trial-specific ANCOVA models will first be fitted:

$$Y_{ir} = \alpha_r + \beta_r T_{ir} + \gamma_r B_{ir} + \epsilon_{ir}, \qquad i \in r.$$

Within each trial r, the adjusted variability ratio will be defined as

$$\mathrm{VR}_r = \frac{\mathrm{SD}\left(\hat{\epsilon}_{ir} : T_{ir} = 1\right)}{\mathrm{SD}\left(\hat{\epsilon}_{ir} : T_{ir} = 0\right)}.$$

The log variability ratios, $\log(\mathrm{VR}_r)$, will then be synthesized across trials using random-effects meta-analysis:

$$\log(\mathrm{VR}_r) = \theta + u_r + e_r, \qquad u_r \sim N(0, \tau^2),$$

where θ is the average log variability ratio across trials, u_r is the between-trial random effect, and e_r is the trial-specific sampling error. Values of $\mathrm{VR}_r > 1$ indicate greater adjusted outcome variability in the esketamine arm than in the placebo arm. This will be interpreted as evidence compatible with treatment-effect heterogeneity.

Mixture-model analysis. Mixture-model analyses will be used to explore whether the distribution of baseline- and trial-adjusted MADRS residuals can be decomposed into latent response components. First, residuals will be obtained from the pooled ANCOVA model:

$$\tilde{\epsilon}_{ir} = Y_{ir} - \left(\hat{\alpha} + \hat{\beta} T_{ir} + \hat{\gamma} B_{ir} + \hat{\eta}_r\right).$$

A Gaussian finite mixture model will then be fitted to $\tilde{\epsilon}_{ir}$. For K latent components,

$$f(\tilde{\epsilon}_{ir} \mid T_{ir}) = \sum_{\ell=1}^{K} \pi_\ell(T_{ir}) \phi\left(\tilde{\epsilon}_{ir}; \mu_\ell, \sigma_\ell^2\right),$$

where $\phi(\cdot; \mu_\ell, \sigma_\ell^2)$ denotes a Gaussian density with component-specific mean μ_ℓ and variance σ_ℓ^2, and $\pi_\ell(T_{ir})$ denotes the probability of belonging to component ℓ, allowed to differ by treatment arm. Models with $K = 1, 2,$ and 3 components will be compared using BIC, component stability, clinical interpretability, and consistency across trials. A favourable response component will be defined as a component with lower residual MADRS values, because lower endpoint MADRS scores indicate better outcome after adjustment for baseline and trial. Such a component will be interpreted as compatible with treatment-related super-response only if it is more frequent in the esketamine arm than in the placebo arm and is not dominated by a single trial.

6.1.2 Sensitivity Analyses on the Primary Endpoints

As a further exploratory assessment, item-level MADRS analyses may be considered if item-level data are available and sufficiently complete, particularly if the variability-ratio or mixture-model analyses suggest distributional heterogeneity. This analysis will not redefine the primary endpoint and will not use the latent IRT score as a replacement outcome. Instead, it will be used to assess whether the MADRS total-score findings may be related to item-level response patterns, item weighting, or floor/ceiling effects. This sensitivity analysis is based on the graded-response IRT model previously applied to MADRS item data. In this model, the observed MADRS item scores are assumed to arise from an underlying latent depression severity variable, denoted here by ψ. The model relates each ordinal item score to ψ through item-specific discrimination and threshold parameters, as described for the MADRS by Carmody et al. and subsequently used in a ketamine trial by Otto et al. [25, 26].

Let U_{irmv} denote the observed score for participant i in trial r, MADRS item $m = 1, \ldots, 10$, and visit $v \in \{0, 28\}$, where $U_{irmv} \in \{0, 1, \ldots, 6\}$. For item m and category threshold $c = 1, \ldots, 6$, the graded-response IRT model is

$$P(U_{irmv} \geq c \mid \psi_{irv}) = \text{logit}^{-1}\left\{a_m(\psi_{irv} - b_{mc})\right\},$$

where ψ_{irv} is latent depression severity, a_m is the item discrimination parameter, and b_{mc} is the item threshold parameter. The probability of a specific item score can be obtained from adjacent cumulative probabilities:

$$P(U_{irmv} = c \mid \psi_{irv}) = P(U_{irmv} \geq c \mid \psi_{irv}) - P(U_{irmv} \geq c + 1 \mid \psi_{irv}).$$

Item characteristic curves and item/test information curves will be used to assess which MADRS items are most informative across the latent severity range and whether floor or ceiling effects are present. The favourable response component identified from the MADRS total-score mixture model will then be examined at the item level. Specifically, Day 28 item-score distributions, item-level floor proportions, and item-level changes from baseline will be compared across treatment arms and mixture components. If the favourable component is broadly reflected across multiple MADRS items, this will support the interpretation of a general depression-severity response pattern. If the component is concentrated in a small number of items, in low-discrimination items, or in near-floor responses, this will suggest that the apparent treatment-effect heterogeneity in the MADRS total score may partly reflect item weighting, symptom-specific response, or floor effects.

6.1.3 Exploratory analyses

If the variance-ratio and mixture-model analyses of MADRS provide evidence for treatment effect heterogeneity, subsequent exploratory analyses will investigate whether baseline characteristics are associated with larger treatment benefit or membership in a favourable response component. Let $\mathbf{X}_{ir}$ denote the vector of pre-specified candidate baseline covariates for participant i in trial r. Candidate variables may include age, sex, race, employment status, education level, number of previous depressive episodes, duration of the current depressive episode, family history of psychiatric disorders, past substance use disorder, and baseline depression severity.

Univariate treatment-covariate interaction models. Conventional univariate treatment-covariate interaction models will be fitted separately for each candidate baseline variable.

$$Y_{ir} = \alpha + \beta T_{ir} + \gamma B_{ir} + \eta_r + \lambda_k X_{kir} + \delta_k T_{ir} X_{kir} + \epsilon_{ir}.$$

The coefficient δ_k is the treatment-covariate interaction estimate. Because lower MADRS scores indicate better outcome, $\delta_k < 0$ indicates that larger values of X_k are associated with greater predicted benefit from esketamine, assuming $T_{ir} = 1$ denotes esketamine.

However, conventional treatment-covariate interaction analyses may miss more complex forms of HTE because they usually assess one or a few prespecified covariates and often assume linear or additive effect modification. Following the PATH statement [27], we will therefore also consider predictive approaches to estimate treatment benefit across multivariable baseline covariate profiles, including PITE models with penalised interaction selection and BART as a flexible exploratory model.

PITE-LASSO model. A multivariable predicted individual treatment effect (PITE) model with LASSO selection will be considered [28]. The working model is

$$Y_{ir} = \alpha + \beta T_{ir} + \gamma B_{ir} + \eta_r + \mathbf{X}_{ir}^{\top}\boldsymbol{\lambda} + T_{ir}\mathbf{X}_{ir}^{\top}\boldsymbol{\delta} + \epsilon_{ir}.$$

The main effects, treatment term, baseline MADRS, and trial effects will be retained in the model, while the treatment-covariate interaction coefficients $\boldsymbol{\delta}$ may be penalised. The penalised estimation problem can be written as

$$\min \left[\sum_{i,r} \left\{ Y_{ir} - \alpha - \beta T_{ir} - \gamma B_{ir} - \eta_r - \mathbf{X}_{ir}^{\top}\boldsymbol{\lambda} - T_{ir}\mathbf{X}_{ir}^{\top}\boldsymbol{\delta} \right\}^2 + \lambda_{\text{pen}} \sum_{k=1}^{p} |\delta_k| \right],$$

where $\mathbf{X}_{ir}$ denotes candidate baseline covariates other than baseline MADRS, λ_{pen} is the tuning parameter controlling the degree of shrinkage. For participant i in trial r, the predicted individual treatment effect will be defined as

$$\widehat{\text{PITE}}_{ir} = \widehat{Y}_{ir}(T_{ir} = 1) - \widehat{Y}_{ir}(T_{ir} = 0) = \hat{\beta} + \mathbf{X}_{ir}^{\top}\hat{\boldsymbol{\delta}}.$$

Negative PITE values indicate greater predicted benefit from esketamine, because lower MADRS scores indicate better outcome.

BART model. Bayesian additive regression trees (BART) may be fitted as a flexible exploratory model for conditional treatment effects. The BART model will include treatment, baseline MADRS, trial, and candidate baseline covariates $\mathbf{X}_{ir}$:

$$Y_{ir} = f(T_{ir}, B_{ir}, \mathbf{X}_{ir}, r) + \epsilon_{ir}, \qquad \epsilon_{ir} \sim N(0, \sigma^2),$$

where the unknown regression function $f(\cdot)$ is represented as a sum of regression trees:

$$f(T_{ir}, B_{ir}, \mathbf{X}_{ir}, r) = \sum_{h=1}^{H} g\left(T_{ir}, B_{ir}, \mathbf{X}_{ir}, r; \mathcal{T}_h, \mathcal{M}_h\right).$$

Here, $\mathcal{T}_h$ denotes the structure of tree h, $\mathcal{M}_h$ denotes its terminal-node parameters, and H is the number of trees. The individual conditional treatment effect will be estimated by contrasting posterior predictions under esketamine and placebo for the same baseline covariate profile:

$$\Delta_{ir}^{\text{BART}} = E\{Y_{ir} \mid T_{ir} = 1, B_{ir}, \mathbf{X}_{ir}, r\} - E\{Y_{ir} \mid T_{ir} = 0, B_{ir}, \mathbf{X}_{ir}, r\}.$$

Negative values of $\Delta_{ir}^{\text{BART}}$ indicate greater predicted benefit from esketamine. BART will be used to explore nonlinear or multivariable patterns that may not be captured by conventional interaction models or LASSO. Variable-importance and variable-dependence summaries will be interpreted descriptively.

Analyses will be implemented in R using `metafor` for variability-ratio meta-analysis, `flexmix` for mixture models, `mirt` for IRT item-characteristic curves, `glmnet` for PITE-LASSO, and `bartCause`/`tidytreatment` for exploratory BART analyses. Example code is provided in Appendix.

Optional borrowing analysis If the exploratory HTE analyses identify a plausible baseline-defined subgroup or covariate pattern associated with larger predicted treatment benefit, an optional simulation-based borrowing illustration may be conducted. This analysis will illustrate how an identified HTE signal could inform the design or analysis of a future subgroup-enriched trial.

Briefly, a hypothetical new randomized trial enriched for the candidate high-benefit subgroup would be simulated. The subgroup treatment effect from this new trial alone would be compared descriptively with estimates obtained by borrowing historical information in two ways: (i) general trial-level borrowing from the overall historical trial population, and (ii) baseline-informed individual-level borrowing, where historical participants more similar to the subgroup-enriched target population receive greater weight. The purpose is to illustrate that, when treatment benefit is heterogeneous, borrowing should target the covariate distribution of the future trial population rather than the overall historical average. Results would be interpreted as a methodological illustration only.

6.1.4 Distributional Comparison

The full distribution of baseline-adjusted MADRS outcomes will be compared between treatment arms using:

1. density plots;
2. empirical cumulative distribution functions;
3. quantile plots;
4. arm-specific means and variances;
5. graphical summaries by trial and treatment arm.

The purpose is to assess whether treatment differences are compatible with a simple location shift or whether the esketamine arm shows evidence of greater dispersion, asymmetry, or tail effects.

6.2 Module 2: Clinician-Rated Versus Patient-Reported Outcomes and Functional Unblinding

6.2.1 Objective

The objective of this module is to compare treatment effects estimated from clinician-rated and patient-reported depression outcomes and to explore whether discrepancies between these outcomes are associated with treatment assignment, adverse events suggestive of functional unblinding, or trial characteristics.

6.2.2 Harmonization of Outcome Scales

Clinician-rated MADRS and patient-reported PHQ-9 scores will be analysed on their original scales and, where necessary, on standardized scales. Standardization will be performed within trial and time point:
$Z_{ij} = \frac{Y_{ij} - \bar{Y}_j}{SD(Y_j)}$.
This will allow comparison of treatment effects across instruments with different ranges and measurement properties.

6.2.3 Treatment Effects by Outcome Type

Treatment effects will be estimated separately for clinician-rated and patient-reported outcomes using ANCOVA models adjusted for baseline score and trial.

For clinician-rated outcomes:
$Y_{ij}^C = \beta_0^C + \beta_1^C T_{ij} + \beta_2^C Y_{ij0}^C + \gamma_j^C + \varepsilon_{ij}^C$.
For patient-reported outcomes:
$Y_{ij}^P = \beta_0^P + \beta_1^P T_{ij} + \beta_2^P Y_{ij0}^P + \gamma_j^P + \varepsilon_{ij}^P$.
The key comparison will be whether the standardized treatment effect is larger for clinician-rated outcomes than for patient-reported outcomes.

A larger clinician-rated effect may suggest rater sensitivity or rating bias, but alternative explanations such as differences in symptom domains, measurement precision, and patient-clinician perspectives will be considered.

6.2.4 Association and Agreement

Association between clinician-rated and patient-reported outcomes will be assessed using Pearson or Spearman correlations, depending on distributional properties. If meaningful, agreement will be studied by estimating intraclass correlation coefficients (ICCs) which can be derived from random effect regression models and the estimated variances of specific random effects.

Agreement will be assessed using:

1. cross-tabulation of severity categories;
2. Bland–Altman-type plots;
3. ICCs
4. standardized discrepancy scores;
5. trial- and treatment-specific graphical summaries.

6.2.5 Calibration between clinical and patient-reported outcome

For each trial, we will estimate the relationship between MADRS total score and PHQ-9 total score using linear regression with MADRS as the dependent variable and PHQ-9 as the independent variable. Trial-specific intercepts, slopes, residual standard deviations, and predicted MADRS values at selected PHQ-9 scores will be reported. Non-linearity will be explored using restricted cubic splines. Trial-specific calibration parameters will be combined using random-effects meta-analysis, preferably using a bivariate model for intercept and slope. As an alternative clinically interpretable approach, predicted MADRS values at pre-specified PHQ-9 anchor points will be meta-analysed. Heterogeneity across trials will be quantified and displayed using forest plots and prediction intervals. Sensitivity analyses will include change-score models, standardized-score models, and reverse calibration with PHQ-9 as the dependent variable.

The primary interpretation should focus on whether PHQ-9 can provide a stable estimate of clinician-rated depression severity as measured by MADRS.

Possible conclusions include:

1. *Good calibration*: similar intercepts and slopes across trials, low between-trial heterogeneity, narrow prediction intervals.
2. *Moderate calibration*: average relationship is clear, but trial-specific variation limits precise conversion.
3. *Poor calibration*: substantial heterogeneity, non-linearity, or wide prediction intervals prevent use of a common conversion rule.

The analysis estimates how PHQ-9 scores translate, on average, into MADRS scores across trials. However, because both scales capture overlapping but non-identical constructs, the calibration function should not be interpreted as evidence that the scales are interchangeable at the individual-patient level.

6.2.6 Discrepancy Score

A discrepancy score will be defined as:

$D_{ij} = Z^C_{ij} - Z^P_{ij}$,

where Z^C_{ij} is the standardized clinician-rated score and Z^P_{ij} is the standardized patient-reported score.

Positive values indicate higher severity according to the clinician-rated measure relative to the patient-reported measure. Negative values indicate lower clinician-rated severity relative to patient report.

Discrepancy scores will be modeled as:

$D_{ij} = \alpha_0 + \alpha_1 T_{ij} + \alpha_2 D_{ij0} + \alpha_3 Y^C_{ij0} + \gamma_j + \varepsilon_{ij}$.

Additional models may include age, sex, clinical population, and other harmonized baseline characteristics.

6.2.7 Functional Unblinding

Adverse events likely to compromise blinding, including dissociation, sedation, dizziness, and other noticeable acute effects, will be summarized by treatment arm and trial.

Exploratory models will assess if dissociation at the latest time-point before assessment of the primary outcome is associated with:

1. larger clinician-rated improvement;
2. larger patient-reported improvement;
3. greater discrepancy between clinician-rated and patient-reported outcomes.

Because adverse events occur after randomization and may lie on the causal pathway between treatment and outcome, these analyses will be interpreted cautiously. They will be used to explore the plausibility of functional unblinding rather than to establish definitive causal effects.

7 Sensitivity Analyses

Sensitivity analyses may include:

1. separate analyses by clinical population;
2. exclusion of trials with non-comparable design features;

3. alternative adjustment sets for baseline covariates;
4. leave-one-trial-out analyses;
5. analyses restricted to participants with both clinician-rated and patient-reported outcomes;
6. analyses excluding participants with major protocol deviations, if such information is available;

8 Presentation of Results

Results will be reported with emphasis on effect sizes, uncertainty, trial-to-trial consistency, and graphical presentation.

The following displays are planned:

1. trial flow and availability of outcomes by analysis module;
2. baseline characteristics by trial and treatment arm;
3. endpoint MADRS distributions by treatment arm;
4. variance-ratio forest plot;
5. mixture-model component summaries;
6. clinician-rated versus patient-reported treatment effect comparison;
7. Bland-Altman plots;
8. discrepancy plots between clinician-rated and patient-reported outcomes;

All conclusions will distinguish clearly between evidence for average efficacy, evidence for heterogeneity, prediction of response, prediction of treatment benefit, and exploratory signals of rating bias.

9 References

References

[1] Florian Naudet et al. *Efficacy and safety of esketamine for treatment resistant depression: registered report for a Systematic Review with an Individual Patient Data Meta-analysis of Randomized, Double-Blind, Placebo-Controlled Trials*. en. July 2025. DOI: 10.1101/2025.07.17.25331696. URL: http://medrxiv.org/lookup/doi/10.1101/2025.07.17.25331696 (visited on 06/28/2026).

[2] Rebecca L Dean et al. "Ketamine and other glutamate receptor modulators for depression in adults with unipolar major depressive disorder". en. In: *Cochrane Database of Systematic Reviews* 2021.11 (Sept. 2021). Ed. by Cochrane Common Mental Disorders Group. ISSN: 14651858. DOI: 10.1002/14651858.CD011612.pub3. URL: http://doi.wiley.com/10.1002/14651858.CD011612.pub3 (visited on 06/28/2026).

[3] Michael P Hengartner and Martin Plöderl. "Estimates of the minimal important difference to evaluate the clinical significance of antidepressants in the acute treatment of moderate-to-severe depression". en. In: *BMJ Evidence-Based Medicine* 27.2 (Apr. 2022), pp. 69–73. ISSN: 2515-446X, 2515-4478. DOI: 10.1136/bmjebm-2020-111600. URL: https://ebm.bmj.com/lookup/doi/10.1136/bmjebm-2020-111600 (visited on 06/28/2026).

[4] Klaus Munkholm, Asger Sand Paludan-Müller, and Kim Boesen. "Considering the methodological limitations in the evidence base of antidepressants for depression: a reanalysis of a network meta-analysis". en. In: *BMJ Open* 9.6 (June 2019), e024886. ISSN: 2044-6055, 2044-6055. DOI: 10.1136/bmjopen-2018-024886. URL: https://bmjopen.bmj.com/lookup/doi/10.1136/bmjopen-2018-024886 (visited on 06/28/2026).

[5] Jonas Rüppel. "'Now Is a Time for Optimism": The Politics of Personalized Medicine in Mental Health Research". en. In: *Science, Technology, & Human Values* 44.4 (July 2019), pp. 581–611. ISSN: 0162-2439, 1552-8251. DOI: 10.1177/0162243919845049. URL: https://journals.sagepub.com/doi/10.1177/0162243919845049 (visited on 06/28/2026).

[6] Anneka Tomlinson et al. "Personalise antidepressant treatment for unipolar depression combining individual choices, risks and big data (PETRUSHKA): rationale and protocol". en. In: *Evidence Based Mental Health* 23.2 (May 2020), pp. 52–56. ISSN: 1362-0347, 1468-960X. DOI: 10.1136/ebmental-2019-300118. URL: https://mentalhealth.bmj.com/lookup/doi/10.1136/ebmental-2019-300118 (visited on 06/28/2026).

[7] Kelly Perlman et al. "A systematic meta-review of predictors of antidepressant treatment outcome in major depressive disorder". en. In: *Journal of Affective Disorders* 243 (Jan. 2019), pp. 503–515. ISSN: 01650327. DOI: 10.1016/j.jad.2018.09.067. URL: https://linkinghub.elsevier.com/retrieve/pii/S0165032718309947 (visited on 06/28/2026).

[8] Ellen Driessen et al. "Developing a multivariable prediction model to support personalized selection among five major empirically-supported treatments for adult depression. Study protocol of a systematic review and individual participant data network meta-analysis". en. In: *PLOS ONE* 20.4 (Apr. 2025). Ed. by Nishant Premnath Jaiswal, e0322124. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0322124. URL: https://dx.plos.org/10.1371/journal.pone.0322124 (visited on 06/28/2026).

[9] Yajie Xiang et al. "An Individual Participant Data Analysis: Optimizing Pharmacological Treatment in Children and Adolescents With Major Depressive Disorder". en. In: *Journal of the American Academy of Child & Adolescent Psychiatry* (May 2026), S0890856726002339. ISSN: 08908567. DOI: 10.1016/j.jaac.2026.05.010. URL: https://linkinghub.elsevier.com/retrieve/pii/S0890856726002339 (visited on 06/28/2026).

[10] Marie Huc et al. "Sex differences in placebo and antidepressant response to intranasal esketamine for treatment-resistant depression". en. In: *Molecular Psychiatry* 31.7 (July 2026), pp. 3674–3682. ISSN: 1359-4184, 1476-5578. DOI: 10.1038/s41380-026-03493-8. URL: https://www.nature.com/articles/s41380-026-03493-8 (visited on 06/28/2026).

[11] Marc B Stone et al. "Response to acute monotherapy for major depressive disorder in randomized, placebo controlled trials submitted to the US Food and Drug Administration: individual participant data analysis". en. In: *BMJ* (Aug. 2022), e067606. ISSN: 1756-1833. DOI: 10.1136/bmj-2021-067606. URL: https://www.bmj.com/lookup/doi/10.1136/bmj-2021-067606 (visited on 06/28/2026).

[12] Colin Xu et al. "Large responses to antidepressants or methodological artifacts? A secondary analysis of STARD, a single-arm, open-label, nonindustry antidepressant trial". en. In: *Journal of Clinical Epidemiology* 187 (Nov. 2025), p. 111943. ISSN: 08954356. DOI: 10.1016/j.jclinepi.2025.111943. URL: https://linkinghub.elsevier.com/retrieve/pii/S0895435625002768 (visited on 06/28/2026).

[13] Martin Plöderl and Michael Pascal Hengartner. "What are the chances for personalised treatment with antidepressants? Detection of patient-by-treatment interaction with a variance ratio meta-analysis". en. In: *BMJ Open* 9.12 (Dec. 2019), e034816. ISSN: 2044-6055, 2044-6055. DOI: 10.1136/bmjopen-2019-034816. URL: https://bmjopen.bmj.com/lookup/doi/10.1136/bmjopen-2019-034816 (visited on 06/28/2026).

[14] Constantin Volkmann, Alexander Volkmann, and Christian A. Müller. "On the treatment effect heterogeneity of antidepressants in major depression: A Bayesian meta-analysis and simulation study". en. In: *PLOS ONE* 15.11 (Nov. 2020). Ed. by Alan D. Hutson, e0241497. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0241497. URL: https://dx.plos.org/10.1371/journal.pone.0241497 (visited on 06/28/2026).

[15] Sarah E Hetrick et al. "New generation antidepressants for depression in children and adolescents: a network meta-analysis". en. In: *Cochrane Database of Systematic Reviews* 2021.5 (May 2021). Ed. by Cochrane Common Mental Disorders Group. ISSN: 14651858. DOI: 10.1002/14651858.CD013674.pub2. URL: http://doi.wiley.com/10.1002/14651858.CD013674.pub2 (visited on 06/28/2026).

[16] Glen I. Spielmans and Joseph P. McFall. "A Comparative Meta-Analysis of Clinical Global Impressions Change in Antidepressant Trials". en. In: *Journal of Nervous & Mental Disease* 194.11 (Nov. 2006), pp. 845–852. ISSN: 0022-3018. DOI: 10.1097/01.nmd.0000244554.91259.27. URL: https://journals.lww.com/00005053-200611000-00006 (visited on 06/28/2026).

[17] Clara Miguel et al. "Self-reports vs clinician ratings of efficacies of psychotherapies for depression: a meta-analysis of randomized trials". en. In: *Epidemiology and Psychiatric Sciences* 34 (2025), e15. ISSN: 2045-7960, 2045-7979. DOI: 10.1017/S2045796025000095. URL: https://www.cambridge.org/core/product/identifier/S2045796025000095/type/journal_article (visited on 06/28/2026).

[18] Meike Shedden Mora, Yvonne Nestoriuc, and Winfried Rief. "Lessons learned from placebo groups in antidepressant trials". en. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 366.1572 (June 2011), pp. 1879–1888. ISSN: 0962-8436, 1471-2970. DOI: 10.1098/rstb.2010.0394. URL: http://royalsocietypublishing.org/rstb/article/366/1572/1879-1888/21633 (visited on 06/28/2026).

[19] Asbjørn Hróbjartsson et al. "Observer bias in randomized clinical trials with measurement scale outcomes: a systematic review of trials with both blinded and nonblinded assessors". en. In: *Canadian Medical Association Journal* 185.4 (Mar. 2013), E201–E211. ISSN: 0820-3946, 1488-2329. DOI: 10.1503/cmaj.120744. URL: http://www.cmaj.ca/lookup/doi/10.1503/cmaj.120744 (visited on 06/28/2026).

[20] Martin Plöderl et al. *Effects of ketamine and esketamine on death, suicidal behaviour, and suicidal ideation in psychiatric disorders: A systematic review and meta-analysis*. en. Aug. 2025. DOI: 10.1101/2025.08.19.25333796. URL: http://medrxiv.org/lookup/doi/10.1101/2025.08.19.25333796 (visited on 06/28/2026).

[21] E Petkova. "A Method to Quantify Rater Bias in Antidepressant Trials". In: *Neuropsychopharmacology* 22.6 (June 2000), pp. 559–565. ISSN: 0893133X. DOI: 10.1016/S0893-133X(99)00154-2. URL: https://www.nature.com/doifinder/10.1016/S0893-133X(99)00154-2 (visited on 06/28/2026).

[22] Lysbeth Floden et al. "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". en. In: *CNS Drugs* 36.6 (June 2022), pp. 649–658. ISSN: 1172-7047, 1179-1934. DOI: 10.1007/s40263-022-00916-2. URL: https://link.springer.com/10.1007/s40263-022-00916-2 (visited on 06/28/2026).

[23] David S. Mathai et al. "Reconsidering "dissociation" as a predictor of antidepressant efficacy for esketamine". en. In: *Psychopharmacology* 240.4 (Apr. 2023), pp. 827–836. ISSN: 0033-3158, 1432-2072. DOI: 10.1007/s00213-023-06324-8. URL: https://link.springer.com/10.1007/s00213-023-06324-8 (visited on 06/28/2026).

[24] Harlan M. Krumholz and Joanne Waldstreicher. "The Yale Open Data Access (YODA) Project — A Mechanism for Data Sharing". en. In: *New England Journal of Medicine* 375.5 (Aug. 2016), pp. 403–405. ISSN: 0028-4793, 1533-4406. DOI: 10.1056/NEJMp1607342. URL: http://www.nejm.org/doi/10.1056/NEJMp1607342 (visited on 06/27/2026).

[25] Thomas J. Carmody et al. "The Montgomery Äsberg and the Hamilton ratings of depression: A comparison of measures". en. In: *European Neuropsychopharmacology* 16.8 (Dec. 2006), pp. 601–611. ISSN: 0924977X. DOI: 10.1016/j.euroneuro.2006.04.008. URL: https://linkinghub.elsevier.com/retrieve/pii/S0924977X06000769 (visited on 06/27/2026).

[26] Marije E. Otto et al. "Item response theory in early phase clinical trials: Utilization of a reference model to analyze the Montgomery-Åsberg Depression Rating Scale". en. In: *CPT: Pharmacometrics & Systems Pharmacology* 12.10 (Oct. 2023), pp. 1425–1436. ISSN: 2163-8306, 2163-8306. DOI: 10.1002/psp4.13018. URL: https://ascpt.onlinelibrary.wiley.com/doi/10.1002/psp4.13018 (visited on 06/27/2026).

[27] David M. Kent et al. "The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement: Explanation and Elaboration". en. In: *Annals of Internal Medicine* 172.1 (Jan. 2020), W1–W25. ISSN: 0003-4819, 1539-3704. DOI: 10.7326/M18-3668. URL: https://www.acpjournals.org/doi/10.7326/M18-3668 (visited on 06/27/2026).

[28] Nicolás M. Ballarini et al. "Subgroup identification in clinical trials via the predicted individual treatment effect". en. In: *PLOS ONE* 13.10 (Oct. 2018). Ed. by Alan Hubbard, e0205971. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0205971. URL: https://dx.plos.org/10.1371/journal.pone.0205971 (visited on 06/27/2026).

10 Appendix

```
library(metafor)
library(nlme)
library(flexmix)

# Suppose:
# madrs_d28 madrs_bl: MADRS at day 28 and baseline
# trt = 0/1 (0=control, 1=treated)
# x1, x2, x3 = baseline covariates
# trial_id is a factor for different trials
dat$trt_f <- factor(dat$trt, levels = c(0, 1), labels = c("placebo", "esketamine"))

# ============================================================
# ANCOVA for average treatment effect
# ============================================================
fit_ancova <- lm( madrs_d28 ~ trt_f + madrs_bl + trial_id, data = dat )
summary(fit_ancova)

# for each trial data d, extract trial-specific baseline adjusted residuals
fit <- lm( madrs_d28 ~ trt_f + madrs_bl, data = d )
d$resid_trial_adj <- resid(fit)

# ================================================
# Variance ratio analyses
# ================================================
# Trial-specific variability ratios
x_t <- d$resid_trial_adj[d$trt_f == "esketamine"]
x_c <- d$resid_trial_adj[d$trt_f == "placebo"]

# within-trial F tests on adjusted residual variances
test <- var.test(x_t, x_c)

# combine the variability ratios measures of all d into vr.input for
# meta analysis, with the structure:
data.frame(
trial_id = unique(d$trial_id),
n_t = length(x_t),
n_c = length(x_c),
```

```
sd_t = sd(x_t, na.rm = TRUE),
sd_c = sd(x_c, na.rm = TRUE) )

vr_dat <- escalc( measure = "VR", sd1i = sd_t, sd2i = sd_c, n1i = n_t, n2i = n_c, data = vr_input )

# Random-effects meta-analysis of log variability ratios
fit_vr <- rma(yi, vi, data = vr_dat, method = "REML")
summary(fit_vr)

# ============================================================
# Mixture model
# ============================================================
# Pooled baseline and trial adjusted residuals for mixture model
fit_resid_pool <- lm( madrs_d28 ~ trt_f + madrs_bl + trial_id, data = dat )
dat$madrs_resid_pool <- resid(fit_resid_pool)

# Gaussian mixture model, Q: how many mixture component? 2? 3?
set.seed(123)

mix_1 <- flexmix( madrs_resid_pool ~ 1, data = dat, k = 1,
model = FLXMRglm(family = "gaussian") )

mix_2 <- flexmix( madrs_resid_pool ~ 1, data = dat, k = 2,
model = FLXMRglm(family = "gaussian"), concomitant = FLXPmultinom(~ trt_f) )

mix_3 <- flexmix( madrs_resid_pool ~ 1, data = dat, k = 3,
model = FLXMRglm(family = "gaussian"), concomitant = FLXPmultinom(~ trt_f) )

# Compare number of components
BIC(mix_1, mix_2, mix_3)

# Select the preferred model, for example:
mix_best <- mix_2

# Component-specific parameters
parameters(mix_best)

# Posterior class assignment
dat$mix_class <- clusters(mix_best)

# Class proportions by treatment arm
prop.table(table(dat$trt_f, dat$mix_class), margin = 1)

# Check whether latent classes are dominated by specific trials
prop.table(table(dat$trial_id, dat$mix_class), margin = 1)

# ============================================================
# Univariate treatment-covariate interaction #
# ============================================================
lm(madrs_d28 ~ trt * X_k + madrs_bl + trial_id, data = dat)
# The coefficient of trt:X_k is the interaction estimate.

# ============================================================
# PITE-LASSO
# ============================================================
    library(glmnet)
    X_lasso <- model.matrix( ~ trt + madrs_bl + trial_id +
    X1 + X2 + X3 + X4 + X5 + trt:X1 + trt:X2 + trt:X3 +
```

```
    trt:X4 + trt:X5, data = dat )[, -1]
    y <- dat$madrs_d28
    # Penalise treatment-covariate interaction terms.
    # Main effects and adjustment variables are retained.
    penalty <- ifelse(grepl("trt:", colnames(X_lasso)), 1, 0)
    fit_lasso <- cv.glmnet( x = X_lasso, y = y, alpha = 1, penalty.factor = penalty )

# ============================================================
# BART
# ============================================================

library(bartMan)
library(dbarts)
library(bartCause)
library(tidytreatment)

# BART model for nonlinear outcome structure and interaction screening
fit_bart_screen <- bart(
  x.train = X,   # treatment, baseline MADRS, trial, and candidate covariates
  y.train = madrs_d28,
  keeptrees = TRUE
)

# bartMan visualisations may be used to inspect variable importance
# and treatment-covariate interactions.

# BART treatment-effect model
te_bart <- bartc(
  response = madrs_d28,
  treatment = trt,
  confounders = madrs_bl + trial_id + X1 + X2 + X3 + X4 + X5,
  data = dat,
  method.rsp = "bart",
  method.trt = "none",
  p.scoreAsCovariate = FALSE # since we are using trial data I assume randomization works.
)

# Conditional treatment effects
bart_pite <- treatment_effects(te_bart)
```