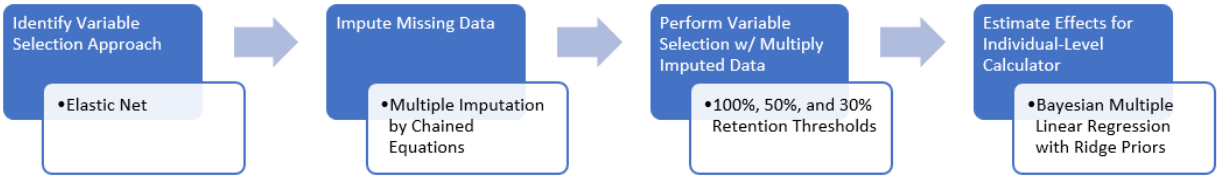
Supplemental File – Statistical Methods

The challenge of the EMBARC study was to account for missing data while implementing deep clinical and biological phenotyping in the setting of a multi-site randomized controlled trial. This further complicated the effort to identify a parsimonious set of meaningful predictors of placebo response (along with their magnitude and direction of effects) given the large set of a priori identified candidate predictors (p=285) exceeded the sample size recruited (n=141). Figure 1 below shows the progression of the analysis.

Supplementary Figure 1 – Analysis Flowchart



Penalized regression techniques such as the lasso and the elastic net are popular tools for variable selection; these methods can be thought of as an extension of ordinary least squares (OLS), which estimates regression parameters by minimizing the sum of squared residuals:

$$\min_{\beta_0,\beta} \sum_{i=1}^{N} \left(y_i - (\beta_0 + \beta^T \boldsymbol{x_i})\right)^2,$$

where y_i is a continuous outcome of interest for the i^{th} subject, β_0 is the intercept term, β is a $1 \times p$ vector of regression parameters, and $\boldsymbol{x_i}$ is an $1 \times p$ vector of predictor variables for the i[th] subject. The elastic net – the method we utilized – also minimizes the sum of squared residuals, but adds additional constraints on both the sum of the squared regression coefficients and the sum of the absolute value of the regression coefficients:

$$\min_{\beta_0,\beta} \sum_{i=1}^{N} \left(y_i - (\beta_0 + \beta^T \boldsymbol{x_i})\right)^2 + \lambda[(1-\alpha)\,\|\beta\|_2^2/2 + \alpha\|\beta\|_1]$$

Mathematically, these additional constraints cause shrinkage in the regression parameters – potentially to 0 – thereby removing these parameters from the model entirely (but only those whose estimates have been shrunken to exactly 0). In other words, variable selection is performed. While this shrinkage does cause the resulting estimates to be biased, the variance of the estimate is also reduced as we are optimizing predictive ability. The ultimate goal is to reduce the overall mean-squared error since it is a function of both bias and variance. Nonetheless, we can still interpret the coefficient estimates from the elastic net in the same way that we could with OLS estimates, making this procedure more useful for research where the magnitude and direction of regression effects are of interest. Additional reasons for utilizing the elastic net are as follows (taken directly from Zou and Hastie, 2005):

1. "In the $p > n$ case, the lasso selects at most n variables before it saturates, because of the nature of the convex optimization problem. This seems to be a limiting feature for a variable selection method. Moreover, the lasso is not well defined unless the bound on the L_1-norm of the coefficients is smaller than a certain value."

2. "If there is a group of variables among which the pairwise correlations are very high, then the lasso tends to select only one variable from the group and does not care which one is selected."

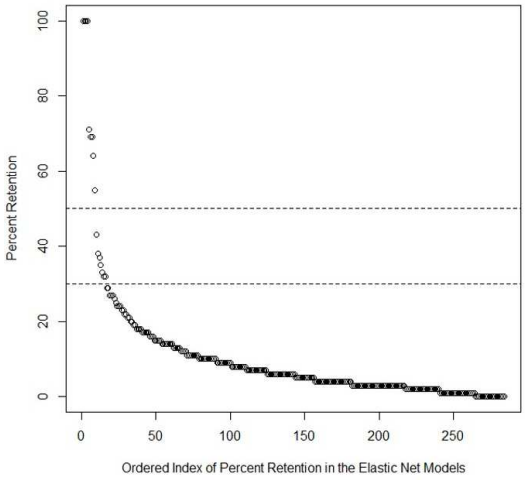
3. "For usual $n > p$ situations, if there are high correlations between predictors, it has been empirically observed that the prediction performance of the lasso is dominated by ridge regression."

It is important to note that the elastic net does not have a closed form solution – that is, calculus cannot be used directly to solve for β in the same way that it can for OLS. Instead, it becomes an optimization problem that depends on the values of two tuning parameters, λ and α, and is solved using a variant of the least-angle regression (LARS) algorithm (Efron et al, 2004). These parameters are typically selected via cross-validation.

The elastic net (and many variable selection techniques) are not able to support missing data while computing the solution path. Multiple imputation (Rubin, 2004; Sterne et al, 2009) is gaining more and more traction as a crucial step to dealing with missing data (Rezvan et al, 2015) and is generally accepted as superior to a complete case analysis when the data are missing at random (White and Carlin, 2010). Multiple imputation by chained equations, or MICE, (Van Buuren, 2000; White et al, 2011) is a flexible approach that can handle mixed types of variables (continuous, binary, unordered categorical and ordered categorical). Hence, we used the MICE procedure with this data to generate 100 imputed data sets.

As one might expect, because multiple imputation gives a range of potential values for missing data, the variable selection results may vary in each imputed data set. For an individual imputed data set, we performed the elastic net using repeated 10-fold cross validation to choose the tuning parameters that give the smallest error, and noted which variables were retained. As expected, the same variables were not retained in each imputed data set; thus, the most realistic goal became to find predictors that provided the *most* evidence of predictive strength, rather than aiming for an exhaustive list. We did this by tracking how often each variable was selected across the 100 data sets. Supplementary Figure 2 displays this information, starting with the most often retained variables.

Supplementary Figure 2 – Ordered Index of Percent Retention in Elastic Net Models



It is clear from this plot that there is a point of diminishing returns – that is, there is definitely a point at which variables are not retained at any kind of meaningful rate – but the challenging question is, when, exactly? A priori we decided that a reasonable threshold would be 50% retention, but we also chose two additional retention thresholds according to the separation of the points in the plot: 100% retention and 30% retention. The benefit to this approach is that it is data driven; the down-side is that the same results may not hold true for new data, limiting generalizability. Nonetheless, this gave us three potential sets of variables to proceed with, summarized below. Note that the 50% retention model also includes those retained 100% of the time, and the 30% model includes all those retained in the 50% retention model.

Table 1 – Elastic Net Screening Results

Variable	**100% Retention**	**≥ 50% Retention**	**≥ 30% Retention**
Baseline HRSD17 score	✓		
Age	✓		
Melancholic depression indicator[1]	✓		
Anhedonia[2]		✓ (71%)	
Anxious arousal[3]		✓ (69%)	
Neuroticism[4]		✓ (69%)	
Physical Abuse[5]		✓ (64%)	
Average theta (defined as 6.5-8hz) current density localized to the rostral anterior cingulate (rACC)		✓ (55%)	
Baseline QIDS Score			✓(43%)
Openness to Experience			✓(38%)
Duration of MDE			✓ (37%)
FMRI Resting Second Block RightVS BR4 Mean			✓(35%)
FMRI Resting First Block LeftVS RightVS Mean			✓(33%)
Flanker Effect All NoOuts ACC			✓(32%)
FMRI Resting First Block LeftInsula RightInsula Mean			✓(32%)

[1] Based on specifier questions on the SCID
[2] Anhedonic depression scale from the Mood and Anxiety Symptom Questionnaire
[3] Anxiety specific scale from the Mood and Anxiety Symptom Questionnaire
[4] Based on 12 neuroticism items from the NEO-FFI
[5] Scale from the CPFQ

After settling on the candidate sets of variables the problem then became estimating the regression coefficients. Outside of the few variables that were selected in 100% of the elastic net models, using simple averages of estimated coefficients across the 100 data sets will introduce additional bias: for example, consider a hypothetical variable that was selected in 75 of the 100 data sets: this means that 25 coefficient estimates are exactly 0 and the other 75 are some non-zero numbers. Simply averaging these 100 values would push the mean towards 0. In reality, the variable either is or is not predictive of the outcome, so either way the inclusion of both zero and non-zero estimates will cause the average estimate to deviate from the truth. This left us with the following needs:

1) To estimate the effects of the regression coefficients (knowing that averaging was not an option).
2) To estimate the variability of the regression coefficients. This is an area of active research in penalized regression (Lockhart et al, 2014; Lee et al, 2016).
3) Derive the above estimates while taking into consideration that we had already used the data to identify three subgroupings of variables. Re-using the data again would increase the risk for overfitting – particularly if we didn't account for the shrinkage that happened when using the elastic net.
4) Resolve needs 1-3 while also incorporating the 100 imputed data sets.

The Bayesian paradigm (McElreath, 2016; Gelman, 2014; Krushke, 2014; for a briefer introduction, see Etz and Vandekerckhove, 2018) offered a solution to these issues: simulated draws from the posterior distributions of the regression parameters derived from Bayesian regression models could be mixed and subsequently sampled from. More on this below.

A Bayesian regression assumes that the outcome has a particular data model – in our case, a normal distribution. The mean of this normal distribution is $\beta_0 + \beta^T \boldsymbol{x_i}$ and the variance is some value, call it σ^2. Additionally, we must assume some prior information about β, β_0, and σ^2: this can either be on the basis of scientific theory, results from previous studies, or can be avoided by assuming "non-informative" information. If we use prior information that is informative, it is in the form of a statistical distribution: in our case (for each element in β), we choose a normal distribution. So, our outcome (depression severity at exit) is said to come from a normal distribution whose mean is $\beta_0 + \beta^T \boldsymbol{x_i}$, and our prior belief about the elements of β is that they each come from a normal distribution. Because they were not of primary interest in this analysis, we used non-informative prior distributions on β_0 and σ^2. To address issue 3) above, each element in β was assigned a normal distribution with mean at 0 and a particular variance structure. Because this prior information said we expected elements from β on average to be 0, shrinkage was applied to the parameter estimates (mimicking that which occurs when using penalized regression) when we combined our prior information with the observed data using Bayes theorem. The result is what is referred to as a joint posterior distribution. Bayes theorem says that the joint posterior distribution of all the parameters we have specified (β, β_0, and σ^2) is proportional to the product of the likelihood of the data (the normal distribution we specified for the outcome variable) and the prior information we assume for our parameters (the normal distributions with mean 0 for each element of β as well as the non-informative priors for β_0 and σ^2). Given below (for those more familiar with Bayesian architecture) is the actual data structure used for each imputed data set:

$$y_i \sim N\left(\delta_i,\ \sigma^2\right)$$

$$\delta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$$

$$\pi(\beta_0) \propto 1$$

$$\pi(\sigma^2) \propto 1/\sigma^2$$

$$\beta_p \sim N(0, \tau^2\sigma^2)$$

$$\tau \sim C^+(0,\ 1)$$

Note that C^+ is a half-cauchy distribution; this distribution is tied to the hyperparameter τ, which is used to create a Bayesian form of ridge regression. The joint posterior distribution that results from this architecture is messy, difficult to derive by hand, and difficult to interpret. However, because it is still a statistical model that is built from both our observed data and prior beliefs about the parameters in the model, we could use this distribution and statistical software to address issues 1-2 and 4 above in the following ways:

- For a given imputed data set, the software was used to isolate each of the parameters of interest (in our case, all elements of β) from the joint posterior distribution and subsequently simulate samples from what is called their full conditional distribution. If we simulated enough samples, the samples could be used as a reasonable proxy for what the true distribution looks like (for example, consider randomly sampling from a normal distribution: if you only drew 10 samples and made a histogram, it may not look like a normal distribution… but if you drew 1,000 samples it probably would). Thus, we could use the mean or median of the sampled data values as our best guess at the true regression parameter.
- Using the same simulated samples above, we constructed a posterior credible interval using the 2.5th and 97.5th percentiles of the posterior samples, which is comparable to a confidence interval. Thus, we have a sense of the variability of the parameters.
- Instead of doing this separately for each imputed data set, we instead drew a number of samples from each and mixed them together, providing a much larger, more comprehensive look at the possible values for each element of β. Then, we provided a single number summary (the median) as the estimate as well as a 95% credible interval. These are the numbers summarized in the first two columns of Table 2.
- Note that we bootstrapped each imputed data set 20 times before fitting the Bayesian model (we were restricted to 20 due to the computational burden) in the hopes of obtaining more generalizable results and avoiding some of the statistical pitfalls associated with using the same data to both choose variables *and* estimate their effects. So, rather than mixing posterior draws from the 100 multiply imputed data sets, we mixed posterior draws from $100 \cdot 20 = 2000$ models.

An additional benefit of shifting to the Bayesian framework is that in addition to having posterior samples for each of the regression coefficients, we were also able to simulate posterior samples for the predicted outcome value for each subject. As we could use these samples to approximate the posterior distributions (that theoretically represent the entire range of potential outcomes, as well as how likely they are), we estimated how likely it was that each person had a predicted outcome that put them in remission or response, whereas this typically is done using three separate models in a more traditional

statistical modeling framework (one multiple linear regression to predict the outcome, one logistic regression to predict remission, and one logistic regression to predict response).

References

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* (Vol. 81). John Wiley & Sons.

White, I. R., & Carlin, J. B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in medicine*, *29*(28), 2920-2931.

Rezvan, P. H., Lee, K. J., & Simpson, J. A. (2015). The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC medical research methodology*, *15*(1), 30.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., ... & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, *338*, b2393.

Van Buuren, S. (2000). *Multivariate imputation by chained equations: MICE V1. 0 user's manual*. Leiden: TNO.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, *30*(4), 377-399.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of statistics*, *32*(2), 407-499.

Lockhart, R., Taylor, J., Tibshirani, R. J., & Tibshirani, R. (2014). A significance test for the lasso. *Annals of statistics*, *42*(2), 413.

Lee, J. D., Sun, D. L., Sun, Y., & Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, *44*(3), 907-927.

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review*, *25*(1), 5-34.

McElreath, R. (2016). Statistical rethinking: A Bayesian course with examples in R and Stan. Boca Raton, FL: Chapman & Hall/CRC.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). Bayesian data analysis (Vol. 2). Boca Raton, FL: CRC press.

Kruschke, J. (2014). Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan. Academic Press.