**Supplementary Appendix**

**Stratification of bipolar patients using polygenic scores increases the accuracy of lithium response prediction: A machine learning approach**

Micah Cearns1, Azmeraw T. Amare1,2, K. Oliver Schubert1,2, Scott Clark1, International Consortium on Lithium Genetics (ConLi+Gen), Bernhard T­. Baune3

1. The discipline of Psychiatry, School of Medicine, University of Adelaide, Adelaide, Australia
2. South Australian Health and Medical Research Institute (SAHMRI), Adelaide, SA, Australia
3. Northern Adelaide Local Health Network, Mental Health Services, Adelaide, Australia
4. Department of Mental Health, University Hospital of Münster, Münster, Germany

**Supplementary methods.**

**Supplementary results and discussion.**

**Table 1.** Missing data table

**Table 2.** Searched hyperparameters for all models

**Table 3.** Extended ridge and elastic net regression results

**Table 4.** Extended random forest regression results

**Table 5.** Extended ridge and elastic net PRS stratified model results

**Table 6.** Extended random forest PRS stratified model results

**Table 7.** Extended classification metrics for the clinical random forest and meta-PRS random forest models

**Table 8.** Missing data table by study site

**Figure 1.** ALDA distributions, stratified meta-PRS distribution, and feature space/R2 relationship

**Figure 2.** Personalised patient stratification and treatment funnel

**Supplementary methods**

**PRS calculation**

PRS were obtained from our previous works published in JAMA Psychiatry1 and Molecular Psychiatry2. For both projects, PRSs were computed at 10 different p-value thresholds (<1 × 10−4, <1 × 10−3, <.01, <.05, <.1, <.20, <.30, <.40, <.50, and <1.0), of which, the best predicting PRS that showed the strongest association with response to lithium was found at p-value threshold < .50 and then used in the current analysis.

This project was conducted through a cross-consortia agreement between the ConLi+Gen and psychiatric genomics consortium (PGC). Any sample overlap between the lithium response dataset and the samples used for SCZ and MDD genotyping were checked before the PRS analysis. No overlapping samples were used in the PRS calculation. In addition, we applied quality control procedures on the genotype data. Samples with low genotype rates < 95%, sex inconsistencies (based on X-chromosome heterozygosity), and one sample from a pair of genetically related individuals was excluded.

**Differences between diagnoses and lithium response**

It is possible that differences between Bipolar 1, 2 and Schizoaffective Bipolar disorders may vary and affect lithium response prediction across cohorts. Therefore, we tested for statistically significant differences using Chi Square tests. No statistically significant differences were found between Bipolar 1 vs Bipolar 2 and Schizoaffective Bipolar (X2 = 1.73, *p* = 0.19) or Bipolar 1 and Schizoaffective Bipolar vs Bipolar 2 (X2 = 3.33, *p* = 0.07).

**Stratified meta-PRS analyses**

For this combined meta-PRS, the MDD and SCZ PRS were first standardised to have a mean of zero and a standard deviation of one and then summed together into a single composite PRS score. These scores were then re-standardised resulting in a meta-PRS representing polygenic loading for MDD and SCZ. Patients with polygenic loadings in the lower and upper quartiles of each PRS distribution were then selected (25th percentile N = 259, 75th percentile N = 259). As per the previous analyses, 33% (N = 171) of patients were partitioned for testing, leaving 347 patients for training and validation. Following this genetic stratification, we re-ran regression analyses using only clinical predictors and then tested the statistical significance of each model using permutation tests (*m* = 10,000). All p-values were FDR corrected using the Benjamini and Hochberg method.

**Assessing sample size effects**

As previous works have shown that smaller samples have a higher risk of systematically misestimating model performance3,4, we devised a Monte-Carlo based statistical test to measure whether the observed changes in performance were attributable to the polygenic stratifications or the subsequent decreases in sample size that resulted from only using patients in the lower and upper quartiles of the PRS distributions. In this procedure, we took 1000 random sub-samples from our full sample (N = 1,034) equal to the N attained in our stratified meta-PRS analysis samples (N = 518) and re-ran the full pipeline on each sub-sample. Following, we used each R2 score to derive a null distribution of scores attainable for each train/validate (N = 347) and test (N = 171) sample combination. This process allowed us to derive a null distribution of scores that were only attributable to decreases in sample size without any PRS stratification. We then took each test score from this distribution that was equal to or greater than the score from our stratified meta-PRS models. We divided the number of occurrences that were greater than or equal to the meta-PRS score by the total number of sub-samples (*m* = 1000) plus 1 (for the meta-PRS model scores), allowing us to derive p-values for the stratified meta-PRS clinical models. Significant p-values represented a difference in scores that were not attributable to changes in sample size.

**Measuring the relationship between the number of predictor variables and R2**

In traditional statistical analysis, every time a predictor variable is added to a model, the model’s R2 increases, therefore, the model R2 needs to be adjusted5. A better approach is to estimate R2 on a held-out test set as was done in the current analyses. However, to further rule out artificial R2 inflation resulting from the increasing number of predictor variables across analyses (as the size of the feature space increases across uni, multi, and interaction term models), we used Pearson’s r to measure the relationship between the number of variables in each model and its subsequent R2 score.

**Clinical data collection procedure**

For the outcome variable (lithium response), the ALDA scale was used across all sites. The minimum data set was the same across all sites following a standardised set of questions / questionnaires and the use of patient medical records. For the use of the ALDA scale, sites were trained and had to undergo training in rating procedures prior to joining the study.

**Supplementary results and discussion**

**Interpreting train, validation, and test set outputs**

In the majority of models, besides the non-linear (random forest) non-stratified PRS and clinical models, the interaction term models perform best in model testing. For the models that did not, it is apparent that the random forest models were unnecessarily complex, leading to overfitting (Interaction PRS train = 0.09 (sd = 0.01), test = 0.0009, interaction clinical train = 0.24 (sd = 0.02), test = 0.0667), interaction clin and PRS train = 0.25 (sd = 0.02), test = 0.052). Here, the models without interaction terms performed best as shown in supplementary table 4 below.

Once genetic stratification was completed (supplementary table 5), it is apparent that higher complexity models become justified as some train/test underfitting arises in these linear models. This is then remedied in supplementary table 6, where the non-linear random forest models demonstrate superior out of sample performance, and the best overall performance across all models.

Regarding train/test underfitting, what is predominantly observed across the supplemental tables is that the validation scores, not the test set scores, are higher than the train scores on some models. This is because we train on all sites in model training, and then validate on only one data collection site when selecting the models hyperparameters, thus, leading to underfitting on some specific sites as the model is being forced to train and learn a multivariate pattern that should generalise to all sites. This becomes evident when then looking at the test set scores (which also included held-out partitions of all sites) and comparing them to the train set scores. It can then be seen that each held out test set score is close to each train score as it is now being tested on the selected set of hyperparameters from validation (leave site-out) across all sites. Thus, any underfitted training scores arise as an artifact of the leave site out cross-validation scheme, that is then remedied when viewing the final test set scores that have used the hyperparameters selected in leave site-out validation.

**Clinical interpretation**

At the current stage, it is unlikely that even the best performing stratified model would be sufficient for clinical use. However, the aim of this work was to lay a foundation for the multimodal use of non-linear transdiagnostic genetic and clinical data, modelled in a stepwise manner, that attempts to follows the natural order of the lithium response system (i.e., genetic polymorphisms *x* environment > central nervous system change > emergent clinical symptoms / deviations in behavioural norms / response profiles).

Moreover, as this work was foundational and based on detecting small variations in underlying genetic profiles and clinical characteristics, we chose to model the full continuous ALDA scale for the main analyses so we could detect small but meaningful effects in genetic and clinical variation that might have otherwise been missed using a clinically derived cut-off of lithium response.

Future works should aim to incorporate a larger range of biomarkers and clinical data related to both bipolar disorder and genetically related conditions. If this proves successful in improving model performance, moving towards a classification framework may be justified, followed by forward tested clinical trials to assess real-world utility.

**Ancestry effects**

It is common practice to control for ancestry effects in PRS through the use of principal components analysis (PCA). This was done in the previous works that were used to derive the polygenic risk scores used in the current study1, 2. In both cases, statistically significant relationships remained between each PRS and lithium response after controlling for these effects. Moreover, these analyses identified two main PCAs related to ancestry, one related to ancestry in the Asian cohorts, and one related to ancestry in the European cohorts. If we included the Asian cohorts and the PCA’s representing the ancestry effects between them and the European cohort and found that these PCA’s were predictive of lithium response, it would be hard to make sense of and interpret them in a predictive study, and this would not be useful in a clinical context. For this reason, we only included the European cohorts in this analysis, and within these European cohorts, used leave site out cross-validation across each European site for training and model selection. However, this does have implications for external validity, with the current work being limited to participants of European descent. Future works should aim to expand the ethnicities of the samples used and create distinct models for participants of similar ancestries.

**Supplementary references**

1. Amare, A.T., Schubert, K.O., Hou, L., Clark, S.R., Papiol, S., Heilbronner, U., Degenhardt, F., Tekola-Ayele, F., Hsu, Y.H., Shekhtman, T. and Adli, M., 2018. Association of polygenic score for schizophrenia and HLA antigen and inflammation genes with response to lithium in bipolar affective disorder: a genome-wide association study. *JAMA psychiatry*, *75*(1), pp.65-74
2. Amare, A.T., Schubert, K.O., Hou, L., Clark, S.R., Papiol, S., Cearns, M., Heilbronner, U., Degenhardt, F., Tekola-Ayele, F., Hsu, Y.H. and Shekhtman, T., 2021. Association of polygenic score for major depression with response to lithium in patients with bipolar disorder. *Molecular psychiatry*, *26*(6), pp.2457-2470
3. Flint, C., Cearns, M., Opel, N., Redlich, R., Mehler, D.M., Emden, D., Winter, N.R., Leenings, R., Eickhoff, S.B., Kircher, T. and Krug, A., 2021. Systematic misestimation of machine learning performance in neuroimaging studies of depression. *Neuropsychopharmacology*, *46*(8), pp.1510-1517.
4. Schnack HG, Kahn RS. Detecting Neuroimaging Biomarkers for Psychiatric Disorders: Sample Size Matters. *Front Psychiatry.* 2016;7:50.
5. Helland IS. On the interpretation and use of R2 in regression analysis. *Biometrics.* 1987:61-69.

**Table 1.** Percentage proportions of missing data stratified by data collection site.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Cagliari****/Sardinian** | **Dresden** | **Geneva** | **Halifax** | **JHU** | **NIMH** | **Poznan** | **Prague****/Czech** | **San Diego****/UCSD** | **Wuerzburg** |
| BP1 vs rest | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| BP1 and SABP vs rest  | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Bipolar age of onset | 0 | 9.3 | 0 | 6.25 | 0 | 5.56 | 2.06 | 0 | 4.63 | 2.2 |
| Depression age of onset | 0 | 9.3 | 2.17 | 25 | 8.33 | 5.56 | 16.49 | 0 | 4.63 | 13.19 |
| Polarity of the 1st episode | 0 | 18.6 | 41.3 | 20.42 | 0 | 5.56 | 34.02 | 2.22 | 1.85 | 12.09 |
| Alcohol dependence | 0 | 4.65 | 0 | 0 | 45.83 | 0 | 17.53 | 0 | 2.31 | 100 |
| Substance dependence  | 0 | 4.65 | 0 | 0 | 54.17 | 0 | 17.53 | 0 | 2.78 | 100 |
| OCD | 0 | 4.65 | 0 | 0 | 91.67 | 0 | 17.53 | 0 | 1.85 | 100 |
| PTSD | 0 | 4.65 | 2.17 | 0 | 100 | 100 | 17.53 | 0 | 1.85 | 100 |
| Panic disorder | 0 | 4.65 | 0 | 0 | 79.17 | 0 | 17.53 | 0 | 1.85 | 100 |
| Any suicidal features? | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Sex | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Age at interview | 0 | 2.33 | 0 | 0 | 0 | 0 | 10.31 | 0 | 0.93 | 100 |
| DSM diagnosis | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| BPD family history | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ALDA scale total score | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Polarity 1st episode (depression onset) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Polarity 1st episode (mania) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Polarity 1st episode (hypomania) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Any suicidal features (No) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Any suicidal features (Yes) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Any suicidal features (Unknown) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DSM Dx (BP1) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DSM Dx (BP2) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DSM Dx (SABP) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DSM Dx (BP3) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DSM Dx (BPNOS) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| BPD family history (Y) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| BPD family history (N) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| BPD family history (Unknown) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**Table 2.** Searched hyperparameter combinations from each model used in analyses.

|  |  |
| --- | --- |
| **Ridge** |  |
| model\_\_alpha | 0.001 | 0.01 | 0.1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| **Elastic net** |  |
| model\_\_alpha | 0.001 | 0.01 | 0.1 | 0.3 | 0.5 | 0.7 | 1 | - | - | - | - | - |
| model\_\_l1\_ratio | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 | 0.1 | - | - |
| **Unimodal random forest** |  |
| model\_\_n\_estimators | 25 | 50 | 75 | 100 | - | - | - | - | - | - | - | - |
| model\_\_max\_features | sqrt | log2 | - | - | - | - | - | - | - | - | - | - |
| model\_\_max\_depth | 4 | 6 | 8 | 10 | - | - | - | - | - | - | - | - |
| model\_\_min\_samples\_leaf | 2 | 5 | 10 | - | - | - | - | - | - | - | - | - |
| **Clinical interaction random forest** |  |
| model\_\_n\_estimators | 25 | 50 | 75 | 100 | - | - | - | - | - | - | - | - |
| model\_\_max\_features | sqrt | log2 | - | - | - | - | - | - | - | - | - | - |
| model\_\_max\_depth | 4 | 6 | 8 | 10 | - | - | - | - | - | - | - | - |
| model\_\_min\_samples\_leaf | 2 | 5 | 10 | - | - | - | - | - | - | - | - | - |
| selector\_\_estimator\_\_n\_estimators | 25 | 50 | 75 | 100 | - | - | - | - | - | - | - | - |
| selector\_\_estimator\_\_max\_features | sqrt | log2 | - | - | - | - | - | - | - | - | - | - |
| selector\_\_estimator\_\_max\_depth | 4 | 6 | 8 | 10 | - | - | - | - | - | - | - | - |
| selector\_\_estimator\_\_min\_samples\_leaf | 2 | 5 | 10 | - | - | - | - | - | - | - | - | - |
| selector\_\_max\_features | 25 | 50 | 75 | 100 | 125 | 150 | 175 | 200 | 225 | 250 | - | - |
| **PRS interaction random forest** |  |
| model\_\_n\_estimators | 25 | 50 | 75 | 100 | - | - | - | - | - | - | - | - |
| model\_\_max\_features | sqrt | log2 | - | - | - | - | - | - | - | - | - | - |
| model\_\_max\_depth | 4 | 6 | 8 | 10 | - | - | - | - | - | - | - | - |
| model\_\_min\_samples\_leaf | 2 | 5 | 10 | - | - | - | - | - | - | - | - | - |
| selector\_\_estimator\_\_n\_estimators | 25 | 50 | 75 | 100 | - | - | - | - | - | - | - | - |
| selector\_\_estimator\_\_max\_features | sqrt | log2 | - | - | - | - | - | - | - | - | - | - |
| selector\_\_estimator\_\_max\_depth | 4 | 6 | 8 | 10 | - | - | - | - | - | - | - | - |
| selector\_\_estimator\_\_min\_samples\_leaf | 2 | 5 | 10 | - | - | - | - | - | - | - | - | - |
| selector\_\_max\_features | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | - | - | - |
| **Multimodal interaction random forest** |  |
| model\_\_n\_estimators | 25 | 50 | 75 | 100 | - | - | - | - | - | - | - | - |
| model\_\_max\_features | sqrt | log2 | - | - | - | - | - | - | - | - | - | - |
| model\_\_max\_depth | 4 | 6 | 8 | 10 | - | - | - | - | - | - | - | - |
| model\_\_min\_samples\_leaf | 2 | 5 | 10 | - | - | - | - | - | - | - | - | - |
| selector\_\_estimator\_\_n\_estimators | 25 | 50 | 75 | 100 | - | - | - | - | - | - | - | - |
| selector\_\_estimator\_\_max\_features | sqrt | log2 | - | - | - | - | - | - | - | - | - | - |
| selector\_\_estimator\_\_max\_depth | 4 | 6 | 8 | 10 | - | - | - | - | - | - | - | - |
| selector\_\_estimator\_\_min\_samples\_leaf | 2 | 5 | 10 | - | - | - | - | - | - | - | - | - |
| selector\_\_max\_features | 62 | 124 | 186 | 248 | 310 | 372 | 434 | 496 | 558 | 620 | - | - |
| **Multimodal random forest** |  |
| model\_\_n\_estimators | 25 | 50 | 75 | 100 | - | - | - | - | - | - | - | - |
| model\_\_max\_features | sqrt | log2 | - | - | - | - | - | - | - | - | - | - |
| model\_\_max\_depth | 4 | 6 | 8 | 10 | - | - | - | - | - | - | - | - |
| model\_\_min\_samples\_leaf | 2 | 5 | 10 | - | - | - | - | - | - | - | - | - |
| selector\_\_estimator\_\_n\_estimators | 25 | 50 | 75 | 100 | - | - | - | - | - | - | - | - |
| selector\_\_estimator\_\_max\_features | sqrt | log2 | - | - | - | - | - | - | - | - | - | - |
| selector\_\_estimator\_\_max\_depth | 4 | 6 | 8 | 10 | - | - | - | - | - | - | - | - |
| selector\_\_estimator\_\_min\_samples\_leaf | 2 | 5 | 10 | - | - | - | - | - | - | - | - | - |
| selector\_\_max\_features | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 | 20 | 22 | 24 |

**Table 3.** Train/validation (N = 692), and test (N = 342) results using linear regression with Ridge regularization. Unimodal, multimodal, and interaction effect models were considered. All p-values were FDR corrected with the Benjamini and Hochberg method.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Train Mean (SD)** | **Validation Mean (SD)** | **Test** | **P-Value** |
| **PRS R2** | **0.01 (0.0)** | **0.2 (0.18)** | **0.012** | **0.0131** |
| PRS RMSE | 3.27 (0.09) | 3.19 (0.45) | 3.365 | - |
| PRS MSE | 10.67 (0.58) | 10.37 (3.04) | 11.324 | - |
| PRS MAE | 2.86 (0.09) | 2.79 (0.44) | 2.969 | - |
| **Interaction PRS R2** | **0.01 (0.0)** | **0.2 (0.18)** | **0.014** | **0.01** |
| Interaction PRS RMSE | 3.26 (0.09) | 3.2 (0.46) | 3.3624 | - |
| Interaction PRS MSE | 10.66 (0.58) | 10.43 (3.08) | 11.306 | - |
| Interaction PRS MAE | 2.86 (0.08) | 2.8 (0.44) | 2.9623 | - |
| **Clin R2** | **0.12 (0.02)** | **0.32 (0.17)** | **0.018** | **0.0004** |
| Clin RMSE | 3.06 (0.07) | 3.33 (0.33) | 3.356 | - |
| Clin MSE | 9.38 (0.4) | 11.22 (2.25) | 11.260 | - |
| Clin MAE | 2.64 (0.06) | 2.85 (0.34) | 2.875 | - |
| **Interaction Clin R2** | **0.12 (0.02)** | **0.26 (0.19)** | **0.045** | **0.0001** |
| Interaction Clin RMSE | 3.07 (0.07) | 3.26 (0.41) | 3.309 | - |
| Interaction Clin MSE | 9.41 (0.4) | 10.83 (2.77) | 10.951 | - |
| Interaction Clin MAE | 2.66 (0.06) | 2.82 (0.41) | 2.867 | - |
| **Clin and PRS R2** | **0.09 (0.02)** | **0.2 (0.17)** | **0.047** | **0.0001** |
| Clin and PRS RMSE | 3.12 (0.07) | 3.2 (0.42) | 3.306 | - |
| Clin and PRS MSE | 9.72 (0.44) | 10.39 (2.78) | 10.927 | - |
| Clin and PRS MAE | 2.72 (0.07) | 2.77 (0.43) | 2.875 | - |
| **Interaction Clin and PRS R2** | **0.13 (0.02)** | **0.27 (0.19)** | **0.051** | **0.0001** |
| Interaction Clin and PRS RMSE | 3.06 (0.07) | 3.27 (0.43) | 3.299 | - |
| Interaction Clin and PRS MSE | 9.34 (0.41) | 10.9 (2.86) | 10.883 | - |
| Interaction Clin and PRS MAE | 2.66 (0.06) | 2.82 (0.43) | 2.856 | - |

**Table 4.** Train/validate (N = 692), and test (N = 342) results using non-linear random forest regression. Unimodal, multimodal, and interaction effect models were considered. All p-values were FDR corrected with the Benjamini and Hochberg method.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Train Mean (SD)** | **Validation Mean (SD)** | **Test** | **P-Value** |
| **PRS R2** | **0.09 (0.01)** | **0.22 (0.18)** | **0.020** | **0.0001** |
| PRS RMSE | 3.13 (0.1) | 3.21 (0.45) | 3.352 | - |
| PRS MSE | 9.81 (0.6) | 10.54 (3.03) | 11.24 | - |
| PRS MAE | 2.74 (0.09) | 2.81 (0.45) | 2.953 | - |
| **Interaction PRS R2** | **0.09 (0.01)** | **0.22 (0.19)** | **0.009** | **0.0001** |
| Interaction PRS RMSE | 3.12 (0.1) | 3.22 (0.46) | 3.371 | - |
| Interaction PRS MSE | 9.75 (0.61) | 10.59 (3.09) | 11.360 | - |
| Interaction PRS MAE | 2.72 (0.09) | 2.81 (0.44) | 2.973 | - |
| **Clin R2** | **0.21 (0.02)** | **0.21 (0.11)** | **0.0812** | **0.0001** |
| Clin RMSE | 2.91 (0.07) | 3.2 (0.35) | 3.245 | - |
| Clin MSE | 8.45 (0.38) | 10.38 (2.33) | 10.531 | - |
| Clin MAE | 2.51 (0.06) | 2.77 (0.38) | 2.801 | - |
| **Interaction Clin R2** | **0.24 (0.02)** | **0.23 (0.14)** | **0.0667** | **0.001** |
| Interaction Clin RMSE | 2.86 (0.06) | 3.23 (0.4) | 3.270 | - |
| Interaction Clin MSE | 8.19 (0.36) | 10.59 (2.63) | 10.695 | - |
| Interaction Clin MAE | 2.47 (0.06) | 2.8 (0.42) | 2.809 | - |
| **Clin and PRS R2** | **0.22 (0.02)** | **0.21 (0.12)** | **0.074** | **0.0001** |
| Clin and PRS RMSE | 2.9 (0.07) | 3.21 (0.37) | 3.258 | - |
| Clin and PRS MSE | 8.41 (0.37) | 10.44 (2.43) | 10.615 | - |
| Clin and PRS MAE | 2.51 (0.06) | 2.77 (0.39) | 2.798 | - |
| **Interaction Clin and PRS R2** | **0.25 (0.02)** | **0.23 (0.16)** | **0.052** | **0.0001** |
| Interaction Clin and PRS RMSE | 2.84 (0.06) | 3.23 (0.42) | 3.296 | - |
| Interaction Clin and PRS MSE | 8.04 (0.33) | 10.61 (2.8) | 10.862 | - |
| Interaction Clin and PRS MAE | 2.45 (0.05) | 2.8 (0.43) | 2.828 | - |

**Table 5.** Train/validate (N = 347), and test (N = 171) results running ML regression analyses on only patients in the lower and upper quartiles of the meta-PRS distribution. Linear models composed of clinical predictors and interaction terms between them were considered. All p-values were FDR corrected with the Benjamini and Hochberg method.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Train Mean (SD)** | **Validation Mean (SD)** | **Test** | **P-Value** |
| **Linear MDD stratified R2** | **0.12 (0.01)** | **0.23 (0.31)** | **-0.028** | **0.0171** |
| Linear MDD stratified RMSE | 3.11 (0.07) | 3.37 (0.34) | 3.309 |  |
| Linear MDD stratified MSE | 9.69 (0.41) | 11.48 (2.27) | 10.947 |  |
| Linear MDD stratified MAE | 2.71 (0.07) | 2.93 (0.37) | 2.903 |  |
| **Linear interaction MDD stratified R2** | **0.13 (0.01)** | **0.21 (0.36)** | **0.027** | **0.0011** |
| Linear interaction MDD stratified RMSE | 3.09 (0.07) | 3.35 (0.45) | 3.22 |  |
| Linear interaction MDD stratified MSE | 9.52 (0.43) | 11.43 (2.87) | 10.368 |  |
| Linear interaction MDD stratified MAE | 2.71 (0.07) | 2.96 (0.43) | 2.81 |  |
| **Linear SCZ stratified R2** | **0.12 (0.01)** | **0.41 (0.42)** | **0.071** | **0.0003** |
| Linear SCZ stratified RMSE | 3.07 (0.07) | 3.42 (0.31) | 3.326 |  |
| Linear SCZ stratified MSE | 9.44 (0.41) | 11.79 (2.1) | 11.059 |  |
| Linear SCZ stratified MAE | 2.63 (0.07) | 2.89 (0.36) | 2.936 |  |
| **Linear interaction SCZ stratified R2** | **0.12 (0.01)** | **0.31 (0.21)** | **0.087** | **0.0001** |
| Linear interaction SCZ stratified RMSE | 3.07 (0.07) | 3.36 (0.47) | 3.293 |  |
| Linear interaction SCZ stratified MSE | 9.43 (0.4) | 11.52 (3.27) | 10.843 |  |
| Linear interaction SCZ stratified MAE | 2.65 (0.07) | 2.87 (0.48) | 2.937 |  |
| **Linear (combined Meta-PRS stratified) R2** | **0.09 (0.01)** | **0.67 (0.39)** | **0.121** | **0.0001** |
| Linear (combined Meta-PRS stratified) RMSE | 3.15 (0.08) | 3.62 (0.38) | 3.075 |  |
| Linear (combined Meta-PRS stratified) MSE | 9.94 (0.5) | 13.25 (2.75) | 9.456 |  |
| Linear (combined Meta-PRS stratified) MAE | 2.73 (0.08) | 3.14 (0.38) | 2.681 |  |
| **Linear interaction (combined Meta-PRS stratified) R2** | **0.1 (0.01)** | **0.51 (0.34)** | **0.092** | **0.0001** |
| Linear interaction (combined Meta-PRS stratified) RMSE | 3.13 (0.08) | 3.46 (0.43) | 3.124 |  |
| Linear interaction (combined Meta-PRS stratified) MSE | 9.79 (0.48) | 12.17 (2.96) | 9.758 |  |
| Linear interaction (combined Meta-PRS stratified) MAE | 2.73 (0.07) | 3.03 (0.41) | 2.724 |  |

**Table 6.** Train/validate (N = 347), and test (N = 171) results running ML regression analyses on only patients in the lower and upper quartiles of the meta-PRS distribution. Non-linear models composed of clinical predictors and interaction terms between them were considered. All p-values were FDR corrected with the Benjamini and Hochberg method.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Train Mean (SD)** | **Validation Mean (SD)** | **Test** | **P-Value** |
| **Random Forest MDD stratified R2** | **0.18 (0.01)** | **0.16 (0.22)** | **0.035** | **0.0001** |
| Random Forest MDD stratified RMSE | 3.01 (0.07) | 3.29 (0.29) | 3.206 |  |
| Random Forest MDD stratified MSE | 9.05 (0.39) | 10.9 (1.99) | 10.281 |  |
| Random Forest MDD stratified MAE | 2.64 (0.07) | 2.9 (0.33) | 2.826 |  |
| **Random Forest interaction MDD stratified R2** | **0.23 (0.01)** | **0.2 (0.29)** | **0.018** | **0.0001** |
| Random Forest interaction MDD stratified RMSE | 2.9 (0.06) | 3.33 (0.31) | 3.234 |  |
| Random Forest interaction MDD stratified MSE | 8.43 (0.36) | 11.18 (2.07) | 10.458 |  |
| Random Forest interaction MDD stratified MAE | 2.54 (0.07) | 2.93 (0.33) | 2.859 |  |
| **Random Forest SCZ stratified R2** | **0.24 (0.01)** | **0.24 (0.14)** | **0.072** | **0.0001** |
| Random Forest SCZ stratified RMSE | 2.86 (0.07) | 3.28 (0.43) | 3.324 |  |
| Random Forest SCZ stratified MSE | 8.16 (0.4) | 10.96 (2.77) | 11.048 |  |
| Random Forest SCZ stratified MAE | 2.46 (0.07) | 2.82 (0.47) | 2.971 |  |
| **Random Forest interaction SCZ stratified R2** | **0.27 (0.02)** | **0.26 (0.14)** | **0.093** | **0.0001** |
| Random Forest interaction SCZ stratified RMSE | 2.8 (0.07) | 3.3 (0.46) | 3.286 |  |
| Random Forest interaction SCZ stratified MSE | 7.84 (0.39) | 11.13 (2.95) | 10.798 |  |
| Random Forest interaction SCZ stratified MAE | 2.4 (0.07) | 2.81 (0.51) | 2.935 |  |
| **Random Forest (combined Meta-PRS stratified) R2** | **0.23 (0.02)** | **0.4 (0.29)** | **0.137** | **0.0001** |
| Random Forest (combined Meta-PRS stratified) RMSE | 2.9 (0.07) | 3.35 (0.44) | 3.047 |  |
| Random Forest (combined Meta-PRS stratified) MSE | 8.42 (0.4) | 11.39 (3.01) | 9.282 |  |
| Random Forest (combined Meta-PRS stratified) MAE | 2.52 (0.07) | 2.92 (0.44) | 2.650 |  |
| **Random Forest interaction (combined Meta-PRS stratified) R2** | **0.7 (0.01)** | **0.39 (0.35)** | **0.045** | **0.0001** |
| Random Forest interaction (combined Meta-PRS stratified) RMSE | 1.81 (0.07) | 3.33 (0.5) | 3.204 |  |
| Random Forest interaction (combined Meta-PRS stratified) MSE | 3.28 (0.24) | 11.31 (3.42) | 10.262 |  |
| Random Forest interaction (combined Meta-PRS stratified) MAE | 1.51 (0.06) | 2.86 (0.46) | 2.760 |  |

**Table 7.** Clinical predictor variables from the best performing meta-PRS stratified model and the Gini feature importance coefficients for each predictor.

|  |  |
| --- | --- |
| **Predictor variable** | **Gini-importance** |
| Age at interview | 0.2238 |
| Age of onset bipolar | 0.2030 |
| Any suicidal features | 0.1486 |
| Age of onset depression | 0.1013 |
| OCD | 0.0481 |
| Polarity 1st episode | 0.0401 |
| Alcohol dependence | 0.0368 |
| Panic disorder | 0.0335 |
| Substance dependence  | 0.0331 |
| Bipolar family history | 0.0320 |
| PTSD | 0.0316 |
| Bipolar 1 and schizo-affective bipolar vs rest | 0.0242 |
| Sex | 0.0226 |
| Bipolar 1 vs rest | 0.0128 |
| DSM diagnosis | 0.0084 |

**Table 8.** Extended classification metrics for the clinical random forest and meta-PRS random forest models. AUC = Area under the receiver operator characteristic, ACC = Accuracy, BAC = Balanced accuracy, Sens = Sensitivity, Spec = Specificity, PPV = Positive predictive value, NPV = Negative predictive value.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **AUC** | **ACC** | **BAC** | **F1** | **Sens** | **Spec** | **PPV** | **NPV** |
| Clinical RF model | 65.63 | 63.16 | 58.95 | 63.24 | 46.09 | 71.81 | 45.30 | 72.44 |
| Meta-PRS clinical RF model  | 69.88 | 69.59 | 63.65 | 68.93 | 46.43 | 80.87 | 54.17 | 75.61 |

**Table 9.** Whilst we dropped variables with more than 20% missing data overall, the Wuerzburg and JHU sites had disproportionately more missing data within them across certain variables (see Table 1 above). Therefore, we ran a sensitivity analysis by excluding these two sites and re-running the best performing model (Random Forest Meta-PRS stratified). As can be seen, the performance is close to the original model that included both sites and remains one of the best performing models even after their exclusion.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Train mean (SD)** | **Validate mean (SD)** | **Test** | **P-Value** |
| **R2** | 0.2 (0.02) | 0.31 (0.26) | 0.12 | 0.0001 |
| **RMSE** | 2.93 (0.09) | 3.22 (0.38) | 3.06 | 0.0001 |
| **MSE** | 8.57 (0.54) | 10.54 (2.54) | 9.38 | 0.0001 |
| **MAE** | 2.54 (0.09) | 2.84 (0.43) | 2.65 | 0.0001 |

**Figure 1.** Top left: Meta-PRS loading distribution stratified into quartiles. Top right: Scatterplot and Pearson’s r correlation between the number of predictor variables (x-axis) and test set R2 scores (y-axis) across all analyses. Bottom left: ALDA total score distribution used as the regression target in analyses. Bottom right: ALDA total score distribution stratified by combined meta-PRS load for patients in quartiles 1 and 4 of the distribution.

****

**Figure 2.** Schematic representation of the stepwise procedure to parse the transdiagnostic genetic heterogeneity of bipolar disorder using meta-PRS followed by the prediction of lithium response with clinical data. Level 2 of the figure (Further biomarker stratification) introduces the idea of using additional biomarkers (e.g., sMRI and fMRI) to further parse patient heterogeneity. In theory, this step would be secondary to genetic stratification as it would most closely follow the biological order of the system (i.e., genetic polymorphisms *x* environment > central nervous system abnormalities > emergent clinical symptoms / deviations in behavioural norms).

****