Development and validation of genomic biotypes for schizophrenia susceptibility from multiple polygenic scores

Ersoy Kocak[*1,2], Joonas Naamanka[*1,2,3], Tobias Gradinger[*1,2], Fiona Klaassen[1,2], Johannes Nitsche[1,2], Philipp Grotehusmann[1,2], Kristina Adorjan[4,5], Linda A Antonucci[6], Giuseppe Blasi[6,7], Monika Budde[4], Piergiuseppe Di Palo[6,7], Maria Heilbronner[4], Gianluca C Kikidis[6], Alba Navarro-Flores[4,8], Mojtaba Oraki Kohshour[4,9], Sergi Papiol[4,10,11], Alessandra Raio[6], Antonio Rampino[6,7], Daniela Reich-Erkelenz[4,10], Eva C. Schulte[4,10,12,13,14], Fanny Senner[4,10,15], Leonardo Sportelli[16], FinnGen[17], Alessandro Bertolino[6,7], Peter Falkai[10,11,14], Urs Heilbronner[4], Giulio Pergola[6,16,18], Thomas G. Schulze[4,14,18,19], Andreas Meyer-Lindenberg[1], Fabian Streit[#1,2,20], Emanuel Schwarz[#1,2,20]

1, Department of Psychiatry and Psychotherapy, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany
2, Hector Institute for Artificial Intelligence in Psychiatry, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany
3, SleepWell Research Program, Faculty of Medicine, University of Helsinki, Helsinki, Finland
4, Institute of Psychiatric Phenomics and Genomics (IPPG), LMU University Hospital, Munich, Germany
5, University Hospital of Psychiatry and Psychotherapy, University of Bern, Bern, Switzerland
6, Department of Translational Biomedicine and Neuroscience, University Of Bari Aldo Moro, Bari, Italy
7, U.O.C. Psichiatria Universitaria, Azienda Ospedaliero-Universitaria Consorziale Policlinico, Bari, Italy
8, International Max Planck Research School for Translational Psychiatry (IMPRS-TP), Max-Planck-Institute of Psychiatry, Munich, Germany
9, Department of Immunology, Faculty of Medicine, Ahvaz, Iran
10, Department of Psychiatry and Psychotherapy, LMU University Hospital, Munich, Germany
11, Department Clinical Translation, Max Planck Institute of Psychiatry, Munich, Germany, Munich, Germany
12, Department of Psychiatry and Psychotherapy, Faculty of Medicine and University Hospital Bonn, Bonn, Germany
13, Institute of Human Genetics, Faculty of Medicine and University Hospital Bonn, Bonn, Germany
14, German Center for Mental Health (DZPG), partner site Munich/Augsburg, Germany
15, Centres for Psychiatry Suedwuerttemberg, Centres for Psychiatry Suedwuerttemberg, Ravensburg, Germany
16, Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore, MD, USA
17, The full list of FinnGen members can be found as a supplementary table
18, Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, USA
19, Department of Psychiatry and Behavioral Sciences, SUNY Upstate Medical University, NY, USA
20, German Center for Mental Health (DZPG), Partner Site Mannheim - Heidelberg - Ulm, Germany

* Contributed equally
# Co-senior authors

Corresponding author
Prof. Dr. Emanuel Schwarz
Phone: +4962117032390
Email: emanuel.schwarz@zi-mannheim.de

Abstract

Understanding the genetic architecture of schizophrenia (SCZ) is invaluable for the development of personalized treatment. In three independent cohorts, using an automated pipeline, we calculated 413 psychiatry-relevant polygenic scores. Using these scores, machine learning was applied to stratify SCZ patients into two biotypes in FinnGen (n_{SCZ}=7486), and to validate these results in the PsyCourse Study (n_{SCZ}=421) and Bari (n_{SCZ}=531). While the two biotypes showed comparable polygenic SCZ risk, they were primarily distinguished by a greater predisposition for neuroticism, depression-related traits, and lower cognitive performance in Biotype 1. The genetic prediction of Biotype 1 was phenotypically characterized by an increased prevalence of SCZ and a more severe and complex clinical manifestation. This illustrates that the penetrance of genetic SCZ risk might partially depend on the predisposition for the aforementioned traits through pleiotropic variants. Our results provide novel, replicable insights into the genetic architecture of SCZ and might inform future personalized treatments.

Introduction

Schizophrenia (SCZ) is a severe and complex mental disorder with a lifetime prevalence estimate of up to 1%, an incidence peak in early adulthood, and an often chronic course[1]. The severity of the disease is evident in many epidemiological metrics, e.g., an approximately 15-year reduction in life expectancy[2] and the fact that only 10-20% of affected individuals are in competitive employment[1,3,4].

The phenotypic complexity of the disease is reflected in its manifestations with multiple domains like perception (e.g., hallucinations), affect (e.g., flattening of affect), cognition (e.g., cognitive deficits), behavior (e.g., loss of drive), speech (e.g., schizophasia), motor skills (e.g., catatonia), and autonomic regulation (e.g., pernicious catatonia) being affected to varying degrees among patients. Although diagnostic systems have long specified a phenotypic subclassification of SCZ, these classifications have been dropped from ICD-11 and DSM-V for lack of reliability and clinical utility[5]. Furthermore, the complexity and heterogeneity of SCZ is reflected in the treatment response: approximately one third of the patients do not respond to standard treatment[6,7], and especially a subgroup of symptoms, commonly referred to as negative symptoms (e.g. flattening of affect and avolition[8,9]) often persist..

The biological complexity of the disease warrants strongly multimodal and multi-level approaches[1]. Evidence suggests that SCZ encompasses a multitude of overlapping neurobiological processes[10]. SCZ cases that are similar regarding those processes can be considered to belong to a biotype, i.e., a biologically more homogeneous subgroup within the disorder. Identifying those biotypes is an approach towards personalized medicine and a more refined understanding of the disease etiology[11]. Across a number of family and twin studies, heritability for SCZ has been estimated at 60%-80% overall[12–14], among the highest for psychiatric disorders, and shows the strongest positive genetic correlation with bipolar disorder (BIP, genetic correlation (rg) ~ .7) , but also with other mental disorders such as major depression (MDD; rg ~ .35)[14,15]. Additionally the phenotypic and biological heterogeneity of SCZ has been hypothesized to depend in part on genetic heterogeneity[16–18]. Therefore, the investigation of genetic variation is a promising base for identifying its possible biotypes[1,19,20]. Genome-wide association studies (GWAS) of SCZ indicate it is a highly polygenic disorder with thousands of common risk variants, i.e. primarily single nucleotide polymorphisms (SNPs), each exerting a small effect on overall disorder risk[15,21]. To estimate the genetic predisposition of an individual to a disease or a trait, the effects conferred by individual SNPs can be summed up into polygenic scores (PGS), also often referred to polygenic risk scores (PRS) in the case of diseases and adverse traits. These indices of polygenic predisposition

show robust predictive performance in independent samples but they fall far short of explaining disease liability to the degree suggested by heritability estimates, including the ones accounting for common variation only[19,22].

Therefore, considered together with the high genetic intercorrelations and overlapping phenotypic constructs in psychiatry[22], an approach based on a broad spectrum of PGS seems promising. Furthermore, SCZ PGS are primarily designed to distinguish cases from controls which makes a multi-PGS strategy better suited for discriminating cases from other cases. While the PGS of SCZ, combined with additional psychiatric PGS with known comorbidity with SCZ, have been used as the basis for identifying biotypes[23], recent advances in machine learning and data harmonization have been demonstrated to allow for a data-driven multi-PGS approach, resulting in improved predictive performance compared to single-PGS models[24].

Building on this recent research we aimed to: i.) identify genetic risk biotypes of SCZ by integrating large numbers of PGS of psychiatry-relevant phenotypes using robust, reproducible machine learning methods; ii.) characterize these biotypes as patient subgroups that demonstrate specific patterns of polygenic risk contributions from a potentially large number of phenotypes iii.) characterization of these biotypes in terms of clinical manifestations.

To achieve this, we developed *PRS*cope, a new automated framework for biotype identification and validation using multi-PGS data. Our analysis included data from a total of 8,442 patients with SCZ, and 28,327 healthy controls (HC) from three independent datasets, to identify and validate biotypes. In addition we used 7,555 patients with BIP, 27,187 patients with MDD, and 41,998 HCs for a transdiagnostic analysis.Our method offers a novel approach to deconstructing the genetic heterogeneity of complex human disorders and may contribute to the personalization of medical care in SCZ and other mental disorders.

Results

The PRScope procedure comprises three overarching steps: I) the retrieval of information from publicly available “source” GWAS, to compute quality-controlled PGS; II) the training of machine learning models using PGS derived from different numbers of source GWAS that are associated with SCZ; and III) the identification and validation of “biotypes” of SCZ that relate to subgroups of patients that share a similar composition of PGS signals. The number of source GWAS used for machine learning is denoted as *feature size* (though the actual number of features may be higher since PGS were calculated for multiple significance thresholds).

Identification and validation of genetic biotypes of SCZ

PRScope retrieves GWAS summary statistics based on the selected Experimental Factor Ontologies (EFO) categories from the EFO browser (www.ebi.ac.uk/efo) and performs filtering and quality control. We applied it to psychiatry-related categories (selected EFO categories can be found in Supplementary Table 1, for more details see section Methods) which yielded 413 source GWAS for the subsequent PGS calculation in the target datasets ($Train_{FinnGen}$, $Val_{PsyCourse,}$ and Val_{Bari}).

We applied this procedure to data from the FinnGen cohort[25] ($Train_{FinnGen}$; 7486 patients with SCZ and 27288 HC, for more demographic details, see Supplementary Table 2-4) and calculated PGS at 10 significance thresholds (from $5x10^{-8}$ to 1) for 413 source GWAS applying a commonly used simple method, clumping and thresholding[26]. Bayesian approaches for calculating PGS[27–29] may outperform this method, particularly when comparing the single PGS derived from the Bayesian methods with that obtained using the best p-value threshold. However, our approach can offer the advantage of leveraging information from all SNP layers defined by different significance thresholds, combining all PGS derived from a GWAS within a multi-PGS framework. To identify biotypes using the multi-PGS data, we first extracted the features in the training folds of 10-fold cross validation. To do this we determined the best p-value threshold (i.e strongest correlation with SCZ) for PGS derived from each source GWAS and ranked the source GWAS based on the correlations of these best-threshold PGS with SCZ. The PGS of the top 45 source GWAS were selected for further analyses (note that PGS with all p-value thresholds were kept even though the ranking was based on the best ones only, see Methods for details). The following steps were applied for each feature size from 1 to 45 separately in the training folds. First, the SCZ patients were classified into two biotypes via k-means clustering. Second, a random forest (RF) model was trained for each of the two biotypes predicting SCZ vs HC. This approach was used to validate that the identified biotypes reflect disease-specific PGS patterns. Third, SCZ vs HC predictions were made in the test folds with each biotype model, resulting in one prediction for Biotype 1 model and one prediction for Biotype 2 model per feature size. Finally, every SCZ patient being uniquely correctly predicted by one of the biotype models was then assigned to this biotype (for more details see section Methods and for a schematic overview of the procedure, see Figure 1). This cross-validation procedure showed that once at least five of the most relevant features on top of the SCZ PGS were added, the predictions of both Biotype 1 and 2 models were positively associated with the true disease status ($p < 0.05$ after Benjamin-Hochberg (BH) correction; Figure 2a, for uncorrected p-values see Supplementary Table 5). This illustrates that using such numbers of PGS the biotypes could be successfully predicted in independent test data. Next, we evaluated the stability of the biotypes with respect to the selected feature

size. With an increasing feature size the nature of the biotypes changed (Figure 2b), and we observed a substantial reallocation of patients, which started leveling off with approximately 12 PGS (Figure 3).

Thus, we decided to predict the biotypes at feature sizes that represent the points of a high rate of reallocation (see Methods for details) into two validation datasets ($Val_{PsyCourse}$[30], n=421 patients with SCZ and n=300 controls, Supplementary Table 3; Val_{Bari}, n=535 patients with SCZ and n=739 controls, Supplementary Table 4). Similarly to $Train_{FinnGen}$ cross-validation, in both validation datasets ($Val_{PsyCourse}$ and Val_{Bari}), we observed that both Biotype 1 and 2 were significantly positively associated with SCZ from feature size 6 onwards ($p < 0.05$ after BH correction, see for uncorrected p-values and the biotype proportions in patients and HC Supplementary Table 6-10), although for $Val_{PsyCourse}$ the p-value exceeded 0.05 (shown as a dip due to the negative log transformation) for the Biotype 2 model at feature size 24 (Figure 2a). The biotypes did not significantly differ concerning age, sex, or population stratification principal components (PCs) in any of the three datasets.

The top SCZ associated PGS fall into three groups

As many of the top 45 selected features were conceptually similar (e.g., different neuroticism items, see Figure 3) we set out to investigate if they could be grouped empirically. To achieve this we calculated all pairwise genetic correlations for those 45 GWAS using LDSC software[31,32] and applied hierarchical clustering to the resulting correlation matrix (using absolute values). Based on a visual inspection of the dendrogram (Supplementary Figure 1) we assessed a four-cluster solution to be the most interpretable one. The clusters, for the remainder of the article referred to as *PGS groups*, can be characterized as relating to cognition, depression and neuroticism. The final group includes SCZ itself, a gene-environment interaction of neuroticism and two alcohol traits. As they had relatively low genetic intercorrelations (Supplementary Figure 2) and were conceptually not easy to reconcile, we decided not to treat them as a group of their own.

SCZ risk penetrance coincides with the genetic predisposition to neuroticism, depression-related traits, and low cognitive performance.

Figure 2 shows that the PGS driving biotype composition besides SCZ-PGS, were the ones relating to neuroticism, depression-related traits, and cognition. It highlights that while both biotypes featured comparable risk for SCZ towards the higher feature sizes, they substantially diverged concerning the other PGS. Specifically, Biotype 1 was characterized by a markedly elevated neuroticism PGS group and a relatively low cognition PGS group, whereas Biotype

2 displayed an opposite pattern for neuroticism and intermediate (between Biotype 1 and HC) levels for the cognition PGS group.

The contribution of neuroticism to SCZ risk is mediated by shared genetic variation

Since SCZ has well-established genetic correlations with neuroticism, depression, and cognition[21,33], and the better-performing Biotype 1 model is characterized by these dimensions, we investigated whether these aspects were collectively reflected as disproportional risk contributions of pleiotropic SNPs.

For this, we constructed a subset of PGS of SNPs that were associated with SCZ and each of the PGS groups. We utilized the conjunctional false discovery rate analysis (conjFDR)[34] with an FDR of 0.1 (see Methods for details). All three of these subset PGS predicted SCZ better than PGS constructed of p-value matched (as indicated in the SCZ GWAS summary statistics) SNPs outside their respective pleiotropy SNP group (Figure 4). Supplementary Table 11-13 show the pleiotropic variants as identified by conjFDR.

Genetic biotypes have transdiagnostic relevance and differ in clinical measures

To support the clinical validity of the identified biotypes, differences in clinical features were assessed across the investigated datasets. These analyses were, in addition to conventional covariates, corrected for the SCZ-PGS, to capture biotype effects independent of genetic predisposition for SCZ (see section *Phenotypic profiling* in the Methods). All phenotypic results are shown for feature size 24 only for simplicity of illustration (for other feature sizes see Supplementary Figure 3-18). In $Train_{FinnGen}$, Biotype 1 was associated with a more severe and complex clinical manifestation in patients with SCZ as indicated by higher rates of clozapine and antidepressant prescriptions and comorbid psychiatric diagnoses; including depression, neurotic, personality, and eating disorders (Figure 5). Both in $Val_{PsyCourse}$ and Val_{Bari} the statistical power of the phenotypic analysis was substantially lower than in $Train_{FinnGen}$ due to a far smaller sample size. Still in $Val_{PsyCourse}$, the higher rate of clozapine prescription in Biotype 1 could be replicated, while the higher rate of antidepressant prescription could not, though it was showed as a trend (Figure 5). Both in $Val_{PsyCourse}$ and Val_{Bari} the Positive and Negative Syndrome Scale (PANSS) showed no significant differences in any of the scales (Figure 5). In $Val_{PsyCourse}$ we further investigated a collection of phenotypic variables representing the major risk components of the biotypes as represented by the PGS groups defining them. We found a significantly (FDR corrected p-values < 0.05) higher score for Biotype 1 in the Inventory of depressive symptomatology ($IDS\text{-}C_{30}$) (Figure 5). In $Train_{FinnGen}$, Biotype 1 model predictions were also associated with BIP (n cases = 7,555; n controls = 11,198) and MDD (n

cases = 27,187; n controls = 26,993) in a gradient-dependent manner, with the highest OR observed for SCZ, intermediate for BIP, and the lowest for MDD. The corresponding associations for Biotype 2 model predictions followed the same gradient-dependent pattern, but the magnitude of the association was lower, for MDD even in the other direction (OR < 1, p = 0.018) (Figure 6).

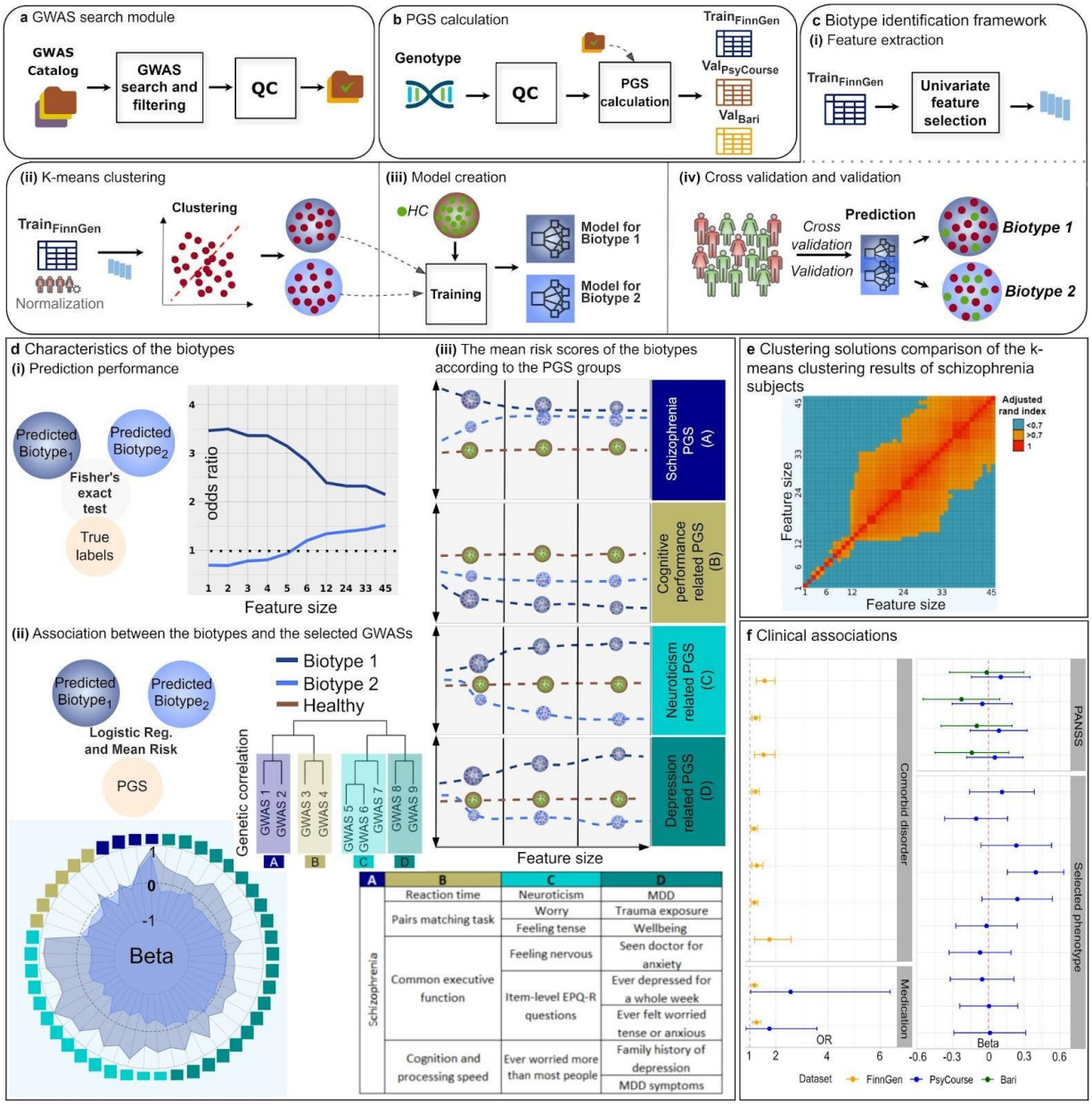
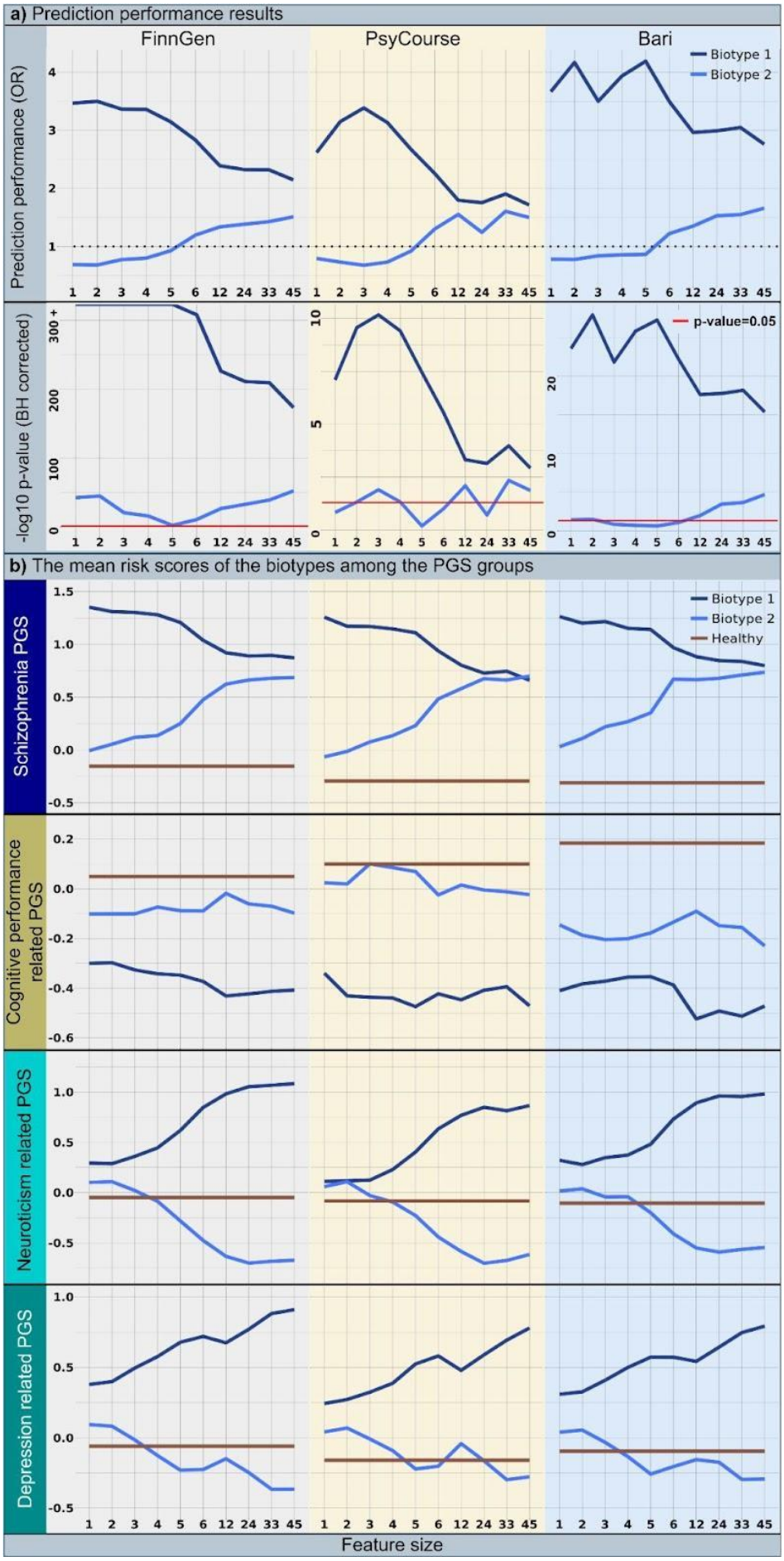


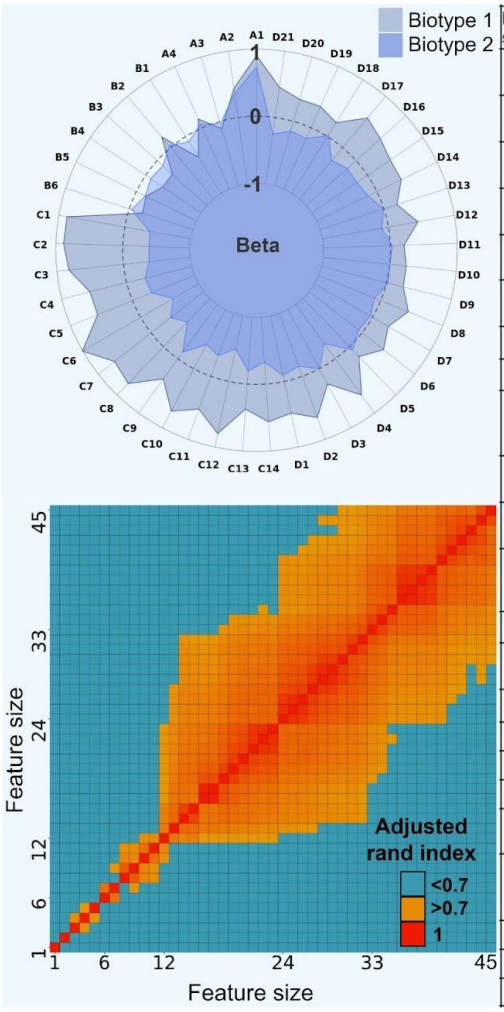
Figure 1. Overview of the PRScope framework and analysis pipeline. a) Psychiatry-related GWASs, annotated with relevant ontologies and fulfilling our filtering criteria, were downloaded from the GWAS Catalog and pre-processed for the subsequent PGS calculation. **b)** Overlapping SNPs across the three datasets were pre-processed and PGS calculations for 413 pre-processed GWAS summary statistics were performed for each GWAS summary statistic using up to 10 different significance thresholds. **c)** Machine learning analysis steps were used to identify two biotypes with differing PGS profiles in a large-scale training dataset, followed by validation in two independent datasets. (i) Selection of GWASs with the strongest association to diagnosis to reduce the dimension to improve the clustering performance and the predictive power of the ML models. To prevent data leakage during cross-validation in the training dataset, feature selection was performed separately for each cross-validation

iteration. For validation, features were extracted using the entire training dataset. (ii) Biotyping was based on different numbers of features (from top 1 to top 45) using k-means clustering. The PGS of selected GWASs for SCZ patients were adjusted by subtracting the median and dividing by the standard deviation of the HC. (iii) Random forest machine learning was applied to characterize the multi-PGS signature of each biotype. (iv) Biotype memberships were predicted by the model in independent data and cross-validation test folds. **d)** The characterization of the predicted biotypes as subgroups of patients that demonstrate specific patterns of polygenic risk contributions from a potentially large number of phenotypes. (i) Associations between the predicted biotype memberships and diagnosis were quantified as the odds ratio and the significance was determined using two-sided Fisher's exact test. (ii) The beta values derived from the logistic regression represent the association between the PGS of selected GWAS summary statistics and biotype membership. The table used as a legend shows the GWASs grouped together based on their genetic correlation. (iii) PCs were calculated for each group separately (Biotype 1, Biotype 2, and HC) using all PGS of a GWAS group, and the mean values of the first PC for each group across different feature sizes were plotted. **e)** The k-means clustering solution for a given feature size was compared with the clustering results of other feature sizes and visualized as a heatmap. **f)** To support the clinical validity of the identified biotypes, differences in clinical features were assessed across the investigated datasets.


a) Prediction performance results
FinnGen
PsyCourse
Bari
Biotype 1
Biotype 2
Prediction performance (OR)
-log10 p-value (BH corrected)
p-value=0.05
b) The mean risk scores of the biotypes among the PGS groups
Biotype 1
Biotype 2
Healthy
Schizophrenia PGS
Cognitive performance related PGS
Neuroticism related PGS
Depression related PGS
1 2 3 4 5 6 12 24 33 45
Feature size

275

Figure 2. Results of biotype validation. a) The odds ratios (OR) and the significance (adjusted p-values) across the three datasets were determined using two-sided Fisher's exact test where the input was a cross table of either Biotype 1 or Biotype 2 model and SCZ. The (BH)-adjusted p-values were -log10 transformed to increase the readability of the plot. The raw p-values can be found in Supplementary Tables (5-7) and the biotype proportions in patients and HC can be found in Supplementary Tables (8-10). Feature sizes 1 to 45 (x-axis) were plotted so that the development of the profiles of the clusters could be demonstrated when more and more features were added to the clustering/prediction procedure. An OR larger than 1 (dashed line) means that a prediction of having SCZ by a given model increases the odds of truly having SCZ. The red lines on the p-value plots denote the BH-adjusted p-value of the significance threshold (0.05). **b)** The PGS compositions of the biotypes and HC were visualized across the three PGS groups (neuroticism, depression, and cognition-related PGS groups) and SCZ PGS. The mean of the first PC (to ensure comparability across units of analysis, all measures were scaled) of each PGS group was plotted across the feature sizes from 1 to 45 (x-axis) for Biotype 1, Biotype 2, and HC.



Feature size	GWAS	Phenotype	Feature size	GWAS	Phenotype
1	A1	Schizophrenia (PGC-2022, Bari excluded)	24	C8	Endorsement
2	A2	Gene-environment interactions in neuroticism	25	D10	Illnesses of siblings: Severe depression
3	D1	Exploring the genetic overlap between 12 psychiatric disorders	26	D11	Illnesses of mother: Severe depression
4	D2	Seen psychiatrist for nerves, anxiety, tension or depression	27	C9	Ever worried more than most people would in similar situation
5	C1	Would you call yourself tense or highly strung?	28	D12	Ever unenthusiastic/disinterested for a whole week
6	C2	Tense / highly strung	29	A3	Reason for reducing alcohol: Other reason
7	B1	Common executive function	30	D13	Ever depressed for a whole week
8	C3	Feeling tense	31	C10	Do you suffer from nerves?
9	D3	MDD symptoms depressed (MDD-2024)	32	D14	Endorsed measures of depression (two) vs mental health questionnaire controls
10	C4	Would you call yourself a nervous person?	33	D15	Depression (broad)
11	B2	Pairs matching task	34	D16	Ever suffered mental distress preventing usual activities
12	C5	Nervous feelings	35	D17	Endorsed measures of depression (one) vs mental health questionnaire controls
13	C6	Worry	36	C11	Neuroticism
14	D4	Seen doctor for nerves, anxiety, tension or depression	37	A4	Reason for reducing alcohol: Health precaution
15	B3	Cognition and processing speed	38	B6	Common executive function
16	C7	Feeling nervous	39	D18	Major depression (MDD 2018)
17	D5	Trauma	40	C12	Neuroticism (Big five 2024)
18	D6	Depression/anxiety	41	D19	Depression
19	D7	MDD symptoms anhedonia (MDD-2024)	42	C13	Guilty feelings
20	D8	Ever felt worried, tense or anxious for most of a month or longer	43	D20	Ever sought or received professional help for mental distress
21	B4	Common executive function	44	C14	Worrier / anxious feelings
22	D9	Trauma exposure	45	D21	Are you a worrier?
23	B5	Reaction time			

Figure 3. PGS composition of the biotypes and stability across the feature sizes. The polar plot was generated using the FinnGen cohort's association results for the features size 45. To facilitate comparison across PGS groups, the PGS of a PGS group were successively plotted. The table in the figure shows the ranked order of the selected GWASs that showed the strongest correlation with the diagnosis. The associations between each GWAS summary statistic from the top 45 GWASs and biotype membership predicted by the model trained on these features were estimated using logistic regression. ARI heatmap was generated using the FinnGen cohort's k-means clustering results for the feature sizes from 1 to 45. ARI was calculated directly using the whole sample prediction results. To facilitate comparison across features, 0.7 is selected as the threshold for high similarity. Detailed information about the GWASs can be found in Supplementary Table 14 and the values used to create the polar plot can be found in Supplementary Table 15. The ARI values used to generate the heatmap can be found in Supplementary Table 16.

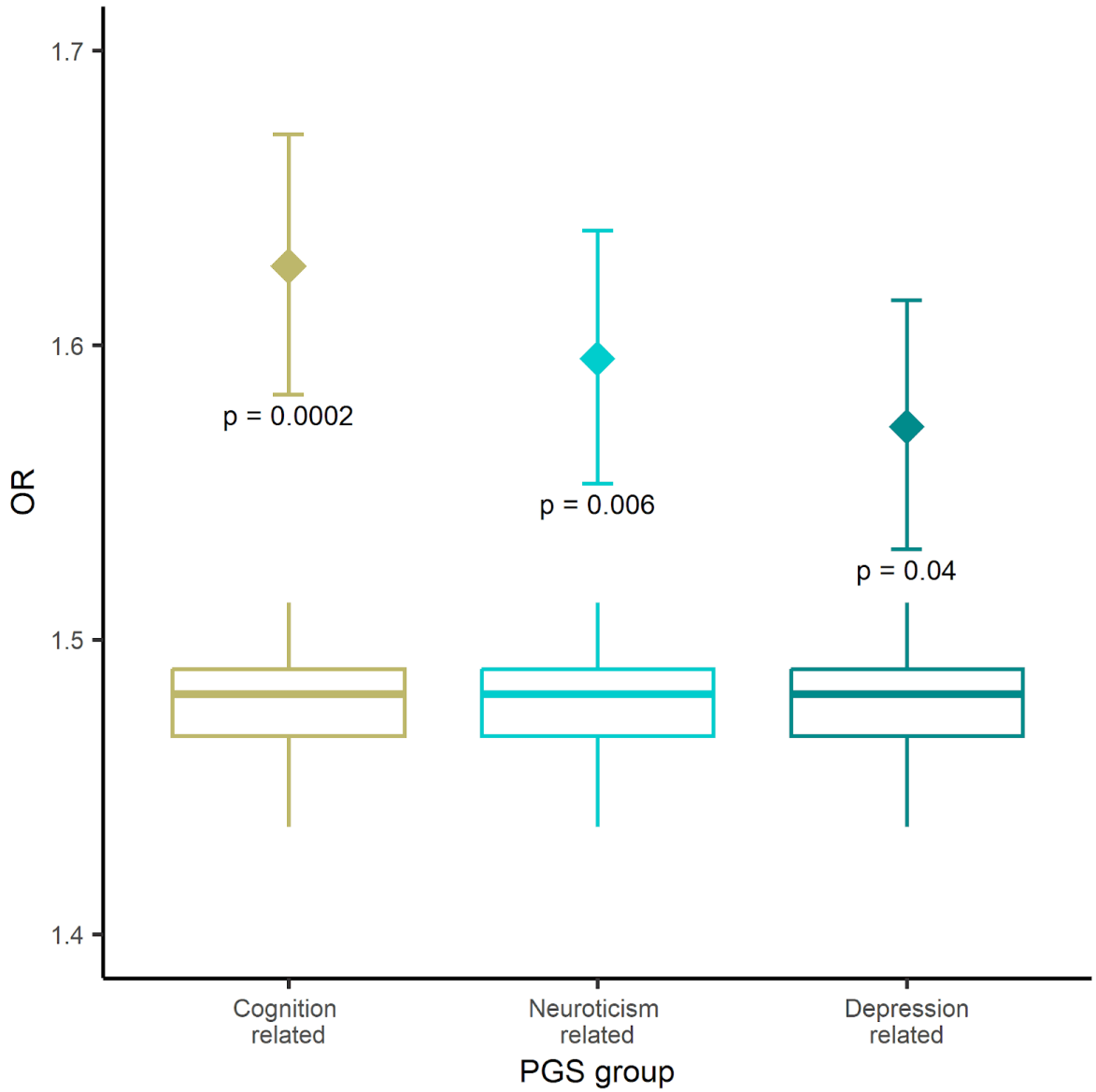


Figure 4. Predictive performance of the conjFDR-based PGS in Train$_{FinnGen}$. The diamonds indicate the odds ratios for a PGS constructed with conjunctional FDR with a given GWAS group (see Methods) in a logistic regression where true SCZ status was the dependent variable. The boxplots represent estimates from similar models with different PGS serving as predictors. Each PGS was constructed on a p-value-matched random sample of SCZ risk SNPs. The p-values are derived from Cochran's Q test of heterogeneity of effects, where the maximum estimate and the corresponding standard error from the 100 respective comparative PGS are contrasted with the conjFDR PGS.

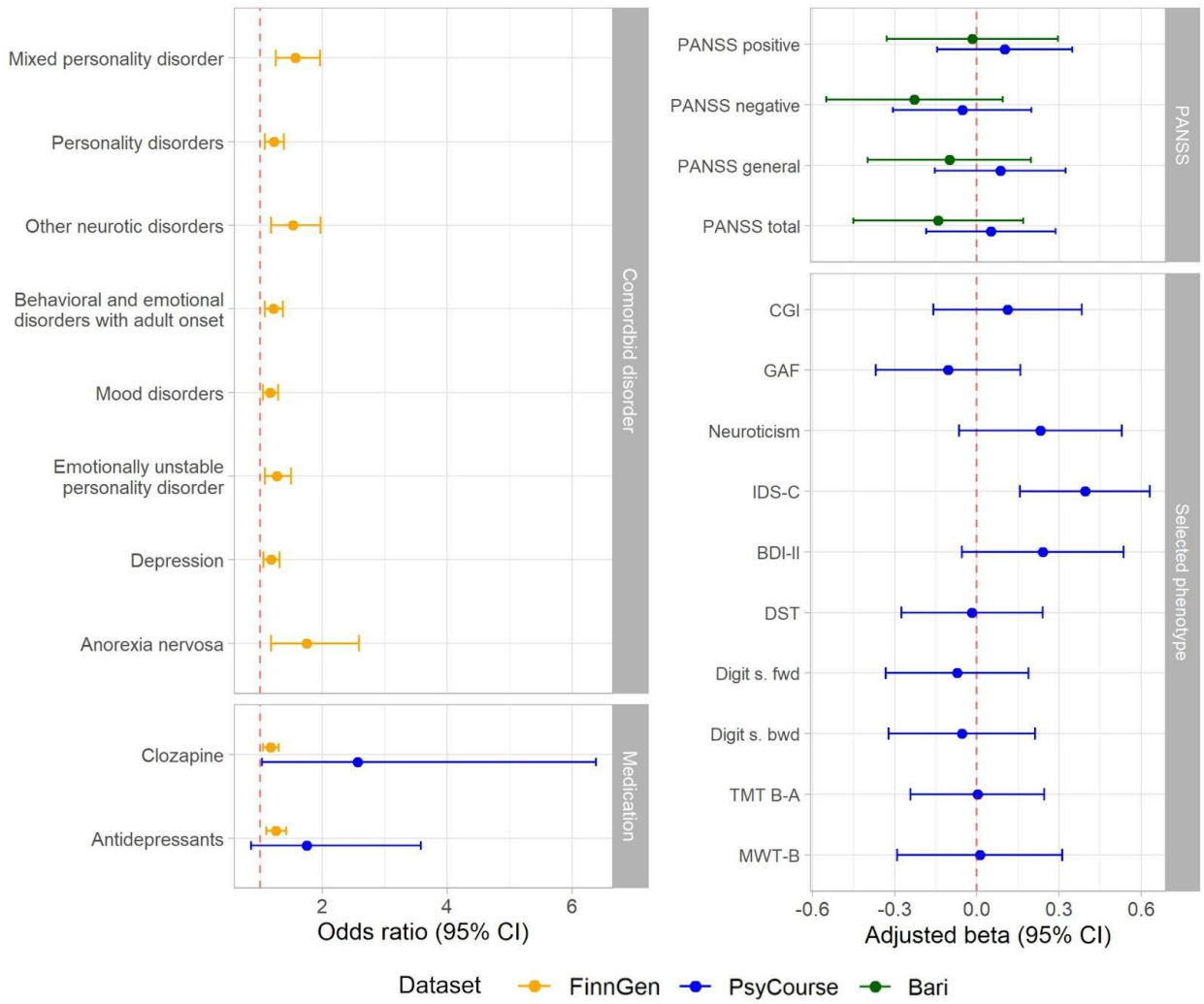


Figure 5. Phenotypic profiling of Biotype 1 and Biotype 2. Feature size = 24. Baseline: Biotype 2. All comparisons are adjusted for the SCZ PGS, sex, and age. In $Val_{PsyCourse}$ and Val_{Bari} additionally adjusted for center/study; all results with the 95% CI (displayed) not overlapping the null effect are also significant with an FDR of 0.05 applied within each category (for comorbid disorders only the significant ones are shown); analyzed case numbers of Biotype 1 and Biotype 2 are listed in brackets ($n_{Biotype\ 2}$ / $n_{Biotype\ 1}$): Clozapine (105/88), Antidepressants (105/88), PANSS = Positive and Negative Syndrome Scale, positive, negative, general and total score ($Val_{PsyCourse}$ 98/115, Val_{Bari} 84/86). CGI = Clinical global impression (99/115), GAF = Global Assessment of Functioning (98/115), Neuroticism = Dimension from Personality (74/85), IDS-C_{30} = Inventory of depressive symptomatology (97/113), BDI-II = Beck Depression Inventory-II (88/103), DST = Digit-Symbol-Test (98/110), Digit s. fwd = Verbal digit span forward (98/112), Digit s. bwd = Verbal digit span backward (98/112), TMT B-A = Trail making test part B - part A ($Val_{PsyCourse}$ 96/104), MWT-B = Multiple-Choice Vocabulary Intelligence Test (67/81). For all FinnGen analyses, the corresponding group sizes are (2621/3044).

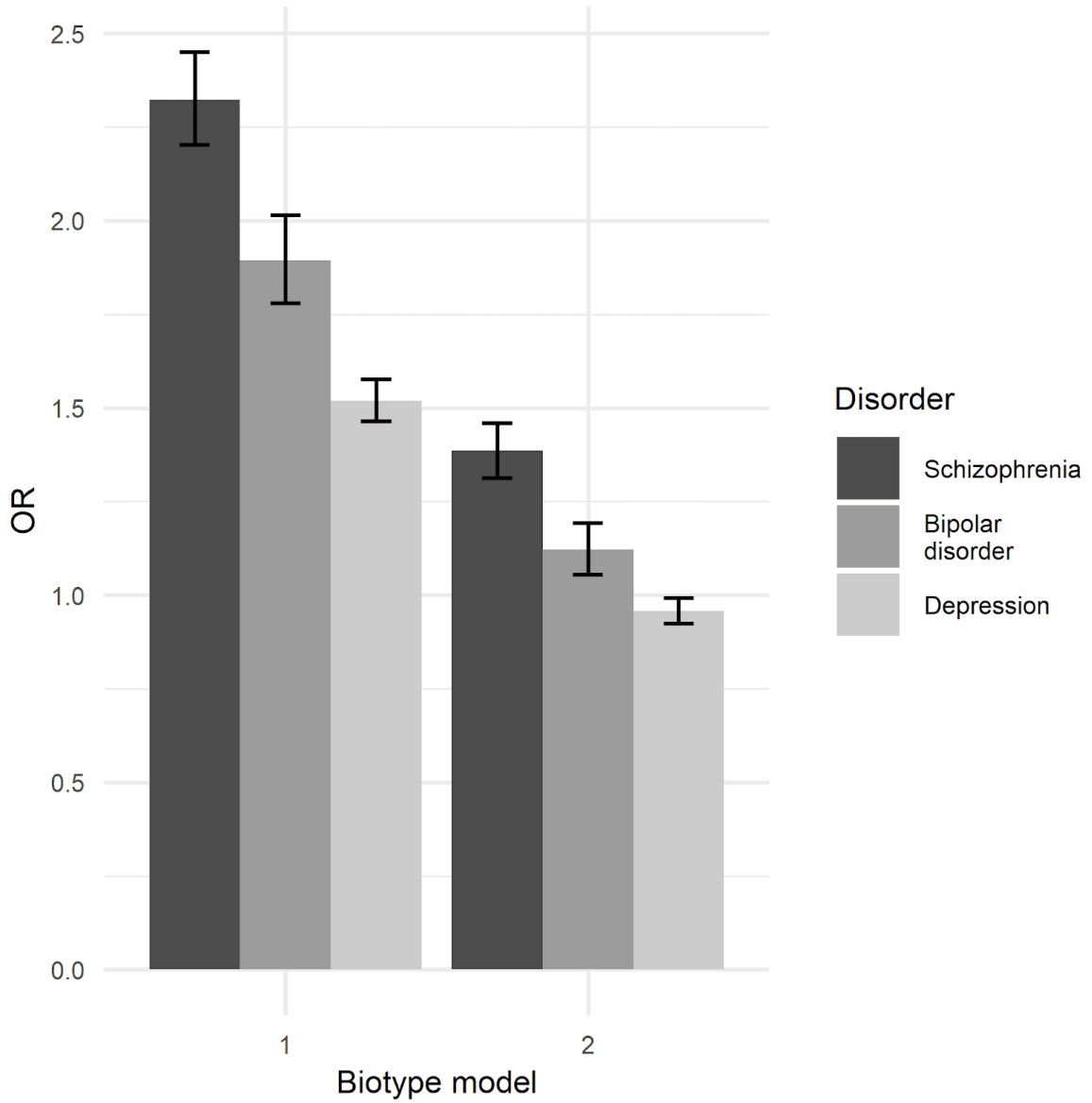


Figure 6. Differential diagnostic associations of SCZ biotypes. The odds ratios (OR) are derived from two-sided Fisher's exact test where the input was a cross table of either Biotype 1 or Biotype 2 model (x-axis) and a disorder. The error bars show 95% confidence intervals.

Discussion

In the present study, we aimed to comprehensively capture the genetic risk architecture of SCZ through the simultaneous investigation of patients' susceptibility to psychiatry-relevant phenotypes using 413 publicly available GWAS, leveraging an exceptionally large training dataset ($Train_{FinnGen}$). We asked whether patient subgroups exist that differ in the risk patterns related to these phenotypes and identified two reproducible biotypes in large-scale data. Our approach leverages the large statistical power of the original genotype data the PGS were derived from, as well as the power of machine learning to identify clinically meaningful biotypes from the large number of PGS. We classified patients into two biotypes, one of which was characterized by high common-variation genetic risk for neuroticism-related phenotypes and depression-related phenotypes and low PGS for cognitive traits (Biotype 1), whereas the other biotype (Biotype 2) had a polygenic profile similar to HC in those dimensions. While the biotypes exhibited approximately the same elevated levels of genetic risk for SCZ, a machine-learning model based on Biotype 1 clearly outperformed a model based on Biotype 2 in predicting SCZ in independent validation datasets ($Val_{PsyCourse}$ and Val_{Bari}). Furthermore, Biotype 1 patients tended to have clinical manifestations that indicate higher rates of treatment resistance and complexity. The predictions of the machine learning models based on both biotypes had the strongest association with SCZ, an intermediate one with BIP, and the weakest with MDD.

While previous studies have demonstrated that simple disease prediction can be improved with a multi-PGS approach[24], our results demonstrate the utility of using this approach for data-driven clustering of patients concerning differing genetic biotypes. Notably, the results indicate that stable and reproducible biotypes could be identified once several PGS (more than 5) were included in the machine-learning procedure. With a lower number, the biotypes were largely defined by their level of SCZ PGS and failed to capture the multi-phenotypic risk architecture observed at higher feature sizes. One remarkable observation of our study was, that while not substantially differing in SCZ risk load, the model based on Biotype 2 was a significantly weaker predictor of SCZ case status than the model based on Biotype 1. There are several potential explanations for this finding that are not mutually exclusive. First, the genetic risk contributions for neuroticism, depression-related traits, and low cognitive performance may reflect biological mechanisms necessary for the disease risk to elevate significantly. While the corresponding genetic risk may be captured by SCZ PGS, it may be that it is merely underweighted in the PGS calculation. Another hypothesis would be one where the SCZ genetic risk estimation is more accurate in Biotype 1 than in Biotype 2 patients. In this scenario, the higher levels of polygenic risk of non-schizophrenia traits/disorders in

Biotype 1 would suggest higher prevalence of alleles that confer risk to both SCZ and some other related trait (i.e they are pleiotropic). These pleiotropic SNPs could be more often truly causal for SCZ than the ones only associated with SCZ risk as the additional associations can be considered as an increase in the strength of statistical evidence, making such SNPs less likely to be false positives even when the p-value in SCZ GWAS summary statistics would not suggest so. Our finding that the PGS based on conjFDR SNPs are slightly better predictors than PGS based on comparable subsets of SCZ risk SNPs lends support for this. However, as the differences are small the conclusion must be considered suggestive.

Biotype 1 yielded a superior predictive model compared to Biotype 2 and was also associated with multiple clinically relevant measures, which collectively point towards higher illness severity and complexity in Biotype 1. Notably, the observed phenotypic profiles of the biotypes closely reflect the PGS groups defining them. This link between genetic liability and actual phenotype is further substantiated by the gradual emergence of the phenotypic profiles with increasing feature sizes (Supplementary Figure 3-18). The explanation for Biotype 1 being the more predictive model and having a higher likelihood for treatment resistant cases (as suggested by the higher likelihood for clozapine use) while having the same genetic SCZ liability, it can be speculated that this might be a result of the interaction between the predisposition for neurotic, depressive, and cognitive traits associated with Biotype 1. Higher neuroticism and lower cognitive ability may reduce stress-coping ability, leading to higher SCZ penetrance in Biotype 1, as stress is one of the main environmental risk factors for SCZ[35–37]. Lower stress coping ability can also lead to maladaptive coping strategies like drug use, a well established trigger for psychosis[38]. Accordingly, the link between Biotype 1 and depression-related phenotypes (increased rates of affective comorbidities and antidepressant use and a significantly higher score on a depression specific questionnaire; Figure 5) might be driven by both a higher genetic liability for depression related traits and neuroticism, as neuroticism is a known risk factor for MDD and they have a high positive genetic correlation[39–42]. An important notion is that while SCZ PGS has been linked to clinical characteristics within SCZ patients (Bipolar Disorder and Schizophrenia Wo...), the current approach goes beyond the identification of subtypes that would have been possible using the SCZ PGS This is also indicated by the observation that all of the phenotypic associations survived adjustment for SCZ PGS. For example, the association with clozapine use was not attenuated by controlling for SCZ PGS (though, SCZ PGS has previously been linked to clozapine dosage within SCZ patients who take it, see [43]). Using a set of well-selected PGS for cognition and neuroticism may have produced similar results but the fact that our data-driven approach arrived at these theoretically sensible dimensions can only be seen as a strength. Additionally the data-driven

approach allowed us to determine which PGS are not important in the context of SCZ biotype identification.

For all its strengths, the current study has several limitations. First, the method by which we assigned the patients to biotypes left some of them without one, as a correct classification was required by exactly one of the prediction models. Second, by no means do our findings suggest there are exactly two biotypes. As one of the main goals of this study was reproducibility and thus considering the sample size limitations of the validation datasets, we restricted the analysis to classify patients into two biotypes. Third, our method only considered genetic predisposition to psychiatry-related traits. It is possible that the inclusion of genetic risk in somatic domains would improve the classification and provide more insight into underlying biological mechanisms[44–46]. Fourth, the heterogeneity between the datasets we used in terms of the availability of phenotypic data hampered us from conducting more harmonized analyses in that regard, drawing less definite conclusions about generalizability. On the other hand, the fact that our procedure works across such different datasets can be seen as an indication of good portability. Finally, as only samples of European descent and GWAS from that ancestry were used, the findings presented in this paper are not valid for non-European ancestry individuals. For our pipeline to become applicable to those populations better-powered GWAS from non-European ancestries are needed[47,48].

In summary, our data-driven machine learning scheme successfully leveraged a large number of psychiatry-related PGS, for the first time, to derive two biotypes of SCZ patients. A prediction model based on Biotype 1, broadly speaking characterized by high neuroticism and depression related traits and low cognitive functioning, is a promising step toward identifying SCZ patients with a poor prognosis at an early stage. It may also prove to be useful in personalizing treatments, a prospect that would need to be evaluated in a randomized and controlled clinical trial. The current study was focused on SCZ but in the future, it could be applied to any phenotype, psychiatric or somatic.

Methods

Overview of method

The aim of the multi-PGS approach, pursued in this study, was to better characterize the genetic architecture of SCZ by the combined analysis of polygenic risk of a large number of psychiatry-relevant phenotypes. For this, we implemented PRScope which extracts GWAS summary statistics from the GWAS Catalog[49]. Phenotypes were automatically selected based on study annotations, the corresponding GWAS summary statistics were quality controlled

and the PGS was then determined in a target genotype dataset. Care was taken to ensure that the target genotype datasets did not contain data overlapping with that of the GWAS summary statistics in the GWAS Catalog, to prevent potential bias. The PGS for the large number of phenotypes (multi-PGS) was then used for machine learning, to identify biotypes of SCZ that differed in their respective composition of associated PGS. These biotypes were identified in a large-scale training dataset, and validated in two independent validation datasets.

Multi-PGS calculation framework

The combined analysis of PGS for different phenotypes faces several challenges, including data consistency issues arising from heterogeneous file formats and naming conventions (related to data as well as phenotypes), differences in genome assemblies, and variant orientation regarding DNA strands. The GWAS Catalog addresses these issues by harmonizing and categorizing summary statistics from published GWASs and making them accessible on a single platform. The harmonization is carried out by the *harmonizer pipeline*[50], which aligns variants to the desired genome assembly and reference data. The categorization is performed using the EFO[51], which annotates GWAS summary statistics with consistent terminology. Furthermore, the GWAS Catalog provides a user-friendly interface for querying the database by criteria such as phenotype, gene, SNP, trait, or study. These features of the GWAS Catalog enabled us to search for GWAS summary statistics related to the disease of interest, automate QC steps, and determine PGS. Utilizing these features, we implemented a pipeline consisting of four modules, each designed for independent use. Marees et al.[52] and more recently Choi et al.[53] have published detailed protocols for the calculation of PGS. These protocols include detailed instructions for performing quality control on genotype data and GWAS summary statistics, as well as standards for interpreting PGS. Our approach followed the guidelines presented by Marees, et al. and Choi, et al.

Automated GWAS search module

A primary challenge for researchers conducting multi-PGS studies is selecting source GWAS summary statistics from among 100,000+ available GWAS summary statistics in the GWAS Catalog. It is crucial to include GWAS summary statistics of traits that are of likely relevance to the target phenotype, to meaningfully reduce the dimensionality of the multi-PGS data but still be able to capture as much of the genetic architecture as possible. To address this, we implemented a module to automate this selection process. This module uses the EFO Browser to filter and collect relevant GWAS summary statistics files. Users first perform a preliminary search in the EFO browser to identify relevant ontologies for the target phenotype and save

the selected EFO IDs. In the next step, they can define search criteria in the module configuration file, such as GWAS population, number of SNPs, sample size, and publication date (described in detail in Supplementary Methods). The module uses the R library *gwasrapidd*[54] to search the GWAS Catalog for studies annotated with the specified EFO IDs, applying the user-defined filters (see section *Filtering criteria* in the Supplementary Texts). By running the module, users receive a list of GWAS Catalog summary statistic IDs that meet the specified criteria (described in detail in Supplementary Figure 19).

In the present study, our strategy for ontology selection aimed at including a broad spectrum of phenotypes relevant to SCZ. For this, we focused on psychiatry-related ontologies from different domains such as biological, behavioral, brain development, psychiatric traits, and psychiatric disorders. In total, we selected 23 ontologies (Supplementary Table 1) and performed our GWAS search module to search the GWAS Catalog for GWAS summary statistics annotated with these ontologies or their sub-ontologies. The resulting list included 413 GWAS summary statistics that met our criteria and had a harmonized version of the respective GWAS summary statistics. Using these harmonized versions of GWAS summary statistics led to two favorable outcomes: First, the GWAS summary statistics were transformed into a uniform format, enabling effective calculation of PGS. Second, variants were already mapped to the latest genome assembly and alleles were orientated to the forward strand. Additionally, we found 1,462 relevant GWAS summary statistics that met our filtering criteria but did not have a harmonized version (Supplementary Table 17). Furthermore, 65 GWAS summary statistics not listed in the GWAS Catalog but accessible from PGC, CTG, and other public sources (Supplementary Table 18-22), were prepared and included in our analysis using the same harmonization pipeline as applied by the GWAS Catalog.

GWAS summary statistics QC module

Duplicate, ambiguous, and multi-allelic SNPs, as well as SNPs that lacked both beta value and odds ratio, were removed from the GWAS summary statistics. The p-value was set to 1 for SNPs without a p-value (described in detail in Supplementary Figure 20).

Genotype QC module

First, SNPs and subjects were filtered based on a relaxed threshold (missing call rate of > 20%). Then a filter with a more stringent threshold was applied. Subjects with missing call rates of > 2% were removed. Next, SNPs with missing call rates of > 2%, minor allele frequency < 1%, and deviation from the Hardy-Weinberg equilibrium ($p \leq 1 \times 10^{-6}$) were removed. Additionally, non-autosomal, duplicate, and ambiguous SNPs were excluded. The

R-package used to estimate relatedness was carried out as implemented in plinkQC v0.3.4[55], by estimating the relatedness between pairs of individuals based on identity-by-state similarity (PI_HAT > 0.1875) and retaining the subject with the lowest genotype missingness in pairs or larger groups of related subjects (described in detail in Supplementary Figure 21). Detailed information about the imputation and pre-imputation can be found in Supplementary Texts (Pre-imputation and Imputation).

PLINK v1.9 was used to assess the impact of population stratification via PC analysis as described in RICOPILI's PCA module[56]. In case the sample was bigger than 1,000 subjects, a subset of approximately 1,000 subjects was randomly selected for the pruning of the SNP set prior to the PCA. Similarly, if more than 1 million SNPs were present in the dataset, a random SNP set of ~1 million SNPs was selected for pruning. For the PCA calculation, we excluded SNPs with more than 2% missing call rates, a minor allele frequency below 5%, or a Hardy-Weinberg equilibrium deviation ($p \leq 1 \times 10^{-3}$), and SNPs in the long-range linkage disequilibrium (LD) regions of the Major Histocompatibility Complex region (chr 6, 25-35 MB) and the chromosome 8 inversion region (chr 8, 7-13 MB). The remaining SNPs were pruned using an r-squared threshold of 0.2, a window size of 200 kb, and a step size of 100 to minimize LD, followed by a second round of pruning with the same parameters. Finally, PLINK was used to calculate PCs for ancestry estimation for all subjects using the SNPs generated in the previous steps.

PGS calculation module

The PGS calculation module of PRScope offers the advantage of quickly generating multiple PGS calculations in parallel using PRSice-2[26], requiring minimal configuration. For the present study, PGS calculations for 413 pre-processed GWAS summary statistics were performed for each GWAS summary statistic using up to 10 different significance thresholds, representing different levels of statistical significance for the genetic variants included in each PGS. PGS including fewer than 500 SNPs at a given significance threshold were excluded from the analyses. Detailed information about the PGS calculation can be found in Supplementary Texts (PGS Calculation Parameters).

Cohort overview

The present study included data from three cohorts. Data from FinnGen release 12 ($Train_{FinnGen}$) (7486/27288 cases/controls for SCZ, 7555/11198 for BIP, 27187/26993 for MDD. Controls for BIP and MDD are partly shared, but not with SCZ controls) were used for the identification of multi-PGS biotypes, and to test transdiagnostic associations. Data from the PsyCourse Study ($Val_{PsyCourse}$) was used for biotype validation. Data from the Bari cohort

(Val_{Bari}) was used for additional biotype validation[21]. All cohorts were independent of the original GWAS summary statistics the PGS were derived from, to prevent circularity. Supplementary Table 2-4 show the composition of the datasets after QC with respect to gender and diagnosis. To minimize bias due to the differences in cohorts, only the SNPs that overlap across three datasets were included in the PGS calculation (2.925.749 SNPs from each dataset). The Multi-PGS data in all three datasets were first checked for outliers by robust multivariate outlier detection using the OGK estimator[57]. Based on this method, the log distances above a threshold (≥ 4.5) were classified as outliers, often representing individuals from different populations or those with genotyping errors. Subsequently, the Multi-PGS data were residualized against the first 20 genetic PCs using linear regression and z-transformed.

All participants provided written informed consent (see below). Procedures were approved by Ethics Committee II of the University of Heidelberg Medical Faculty Mannheim (protocol no. 2024-845) and the respective local ethical committees.

Study subjects in FinnGen provided informed consent for biobank research, based on the Finnish Biobank Act. Alternatively, separate research cohorts, collected prior the Finnish Biobank Act came into effect (in September 2013) and start of FinnGen (August 2017), were collected based on study-specific consents and later transferred to the Finnish biobanks after approval by Fimea (Finnish Medicines Agency), the National Supervisory Authority for Welfare and Health. Recruitment protocols followed the biobank protocols approved by Fimea. The Coordinating Ethics Committee of the Hospital District of Helsinki and Uusimaa (HUS) statement number for the FinnGen study is Nr HUS/990/2017.

The FinnGen study is approved by Finnish Institute for Health and Welfare (permit numbers: THL/2031/6.02.00/2017, THL/1101/5.05.00/2017, THL/341/6.02.00/2018, THL/2222/6.02.00/2018, THL/283/6.02.00/2019, THL/1721/5.05.00/2019 and THL/1524/5.05.00/2020), Digital and population data service agency (permit numbers: VRK43431/2017-3, VRK/6909/2018-3, VRK/4415/2019-3), the Social Insurance Institution (permit numbers: KELA 58/522/2017, KELA 131/522/2018, KELA 70/522/2019, KELA 98/522/2019, KELA 134/522/2019, KELA 138/522/2019, KELA 2/522/2020, KELA 16/522/2020), Findata permit numbers THL/2364/14.02/2020, THL/4055/14.06.00/2020, THL/3433/14.06.00/2020, THL/4432/14.06/2020, THL/5189/14.06/2020, THL/5894/14.06.00/2020, THL/6619/14.06.00/2020, THL/209/14.06.00/2021, THL/688/14.06.00/2021, THL/1284/14.06.00/2021, THL/1965/14.06.00/2021, THL/5546/14.02.00/2020, THL/2658/14.06.00/2021, THL/4235/14.06.00/2021, Statistics Finland (permit numbers: TK-53-1041-17 and TK/143/07.03.00/2020 (earlier TK-53-90-20)

TK/1735/07.03.00/2021, TK/3112/07.03.00/2021) and Finnish Registry for Kidney Diseases permission/extract from the meeting minutes on 4th July 2019.

The Biobank Access Decisions for FinnGen samples and data utilized in FinnGen Data Freeze 12 include: THL Biobank BB2017_55, BB2017_111, BB2018_19, BB_2018_34, BB_2018_67, BB2018_71, BB2019_7, BB2019_8, BB2019_26, BB2020_1, BB2021_65, Finnish Red Cross Blood Service Biobank 7.12.2017, Helsinki Biobank HUS/359/2017, HUS/248/2020, HUS/430/2021 §28, §29, HUS/150/2022 §12, §13, §14, §15, §16, §17, §18, §23, §58, §59, HUS/128/2023 §18, Auria Biobank AB17-5154 and amendment #1 (August 17 2020) and amendments BB_2021-0140, BB_2021-0156 (August 26 2021, Feb 2 2022), BB_2021-0169, BB_2021-0179, BB_2021-0161, AB20-5926 and amendment #1 (April 23 2020) and it´s modifications (Sep 22 2021), BB_2022-0262, BB_2022-0256, Biobank Borealis of Northern Finland_2017_1013, 2021_5010, 2021_5010 Amendment, 2021_5018, 2021_5018 Amendment, 2021_5015, 2021_5015 Amendment, 2021_5015 Amendment_2, 2021_5023, 2021_5023 Amendment, 2021_5023 Amendment_2, 2021_5017, 2021_5017 Amendment, 2022_6001, 2022_6001 Amendment, 2022_6006 Amendment, 2022_6006 Amendment, 2022_6006 Amendment_2, BB22-0067, 2022_0262, 2022_0262 Amendment, Biobank of Eastern Finland 1186/2018 and amendment 22§/2020, 53§/2021, 13§/2022, 14§/2022, 15§/2022, 27§/2022, 28§/2022, 29§/2022, 33§/2022, 35§/2022, 36§/2022, 37§/2022, 39§/2022, 7§/2023, 32§/2023, 33§/2023, 34§/2023, 35§/2023, 36§/2023, 37§/2023, 38§/2023, 39§/2023, 40§/2023, 41§/2023, Finnish Clinical Biobank Tampere MH0004 and amendments (21.02.2020 & 06.10.2020), BB2021-0140 8§/2021, 9§/2021, §9/2022, §10/2022, §12/2022, 13§/2022, §20/2022, §21/2022, §22/2022, §23/2022, 28§/2022, 29§/2022, 30§/2022, 31§/2022, 32§/2022, 38§/2022, 40§/2022, 42§/2022, 1§/2023, Central Finland Biobank 1-2017, BB_2021-0161, BB_2021-0169, BB_2021-0179, BB_2021-0170, BB_2022-0256, BB_2022-0262, BB22-0067, Decision allowing to continue data processing until 31st Aug 2024 for projects: BB_2021-0179, BB22-0067,BB_2022-0262, BB_2021-0170, BB_2021-0164, BB_2021-0161, and BB_2021-0169, and Terveystalo Biobank STB 2018001 and amendment 25th Aug 2020, Finnish Hematological Registry and Clinical Biobank decision 18th June 2021, Arctic biobank P0844: ARC_2021_1001.

Machine learning analysis

The goal of the applied machine learning analyses was to identify (using $Train_{FinnGen}$) and validate (using the $Val_{PsyCourse}$ and Val_{Bari} data) subgroups of patients with SCZ with differing multi-PGS profiles (i.e. ‘biotypes’). Using the multi-PGS matrix we first determined the respective z-score of each PGS for patients with SCZ by subtracting the median of the controls and dividing by the standard deviation of the controls. To tune clustering to SCZ-relevant

features, we then selected the GWAS summary statistics that showed the strongest association with the diagnosis. Clustering was based on different numbers of such transformed PGS (from top 1 to top 45) using k-means clustering (k=2, performed using the *kcca* function of the *flexclust* library)[58], to identify 2 subgroups. Clustering was repeated 100 times, the co-occurrence of subjects within a given cluster was recorded for each subject pair, and a consensus was identified by clustering the resulting matrix. Random forest machine learning[59] was applied to characterize the multi-PGS signature of each biotype and to predict this signature in independent data. To characterize the PGS signature associated with each biotype, 51-times repeated random forest machine learning (using the ranger v0.17.0 R library with 1000 trees and default parameters)[60] was applied using under-sampling. Under-sampling was used to balance the unequal group sizes of patients within a given biotype, and the control group. To evaluate the reproducibility of this approach in the training data, the entire procedure was applied within a 10-fold cross-validation. For instance, to train the model for Biotype 1, the training folds comprised an equal number of HCs and SCZ patients assigned to Biotype 1. Then, the Biotype 1 membership was predicted by the model in the test fold and associations between the predicted biotype memberships and diagnosis were quantified as the odds ratio (OR). The significance was determined using two-sided Fisher's exact test. The same approach was applied to Biotype 2. The resulting p-values were adjusted for multiple testing (from top 1 to top 45, i.e., the number of tests: 45) using R's p.adjust function with the BH correction method. The BH-adjusted p-values were -log10 transformed and plotted (Figure 2a). To prevent data leakage, the z-scoring and cluster identification were performed separately for each CV-iteration. To obtain comparable cluster assignments (as the cluster numbering is arbitrary) across CV folds, these were aligned to the first CV-iteration based on linear correlation obtained from predictions in overlapping training subjects.

Feature selection for clustering and prediction

Feature selection aimed at identifying the PGS that showed the strongest association with diagnosis in order to improve the clustering performance and the predictive power of the ML models. Our approach included all PGS calculated at different significance thresholds for GWAS summary statistics, allowing us to fully utilize the full information provided by the corresponding GWAS summary statistics file. To achieve this, we first grouped the PGS of a given GWAS summary statistics into a "PGS block" and ranked these blocks based on the "best PGS" within a PGS block that showed the strongest correlation with the diagnosis (Supplementary Figure 22). A PGS block is composed of PGS calculated at up to 10 different significance thresholds from a GWAS summary statistics (see Supplementary Texts, *PGS*

Calculation Parameters). For downstream analysis, all PGS of the highest-ranked blocks were included, ranging from the top 1 to the top 45.

Investigating cluster profiles

This step aims to characterize the biotype profiles by estimating the associations between selected GWAS summary statistics and the biotypes using logistic regression. Subjects uniquely assigned to each biotype and the healthy control group were sampled to create dependent variables of logistic regression. Logistic regression was then applied separately for each best PGS of selected GWAS summary statistics to associate the biotype groups with GWAS summary statistics as displayed in the formulas (1,2).

$$(1)\ Y_{model\ 1} \sim PGS_j$$

$$(2)\ Y_{model\ 2} \sim PGS_j$$

$Y_{model\ 1}$ and $Y_{model\ 2}$ are binary variables where value of 1 indicates Biotype 1 and Biotype 2, respectively, and value of 0 indicates HC. PGS_j is the best PGS of jth GWAS summary statistics of the top 45.

The beta values calculated by the logistic regression represent the association between the best PGS of selected GWAS summary statistics and the biotype assignments (Figure 3). Similarly, PCs were calculated for each group (Biotype 1, Biotype 2, and HC) using all PGS of a PGS group. The mean values of the first PC for each group across different feature sizes were plotted (Figure 2). For clearer interpretation we display SCZ as a single PGS.

Investigating cluster stability

The k-means clustering solution for a given feature size was compared with the clustering results of all other feature sizes and visualized as a heatmap (Figure 3). Cluster similarities were quantified using the adjusted Rand index[61].

Phenotypic profiling

To compare the phenotypic profiles of the two biotypes, we used a general linear model (GLM) regressing the phenotypic variable *j* on the biotype (*n* = either 1 or 2), with the SCZ PGS on all p-value thresholds and available demographic data used as *k* covariates using (3):

$$(3)\ pheno_j \sim biotype_n + cov_1 + \ldots + cov_k$$

Continuous phenotypes were initially z-standardized ($\frac{x-\mu}{\sigma}$). R version 4.4.1 was used for all statistical analyses. The glm function from the standard stats package was used to apply the GLM as displayed in (3). In the case of binary phenotypic variables, a logistic regression with the same covariates was performed and odds ratios are reported. As the two distinctive groups most representative of each of the biotypes are the SCZ patients uniquely correctly identified by either of the biotypes, we used these groups for the comparisons. All results were displayed as forest plots showing the beta-value (or odds ratio) of the biotype and their respective 95% confidence intervals.

The PsyCourse Study

PsyCourse is a multi-center observational longitudinal deep-phenotyping study of the affective-to-psychotic spectrum, including healthy controls. All participants provided written informed consent, and the study was approved by all relevant Ethics Committees (see [30]). Demographic covariates used for the phenotypic profiling in $Val_{PsyCourse}$ were sex, age, and center of recruitment (in addition to the SCZ PGS on all p-value threshold levels used in all datasets). Due to the large number of recruitment sites (n=20), and the unequal distribution of patients recruited across these sites, we focused on sites that recruited at least 20 patients (resulting in a selection of n = 339 SCZ patients). Due to the longitudinal study design, some of the phenotypic measures are taken at up to four time points (spread over 18 months) for each subject. If this was the case, we used the maximum value over the time points. Due to the sparse pattern of missings in the PsyCourse dataset, the number of available data varies depending on the phenotypic measure. Based on the results from $Train_{FinnGen}$, we compared the number of patients taking clozapine or antidepressants between the biotypes (this analysis was run without controlling for center, due to the low number of cases from each center). The information about the clozapine and antidepressant treatment was derived from the medication dataset, a subset of the PsyCourse data set (see below). In addition to this we analyzed phenotypic measures based on PGS defining the biotypes which could be summarized as SCZ (Positive and Negative Syndrome Scale (PANSS): positive, negative and general symptoms subscale and total score) and three additional major components, which genetic predispositions mainly drive the difference between the biotypes: Cognitive function (verbal digit span forward, verbal digit span backward, Digit-Symbol-Test (DST), Multiple-Choice Vocabulary Intelligence Test (MWT-B), TMT B-A = Trail Making Test part B - part A), depression-related traits (Inventory of depressive symptomatology ($IDS\text{-}C_{30}$), Beck Depression Inventory II (BDI-II)) and neuroticism (Neuroticism = Dimension from BFI-10 questionnaire[62]). On top of these, we also added measures of overall disease severity (Clinical global impression (CGI), Global Assessment of Functioning (GAF)). Variable names as used

in the PsyCourse codebook: https://data.ub.uni-muenchen.de/251/1/210908_PsyCourse_v5.0.html are mapped in Supplementary Table 23. The resulting p-values were separately FDR-corrected for each analysis block (medication, PANSS scales, additional phenotypic variables).

Bari

Demographic covariates used for the phenotypic profiling in Val_{Bari} were sex, age and the original study in which the data was acquired (in addition to the SCZ PGS on all p-value threshold levels used in all datasets). The Bari dataset consists of data from two studies. Clinical phenotypic data in this dataset was the PANSS negative sum, PANSS positive sum, PANSS general sum, and PANSS total sum, which was available for 249 of the SCZ patients. Other phenotypic data was also available but for only a very small number of SCZ patients. Analogous to the approach in the PsyCourse dataset we used the max value whenever values for more than one time point were available. The resulting p-values were separately FDR-corrected for the PANSS scales.

FinnGen

To examine biotype differences in comorbidity of psychiatric phenotypes each phenotype was analyzed with a separate logistic regression with biotype as a dichotomous predictor and SCZ PGS, age, and sex as covariates. Endpoints with less than 100 cases among the SCZ patients were excluded. Benjamini-Hocheberg multiple-testing correction was applied within a given feature size at a false discovery rate of 0.05. The same procedure was applied for the medication analyses (see *Phenotype definitions in FinnGen* for how medication history was determined).

Selection of feature sizes for the full FinnGen data training

To limit the number of models to be tested in the validation data sets ($Val_{PsyCourse}$ and Val_{Bari}), a few feature sizes were selected. This was achieved by first converting the cluster assignment table across all feature sizes to a distance matrix (function *dist* with the method option “binary”, from the R-package *stats*[63]) and then performing hierarchical clustering (function *hclust* from the R-package *stats*[64]) to determine feature sizes which give out similar cluster assignments. Upon visual inspection of the resulting dendrograms, 5 clusters (see Supplementary Figure 1) were deemed appropriate, and for each of the clusters, the lowest feature size was selected for the full data training. Additionally, feature sizes 1 to 5 were included so that the development of the profiles of the clusters could be demonstrated when more and more features were added to the clustering/prediction procedure.

Formation of the PGS groups

The first step in forming the three PGS groups (cognition, depression, and neuroticism) was to calculate genetic correlations using LDSC software for the top 45 GWASs excluding SCZ and a trait incorporating 12 psychiatric disorders[31,32]. The European samples reference panel from the 1000 Genomes project[65] was used for LD scoring. Then, hierarchical clustering was performed on the absolute values of the resulting (quasi-) correlation matrix. Upon visual inspection and theoretical consideration, a three-cluster solution was selected (Supplementary Figure 1-2).

Subset PGS analysis

To provide insight as to why one of the cluster models predicts SCZ better than the other, despite similar overall polygenic risk for SCZ. To investigate this we further analyzed the polygenic SCZ risk, by assessing the contribution of genomic loci shared between the SCZ GWAS summary statistics and the other three GWAS groups. We calculated a corresponding subset SCZ-PGS using the SNPs with conjunction FDR<0.1 for each of the PGS groups, and weighting the SNPs by their association in the SCZ GWAS. To achieve this, we first conducted the conjunctional FDR[34,66] approach between SCZ GWAS summary statistics and all the other GWAS summary statistics among the top 45. If a SNP had a conjFDR of below 0.1 with any of the GWAS summary statistics within a PGS group (see *Formation of the PGS groups*) it was included in the calculation of the subset SCZ-PGS of that PGS group. For example, for the PGS group “Cognition” a corresponding subset PGS was constructed, capturing the genetic variation relating to both SCZ and cognitive traits (i.e. pleiotropic effects). For each of the three SNP subsets 100 p-value matched SNP sets (as indicated in the SCZ summary statistics) of equal size were sampled from the SNPs outside of the corresponding pleiotropic SNP set. The matching was performed with the R function *matchit* (distance = “glm”) from the package *MatchIt*[67] Different matched subsets were achieved by adding slight random error to the p-values against which the matching was done.

After defining the SNP sets, PGS were calculated for each of them. Their predictive performance was assessed with logistic regressions where SCZ was the outcome and a given subset PGS the predictor. Before the analyses, all the PGS were adjusted for the first 20 ancestral PCs and then standardized. Finally, Cochran’s Q test (function *rma* with method = “EE”, from the R package *metafor*[68]) was used to determine a p-value for the difference in effect size between a subset PGS of a given GWAS group and the best corresponding comparative PGS.

Phenotype definitions in FinnGen

SCZ, BIP, and MDD cases were selected using the specialized inpatient and outpatient, cause of death, and drug purchase registers (coded in the data as INPAT, OUTPAT, DEATH and PURCH, respectively). For each disorder, their respective ICD codes served as case criteria. Additionally, individuals who had ever purchased the drug clozapine were determined as having SCZ, as that drug is virtually never prescribed for other disorders. Lithium for bipolar disorder or antidepressants for depression were not utilized to that end as they are more commonly prescribed for other disorders and diseases. The disorders were regarded as hierarchical in the way that SCZ superseded BIP which in turn superseded MDD.

For defining controls, more stringent criteria were applied. In the case of SCZ, all psychotic disorders lead to exclusion, as well as Olanzapine, Risperidone, Amisulpride and Haloperidol purchases. The aforementioned drugs indicate a psychotic disorder or episode, but not necessarily confirmed SCZ. For BIP, the same criteria applied with the addition of not having any mood disorders. Additional registers of primary health care (PRIM_OUT) and drug reimbursement (REIMB) were used for control exclusion to add more certainty that control subjects are truly not cases. The exact ICD codes and ATC drug codes can be found in Supplementary Table 24.

From the remaining eligible controls a sex and age-matched sample was taken (*matchit* function in the R-package *MatchIt*[67]) to simplify downstream analyses and to limit the number of individuals for whom PGS needed to be calculated. The control group sizes were four times the number of cases for SCZ and BIP and of equal size for MDD. Finally, the full sample (cases of all disorder phenotypes and all control groups) was restricted to unrelated individuals so that the maximum allowed degree of relatedness was 3 (e.g. first cousins). 10,000 random selections of removal with a preference for keeping cases were performed and the one that resulted in the least subjects getting excluded was applied.

Phenotypes for comorbidity and medication analyses

All psychiatry-related (category F5) FinnGen core endpoints served as phenotypes in the comorbidity analyses without further modifications. Core endpoints are the ones defined by the FinnGen core analysis team and their descriptions can be browsed at https://risteys.finngen.fi/. For the medication analyses, clozapine use was identified with the ATC code N05AH02, and antidepressants with the ATC category N06A as recorded in the drug purchase register.

Acknowledgments

We thank all research participants and all researchers and clinicians who collected, generated, or processed the data used in this study. The full list of FinnGen banner authors can be found in Supplementary Table 25.

Funding

This work was supported by the Hector foundation II, the German Federal Ministry of Education and Research (BEST project, grant 01EK2101B), the German Research Foundation (DFG) (TRR 265/2 TP A06), and was endorsed by German Center for Mental Health (DZPG). U.H. was supported by European Union's Horizon 2020 Research and Innovation Programme (PSY-PGx, grant agreement No 945151) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, project number 514201724). T.G.S. was supported by the European Union Horizon 2020 Research and Innovation Program (PSY-PGx, grant agreement No 945151), and also by the Deutsche Forschungsgemeinschaft within the framework of the projects www.kfo241.de and www.PsyCourse.de [SCHU 1603/4–1, 5–1, 7–1]. T.G.S. was further supported by the Dr Lisa Oehler Foundation (Kassel, Germany), the Bundesministerium für Bildung und Forschung (BMBF, Federal Ministry of Education and Research; projects: IntegraMent [01ZX1614K], BipoLife [01EE1404H], e:Med Program [01ZX1614K]) and European Union's Horizon 2020 Research and Innovation Programme (ERA-NET Neuron Projects GEPI-BIOPSY [BMBF No 01EW2005] and MulioBio [BMBF No 01EW2009]). G.C.K. 's PhD scholarship is financially supported by Exprivia S.p.A. under the Italian ministerial decree D.M. 351. Furthermore, this study was supported by FAIR - Future AI Research (PE00000013), spoke 6 – Symbiotic AI, under the NRRP MUR program funded by the NextGenerationEU; Project CUP : H97G22000210007, and the Apulian regional government for the project: "Early Identification of Psychosis Risk.

Conflicts of Interest

G.B. has received lecture and consultant fees by Lundbeck. A.R. has received lecture fees by Janssen. A.B. received consulting fees from Biogen and lecture fees from Otsuka, Janssen, and Lundbeck. G.P. has received lecture fees by Lundbeck. A.M.-L. has received consultant fees from: Agence Nationale de la Recherche, Brain Mind Institute, Brainsway, CISSN (Catania International Summer School of Neuroscience), Daimler und Benz Stiftung, Fondation FondaMental, Hector Stiftung II, Janssen-Cilag GmbH, Lundbeck A/S,

Lundbeckfonden, Lundbeck Int. Neuroscience Foundation, MedinCell, Sage Therapeutics, Techspert.io, The LOOP Zürich, University Medical Center Utrecht, von Behring Röntgen Stiftung. A.M.-L.has received speaker fees from: Ärztekammer Nordrhein, BAG Psychiatrie Oberbayern, Biotest¨ AG, Forum Werkstatt Karlsruhe, International Society of Psychiatric Genetics, Brentwood, Klinik für Psychiatrie und Psychotherapie Ingolstadt, Lundbeck SAS France, med Update GmbH, Merz-Stiftung, Siemens Healthineers, Society of Biological Psychiatry. A.M.-L. has received editorial fees from: American Association for the Advancement of Science, Elsevier, Thieme Verlag. E.S. received speaker fees from bfd buchholz-fachinformationsdienst GmbH and Lundbeckfonden as well as editorial fees from Lundbeckfonden and the Wellcome Trust. The other authors have declared that there are no conflicts of interest in relation to the subject of this study.

References

1. Marder, S. R. & Cannon, T. D. Schizophrenia. *N. Engl. J. Med.* **381**, 1753–1761 (2019).
2. Chang, C.-K. *et al.* Life expectancy at birth for people with serious mental illness and other major disorders from a secondary mental health care case register in London. *PLoS One* **6**, e19590 (2011).
3. Fleischhacker, W. W. *et al.* Schizophrenia--time to commit to policy change. *Schizophr. Bull.* **40 Suppl 3**, S165–94 (2014).
4. Kahn, R. S. *et al.* Schizophrenia. *Nat Rev Dis Primers* **1**, 15067 (2015).
5. Braff, D. L., Ryan, J., Rissling, A. J. & Carpenter, W. T. Lack of use in the literature from the last 20 years supports dropping traditional schizophrenia subtypes from DSM-5 and ICD-11. *Schizophr. Bull.* **39**, 751–753 (2013).
6. Fusar-Poli, P. *et al.* Treatments of negative symptoms in schizophrenia: Meta-analysis of 168 randomized placebo-controlled trials. *Schizophr. Bull.* **41**, 892–899 (2015).
7. Potkin, S. G. *et al.* The neurobiology of treatment-resistant schizophrenia: paths to antipsychotic resistance and a roadmap for future research. *NPJ Schizophr* **6**, 1 (2020).
8. Lincoln, T. M., Dollfus, S. & Lyne, J. Current developments and challenges in the assessment of negative symptoms. *Schizophr. Res.* **186**, 8–18 (2017).
9. Dollfus, S. & Lyne, J. Negative symptoms: History of the concept and their position in

diagnosis of schizophrenia. *Schizophr. Res.* **186**, 3–7 (2017).

10. Owen, M. J. New approaches to psychiatric diagnostic classification. *Neuron* **84**, 564–571 (2014).
11. Millan, M. J. *et al.* Altering the course of schizophrenia: progress and perspectives. *Nat. Rev. Drug Discov.* **15**, 485–515 (2016).
12. Hilker, R. *et al.* Heritability of schizophrenia and schizophrenia spectrum based on the nationwide Danish Twin Register. *Biol. Psychiatry* **83**, 492–498 (2018).
13. Cardno, A. G. *et al.* Heritability estimates for psychotic disorders: the Maudsley twin psychosis series. *Arch. Gen. Psychiatry* **56**, 162–168 (1999).
14. Lichtenstein, P. *et al.* Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet* **373**, 234–239 (2009).
15. Legge, S. E. *et al.* Genetic architecture of schizophrenia: a review of major advancements. *Psychol. Med.* **51**, 2168–2177 (2021).
16. Kendler, K. What psychiatric genetics has taught us about the nature of psychiatric illness and what is left to learn. *Mol. Psychiatry* **18**, 1058–1066 (2013).
17. Alnæs, D. *et al.* Brain heterogeneity in schizophrenia and its association with polygenic risk. *JAMA Psychiatry* **76**, 739–748 (2019).
18. Pergola, G., Penzel, N., Sportelli, L. & Bertolino, A. Lessons learned from parsing genetic risk for schizophrenia into biological pathways. *Biol. Psychiatry* **94**, 121–130 (2023).
19. Owen, M. J., Legge, S. E., Rees, E., Walters, J. T. R. & O'Donovan, M. C. Genomic findings in schizophrenia and their implications. *Mol. Psychiatry* **28**, 3638–3647 (2023).
20. Jablensky, A. Subtyping schizophrenia: implications for genetic research. *Mol. Psychiatry* **11**, 815–836 (2006).
21. Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022).
22. O'Donovan, M. C. & Owen, M. J. The implications of the shared genetics of psychiatric disorders. *Nat. Med.* **22**, 1214–1219 (2016).

23. Chen, J. *et al.* Polygenic risk scores for subtyping of schizophrenia. *Schizophr. Res. Treatment* **2020**, 1638403 (2020).
24. Albiñana, C. *et al.* Multi-PGS enhances polygenic prediction by combining 937 polygenic scores. *Nat. Commun.* **14**, 4702 (2023).
25. Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).
26. Choi, S. W. & O'Reilly, P. F. PRSice-2: Polygenic Risk Score software for biobank-scale data. *GigaScience* **8**, (2019).
27. Zheng, Z. *et al.* Leveraging functional genomic annotations and genome coverage to improve polygenic prediction of complex traits within and between ancestries. *Nat. Genet.* **56**, 767–777 (2024).
28. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A. & Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1776 (2019).
29. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. *Bioinformatics* **36**, 5424–5431 (2021).
30. Budde, M. *et al.* A longitudinal approach to biological psychiatric research: The PsyCourse study. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **180**, 89–102 (2019).
31. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
32. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
33. Wootton, O. *et al.* Genomic insights into the shared and distinct genetic architecture of cognitive function and schizophrenia. *Sci. Rep.* **14**, 15356 (2024).
34. Andreassen, O. A. *et al.* Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. *PLoS Genet.* **9**, e1003455 (2013).
35. Pruessner, M., Cullen, A. E., Aas, M. & Walker, E. F. The neural diathesis-stress model of schizophrenia revisited: An update on recent findings considering illness stage and

neurobiological and methodological complexities. *Neurosci. Biobehav. Rev.* **73**, 191–218 (2017).

36. Lederbogen, F., Haddad, L. & Meyer-Lindenberg, A. Urban social stress--risk factor for mental disorders. The case of schizophrenia. *Environ. Pollut.* **183**, 2–6 (2013).
37. Lederbogen, F. *et al.* City living and urban upbringing affect neural social stress processing in humans. *Nature* **474**, 498–501 (2011).
38. Antonucci, L. A. *et al.* Personality changes during adolescence predict young adult psychosis proneness and mediate gene-environment interplays of schizophrenia risk. *Psychol. Med.* **54**, 1–11 (2024).
39. Farmer, A. *et al.* Neuroticism, extraversion, life events and depression: The Cardiff Depression Study. *Br. J. Psychiatry* **181**, 118–122 (2002).
40. Kendler, K. S., Kuhn, J. & Prescott, C. A. The interrelationship of neuroticism, sex, and stressful life events in the prediction of episodes of major depression. *Am. J. Psychiatry* **161**, 631–636 (2004).
41. Nagel, M. *et al.* Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nat. Genet.* **50**, 920–927 (2018).
42. Luciano, M. *et al.* Association analysis in over 329,000 individuals identifies 116 independent variants influencing neuroticism. *Nat. Genet.* **50**, 6–11 (2018).
43. Kappel, D. B. *et al.* Genomic stratification of clozapine prescription patterns using schizophrenia polygenic scores. *Biol. Psychiatry* **93**, 149–156 (2023).
44. Uellendahl-Werth, F. *et al.* Cross-tissue transcriptome-wide association studies identify susceptibility genes shared between schizophrenia and inflammatory bowel disease. *Commun. Biol.* **5**, 80 (2022).
45. Gong, W. *et al.* Role of the gut-brain axis in the shared genetic etiology between gastrointestinal tract diseases and psychiatric disorders: A genome-wide pleiotropic analysis. *JAMA Psychiatry* **80**, 360–370 (2023).
46. Tylee, D. S. *et al.* An atlas of genetic correlations and genetically informed associations

linking psychiatric and immune-related phenotypes. *JAMA Psychiatry* **79**, 667–676 (2022).

47. Peterson, R. E. *et al.* Genome-wide association studies in ancestrally diverse populations: Opportunities, methods, pitfalls, and recommendations. *Cell* **179**, 589–603 (2019).
48. Fatumo, S., Chikowore, T., Choudhury, A. & Ayub, M. A roadmap to increase diversity in genomic studies. *Nat. Med.* (2022).
49. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).
50. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
51. Malone, J. *et al.* Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* **26**, 1112–1118 (2010).
52. Marees, A. T. *et al.* A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *Int. J. Methods Psychiatr. Res.* **27**, e1608 (2018).
53. Choi, S. W., Mak, T. S.-H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.* **15**, 2759–2772 (2020).
54. Magno, R. & Maia, A.-T. gwasrapidd: an R package to query, download and wrangle GWAS catalog data. *Bioinformatics* **36**, 649–650 (2020).
55. HannahVMeyer. *Meyer-Lab-cshl/plinkQC: plinkQC 0.3.2*. (Zenodo, 2020). doi:10.5281/ZENODO.3934294.
56. Lam, M. *et al.* RICOPILI: Rapid imputation for COnsortias PIpeLIne. *Bioinformatics* **36**, 930–933 (2020).
57. Gnanadesikan, R. & Kettenring, J. R. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics* **28**, 81 (1972).
58. Leisch, F. Visualizing cluster results using package flexclust and friends. 116 (2009).
59. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).

60. Wright, M. N. & Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *arXiv [stat.ML]* (2015).
61. Hubert, L. & Arabie, P. Comparing partitions. *J. Classif.* **2**, 193–218 (1985).
62. Rammstedt, B. & John, O. P. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *J. Res. Pers.* **41**, 203–212 (2007).
63. The R Project for Statistical Computing. http://www.R-project.org/.
64. The R Project for Statistical Computing. http://www.R-project.org/.
65. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
66. Smeland, O. B. *et al.* Discovery of shared genomic loci using the conditional false discovery rate approach. *Hum. Genet.* **139**, 85–94 (2020).
67. Ho, D. E., Imai, K., King, G. & Stuart, E. A. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *J. Stat. Softw.* **42**, (2011).
68. Viechtbauer, W. Conducting Meta-Analyses inRwith themetaforPackage. *J. Stat. Softw.* **36**, (2010).

Supplementary texts

Filtering criteria

The following filtering was used to include summary statistics in the analyses:

- European ancestry samples
- Total sample size > 10,000
- Summary statistics containing at least 250,000 SNPs
- The publication date is newer than January 1, 2018.
- Harmonized version.
- Number of cases > 5,000

Pre-imputation and imputation

FinnGen

FinnGen is a public-private partnership project between Finnish research institutes and biobanks and private pharmaceutical companies. It leverages comprehensive Finnish digital health care registers combined with genotypic data from 500k individuals. The diagnoses in the registers are based on ICD-10, ICD-9 and ICD-8 criterion.
Samples were genotyped using Affymetrix (ThermoFisher Scientific) and Illumina arrays. Imputation was performed using a reference panel specific to the Finnish population (SISu v.4) based on 8,554 whole-genome-sequenced individuals resulting in 21 million imputed variants for each participant. Sample QC involved the removal of participants whose genetic sex did not match the ones provided in the registers, with a variant missingess of >2%, heterozygosity of more than 4 SD from the mean in common variant (MAF > 0.05) or excess relatedness to other samples (more than 500 relatives with PI_HAT > 0.1). Genetic variants were pruned out if their allele frequencies significantly differed from the reference panel (p < 5e-8) adjusted for the first 10 principal components, it they had HWE p < 1e-10 or 15% of the batches had a missinges > 0.04. Additionally, variants were removed within a batch if HWE p-value was below 1e-6 or missingess was below 0.02. Each genotyping batch consisted of ~5000 participants.

PsyCourse Study SCZ - Healthy control

The pre-imputation genotype data provided by the PsyCourse study included 428,904 genome-wide SNPs (GSA chip) from 1,600 individuals diagnosed with bipolar disorder, SCZ, schizoaffective disorder, schizophreniform disorder and HC. From this dataset, 26 related individuals were excluded, and SCZ patients (ICD-10 Schizophrenia and Schizophrenia according to DSM-IV-TR) and HC born in European countries (PsyCourse codebook variable : v1_cntr_brth) were selected, resulting in a final dataset of 755 individuals.

In the first QC-cycle of the imputation, Ricopili's pre-imputation module with default parameters was executed. Thus, 500 SNPs and one individual failing to meet the criteria were excluded from the dataset. In the next step, Ricopili's PCA module using default parameters was executed to identify the outlier subjects. 31 outlier subjects were identified and excluded that exceeded ±0.1 of the first two PCs. In the second QC-cycle, Ricopili's pre-imputation module with default parameters was executed again and 84 SNPs failing to meet the criteria were excluded from the dataset. The final pre-imputation dataset included 723 subjects and 428,320 SNPs.

As the final data met quality control criteria (lambda1000 = 1.014 and no genome-wide significant SNPs were found), imputation was performed using the 1000 Genomes European population as reference. Once the imputation was completed, SNPs with info scores above 0.8 (n = 4,883,969) were selected for post-imputation QC using our in-house genotype data pre-processing module.

After completing the pre-processing module, the imputed and quality-controlled dataset consisted of 3,470,953 SNPs and 725 individuals. That is, 1 individual (PSYCOURSE_0708) and 1,413,016 SNPs, which did not meet the criteria specified in the genotype pre-processing module, were removed from the post-imputation dataset. The final data included 721 individuals.

Bari Cohort

Subjects with European ancestry, comprising both healthy controls and individuals diagnosed with SCZ or schizoaffective disorder, were recruited from the Apulia region in Italy. The recruitment process followed the ethical standards outlined in the World Medical Association's Declaration of Helsinki and was approved by the local ethics committee ("Comitato Etico Indipendente Locale - Azienda Ospedaliero-Universitaria Consorziale Policlinico di Bari"). Healthy controls were screened using the Non-Patient Structured Clinical Interview for DSM-IV to ensure they were unaffected by any psychiatric condition. Diagnosis of SCZ was made using the Structured Clinical Interview for the DSM-IV, Axis 1 disorders by board-certified psychiatrists. Participants were excluded if they had a significant history of drug or alcohol abuse, active substance use within the past year, a history of head trauma with loss of consciousness, or any other serious medical condition. Genotype data from Bari was collected over four years: 2015, 2016, 2017, and 2019. Genotyping for all participants was conducted using the Illumina HumanOmni2.5-8 v1 BeadChip platform. From the pre-imputation genotype data, subjects with SCZ and HC were selected from all batches. SNP overlaps between batches were then compared, and those with the highest overlap were combined, and imputed together. Accordingly, the batches from 2015 (557 individuals, 1,258,881 SNPs) and 2019 (154 individuals, 1,454,515 SNPs) were combined and imputed together. Similarly, the batches from 2016 (139 individuals, 1,424,791 SNPs) and 2017 (501 individuals, 1,431,583 SNPs) were merged and imputed together.

Dataset 1 (2015 together with 2019)

Pre-imputation data consisted of 1,197,491 overlapping SNPs between two batches and 711 individuals from Italy diagnosed with SCZ and HC.

In the first QC-cycle, Ricopili's pre-imputation module with default parameters (except minor allele frequency (MAF) > 0.02, because the lambda1000 value was high due to the genetic inflation of SNPs with MAF values lower than this.) was executed. Thus, 59,089 SNPs failing to meet the criteria were excluded from the dataset. In the next step, Ricopili's PCA module using default parameters was executed to identify the related and outlier individuals. In total, 34 subjects were excluded from the data which were identified as related (pi-hat > 0.2). The final pre-imputation dataset included 677 subjects and 1,138,402 SNPs.

As the final data met quality control criteria (lambda1000 = 1.022 and one genome-wide significant SNPs was removed), imputation was performed using the 1000 Genomes European population as reference. Once the imputation was completed, SNPs with info scores above 0.8 (n = 7,892,583) were selected for merging with the Dataset 2.

Dataset 2 (2016 together with 2017)

Pre-imputation data consisted of 1,357,782 overlapping SNPs between two batches and 640 individuals from Italy diagnosed with SCZ and HC.

In the first QC-cycle, Ricopili's pre-imputation module with default parameters (except MAF > 0.02, because the lambda1000 value was high due to the genetic inflation of SNPs with MAF values lower than this.) was executed. Thus, 57,412 SNPs failing to meet the criteria were excluded from the dataset. In the next step, Ricopili's PCA module using default parameters was executed to identify the related and outlier individuals. In total, 21 subjects were excluded from the data which were identified as related (pi-hat > 0.2). The final pre-imputation dataset included 619 subjects and 1,300,370 SNPs.

As the final data met quality control criteria (lambda1000 = 0.979 and one genome-wide significant SNPs was removed), imputation was performed using the 1000 Genomes European population as reference. Once the imputation was completed, SNPs with info scores above 0.8 (n = 8,147,504) were selected for merging with the Dataset 1.

Merging Dataset 1 and Dataset 2

Merging Dataset 1 and Dataset 2 resulted in a final dataset with 7,458,687 overlapping SNPs and 1,296 subjects. Post-imputation QC was performed using our in-house genotype data pre-processing module.

After completing the pre-processing module, the imputed and quality-controlled dataset consisted of 5,432,607 SNPs and 1273 individuals. That is, 23 individuals and 2,026,080 SNPs, which did not meet the criteria specified in the genotype pre-processing module, were removed from the post-imputation dataset. Four subjects whose deviations exceeded four

times the standard deviation based on PCs 1 and 2 were excluded, resulting in a final sample of 1,269 subjects for analysis.

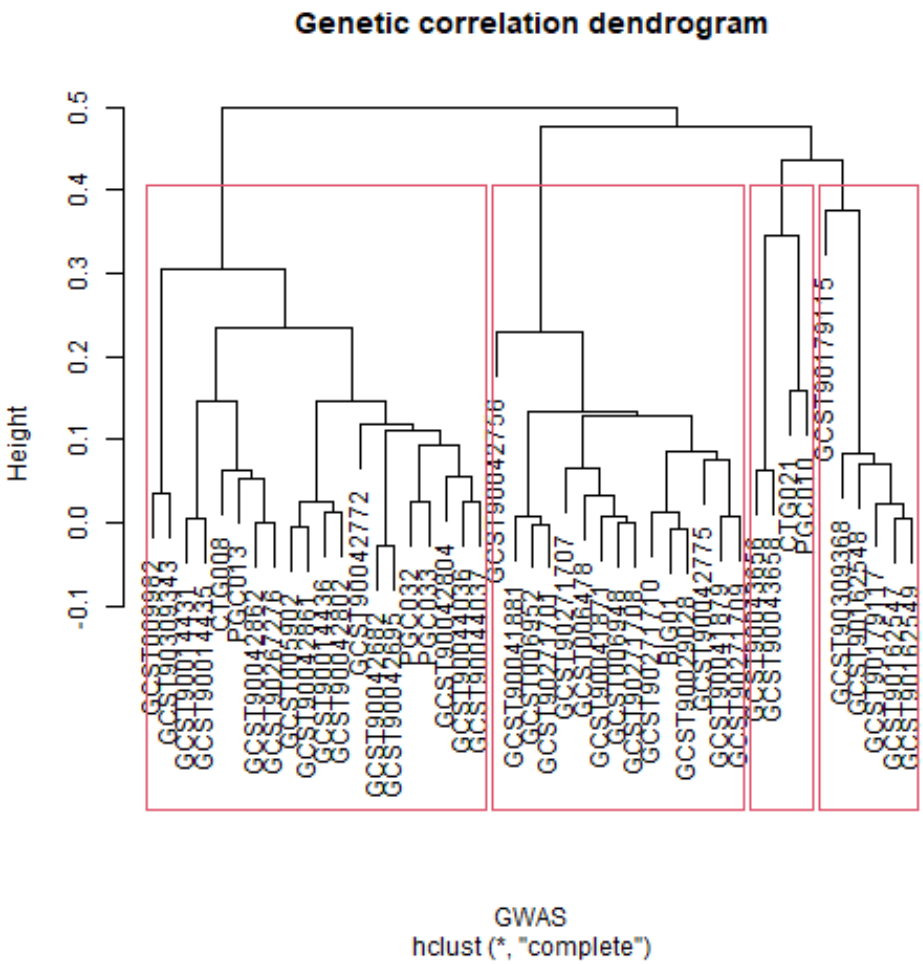
Intersecting SNPs across datasets

To minimize bias due to the differences in cohorts, only the SNPs that overlap across three datasets were included in the PGS prediction. That is, 2,925,749 SNPs were included from each dataset.

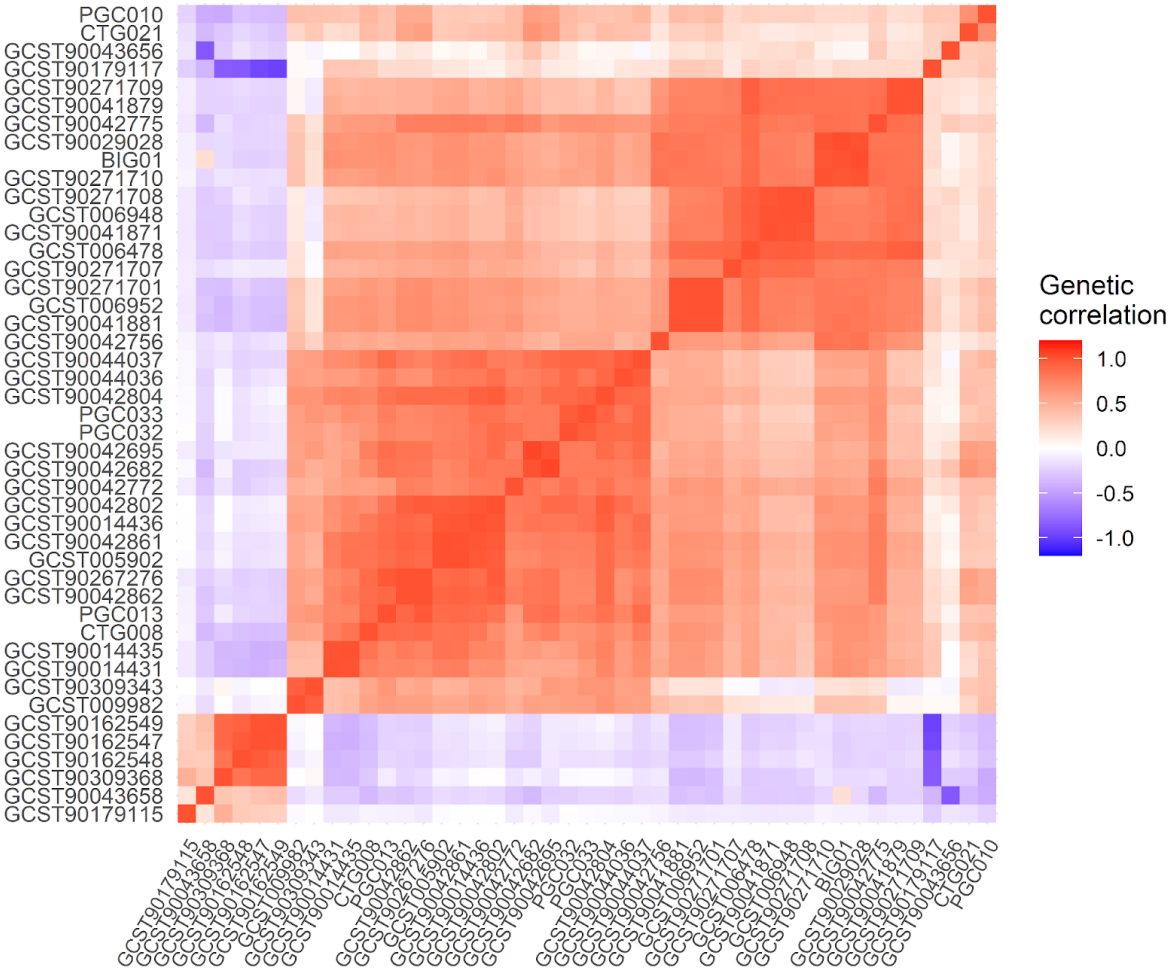
PGS Calculation Parameters

P-value thresholds : $5x10^{-8}$, $1x10^{-7}$, $1x10^{-6}$, $1x10^{-5}$, $1x10^{-4}$, 0.001, 0.01, 0.05, 0.1, 0.5, 1 (--bar-levels)
Clumping : 500 (--clump-kb), 1 (--clump-p), 0.2 (clump-r2)
MAF filtering : 0.01 (--base-maf), MAFs of the SNPs in a GWAS summary statistics were calculated from 1000 Genome European population reference dataset.
LD reference file : 1000 Genome European population reference dataset.

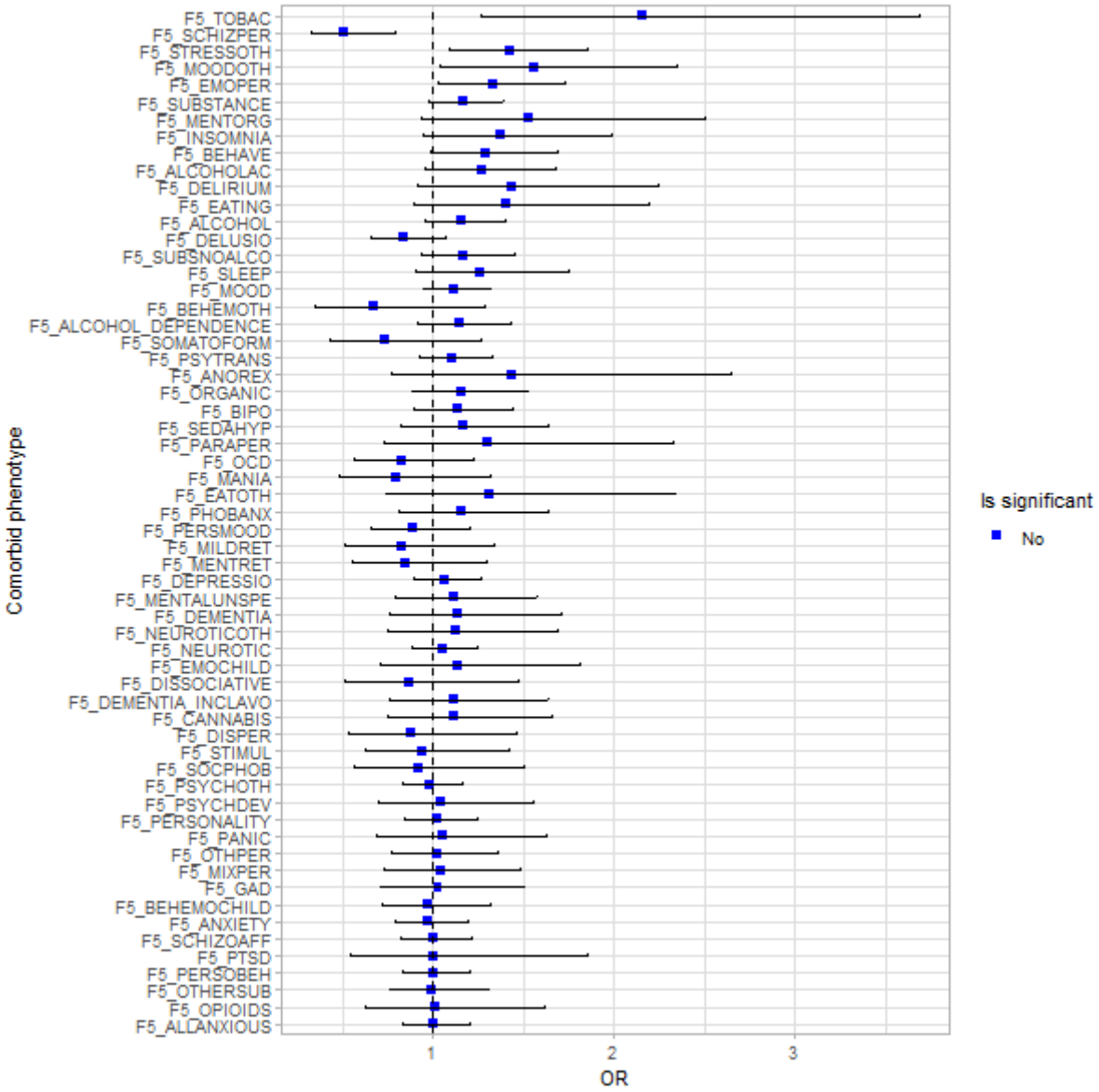
Supplementary Figures



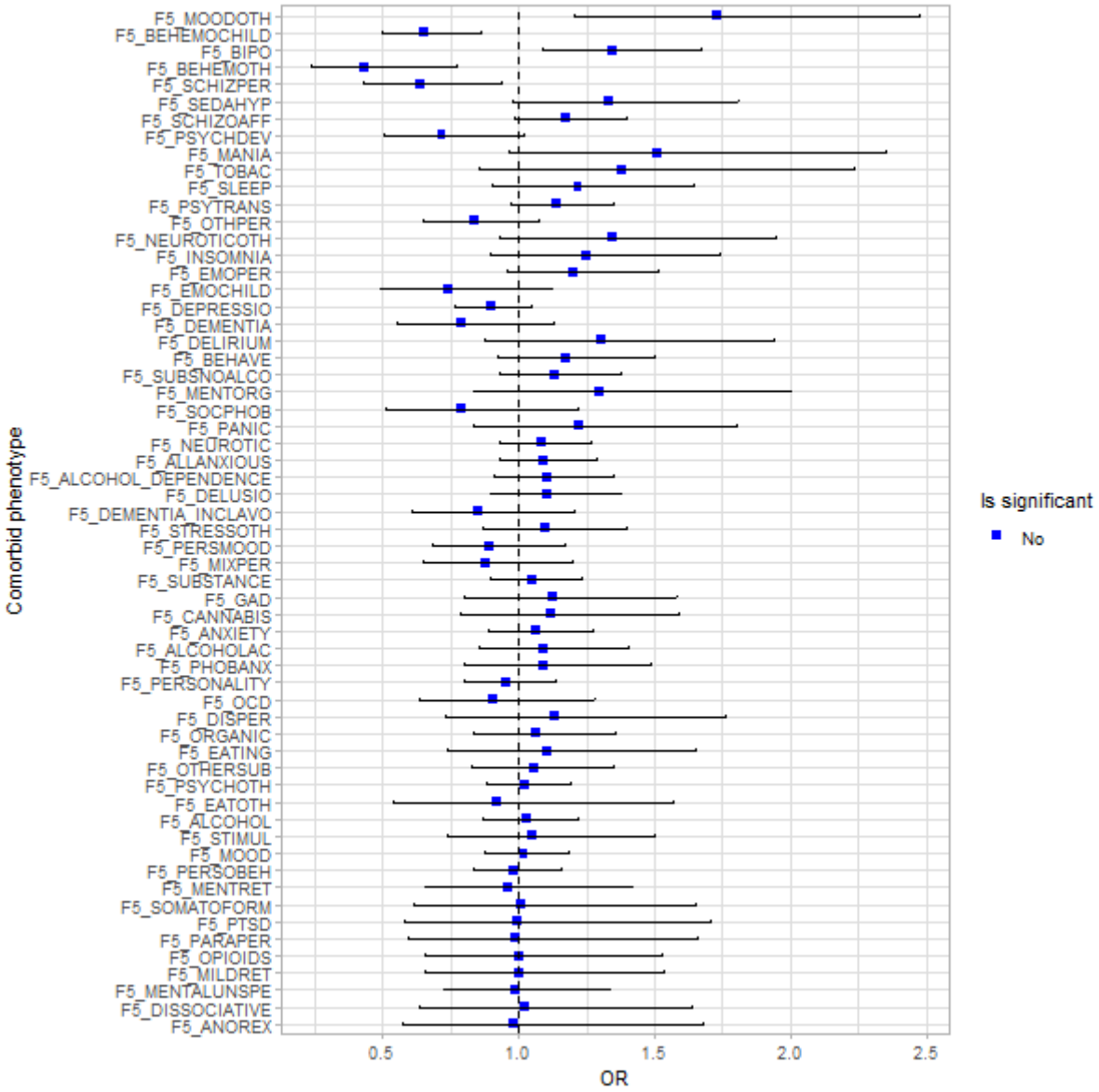
Supplementary Figure 1. The dendrogram of the hierarchical clustering of the absolute genetic correlation matrix. The group second on the right contains SCZ itself, a gene-environment interaction GWAS and two measures related to alcohol use. For clearer interpretation of selected downstream analyses characterizing the contribution of the different PGS-groups, we investigated SCZ as a single PGS.



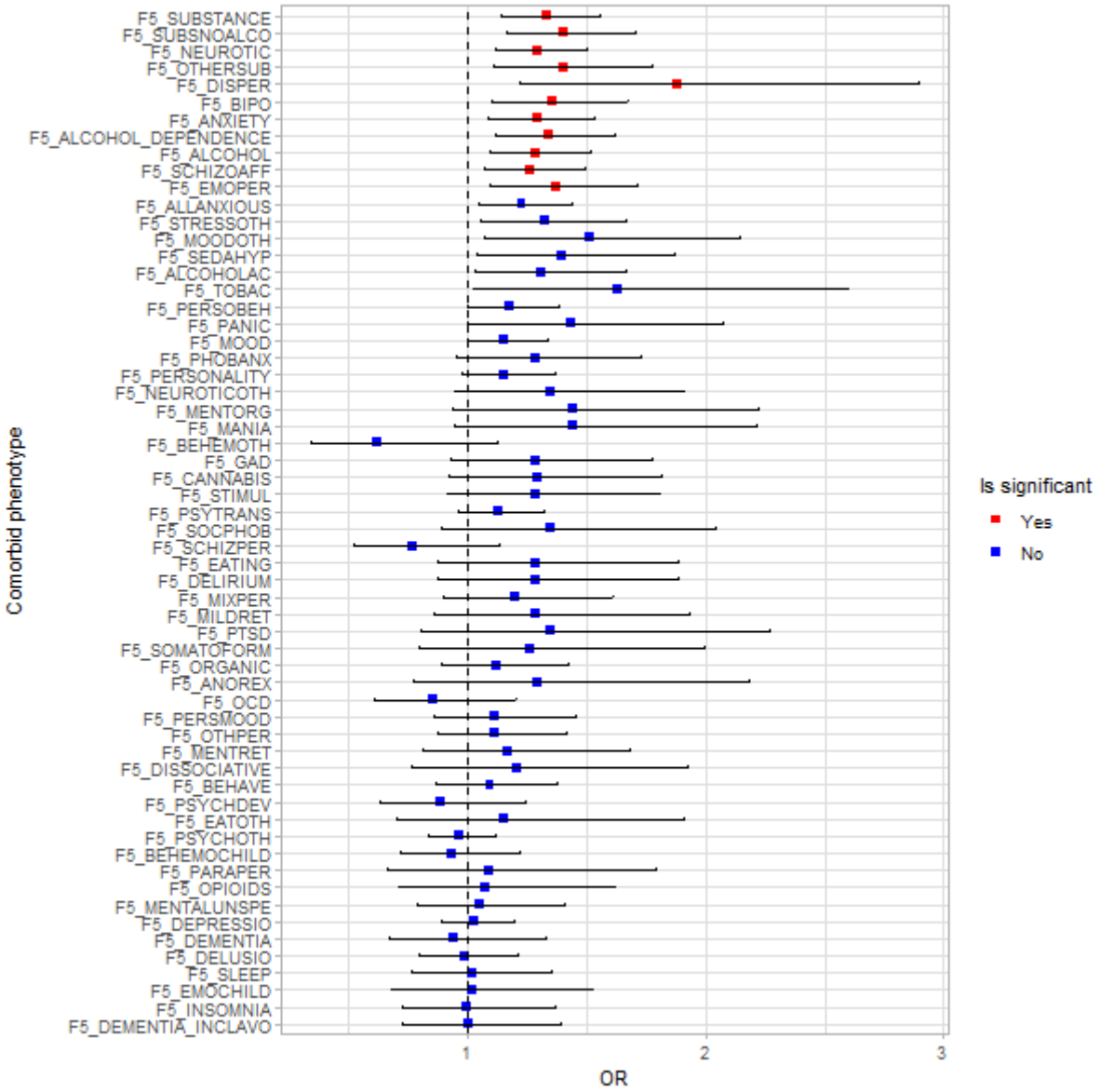
Supplementary Figure 2. Genetic correlations between the top 45 source GWAS. Note that Supplementary Figure 21 is not based on the genetic correlation values plotted here but on the absolute values.



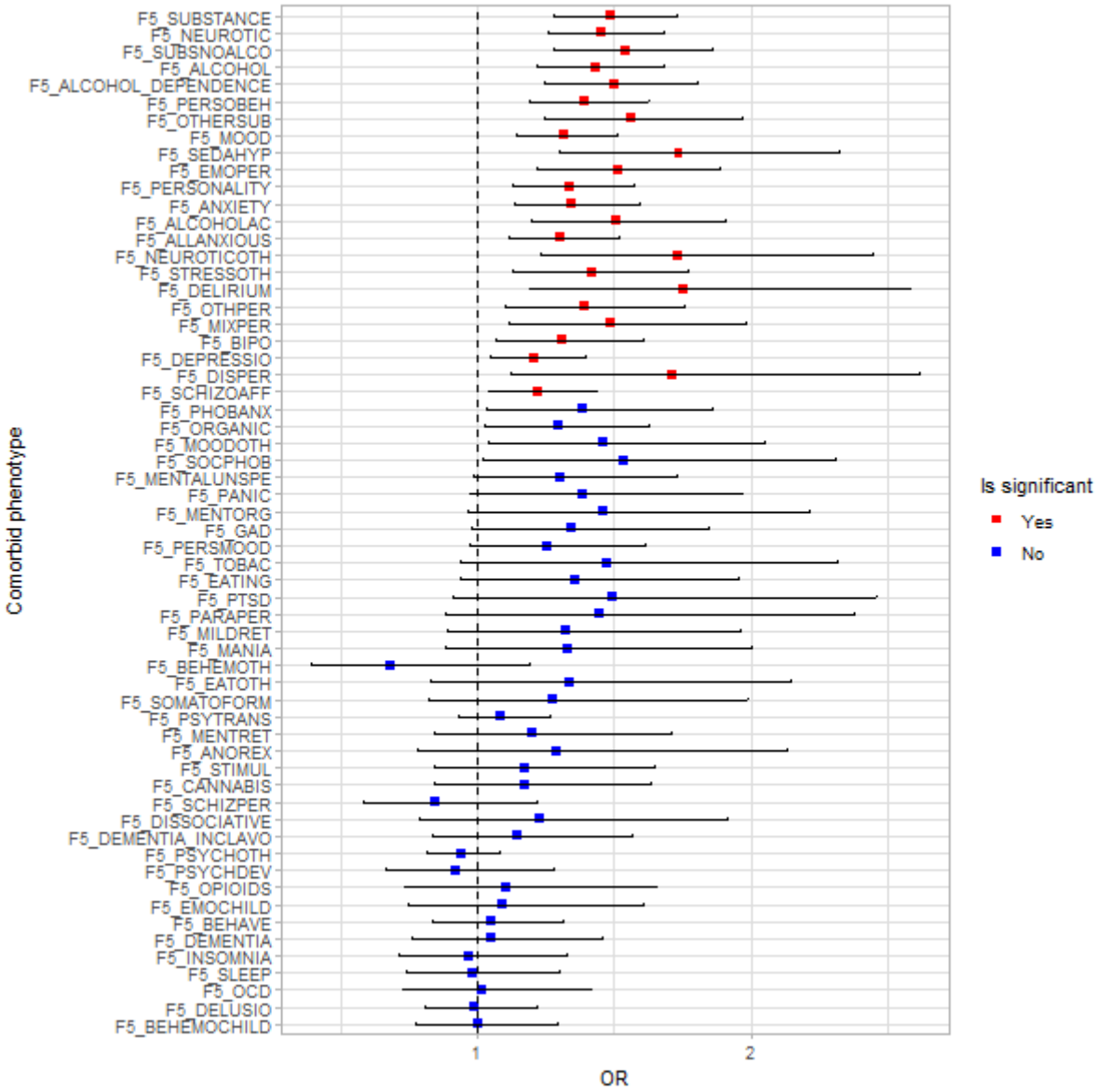
Supplementary Figure 3. Feature size 1: Differences in comorbidity rates of mental disorders and disorder categories between the biotypes. Red color indicates statistical significance at FDR of 0.05. The error bars show 95 % confidence intervals.



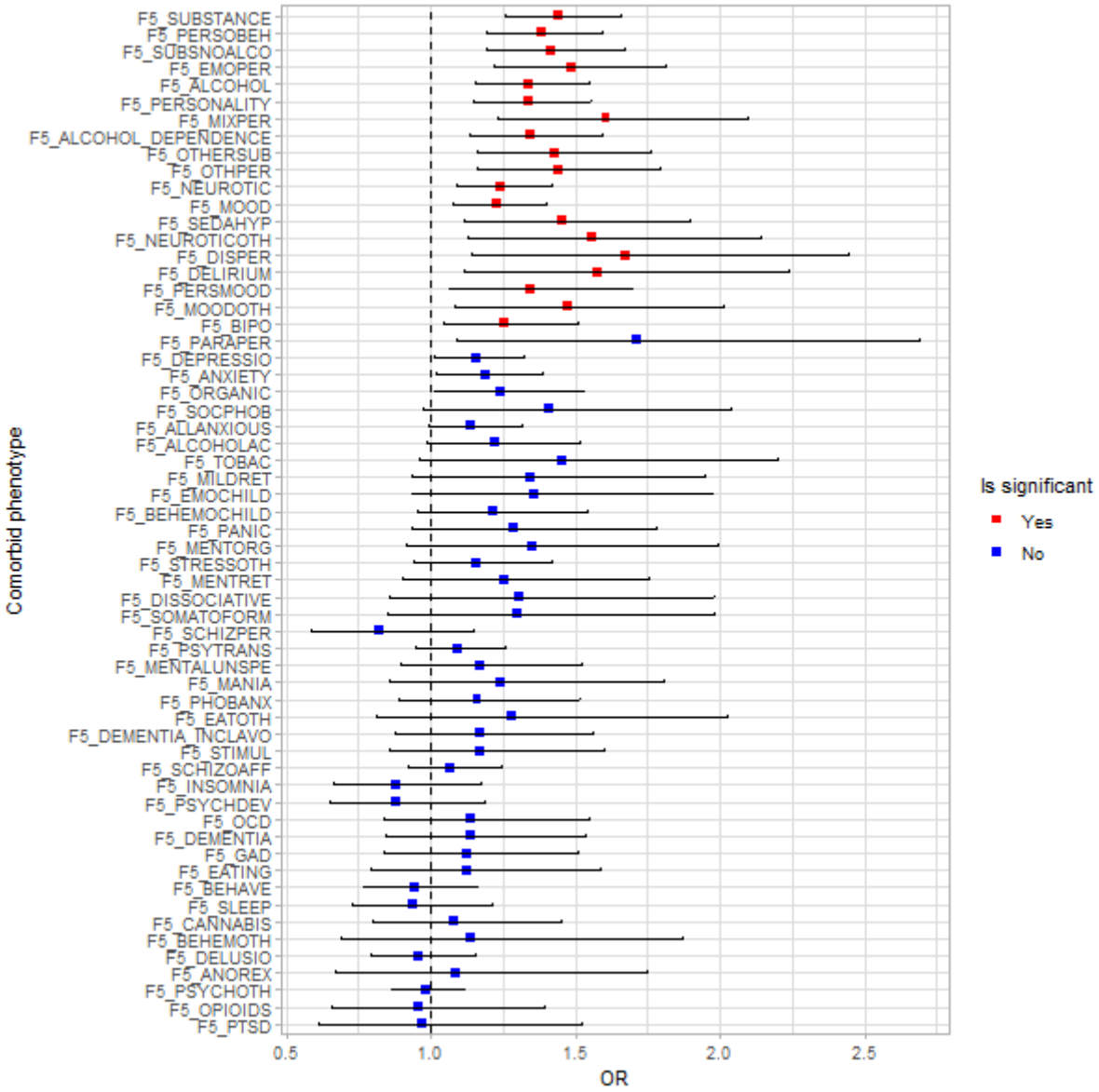
Supplementary Figure 4. Feature size 2: Differences in comorbidity rates of mental disorders and disorder categories between the biotypes. Red color indicates statistical significance at FDR of 0.05. The error bars show 95 % confidence intervals.



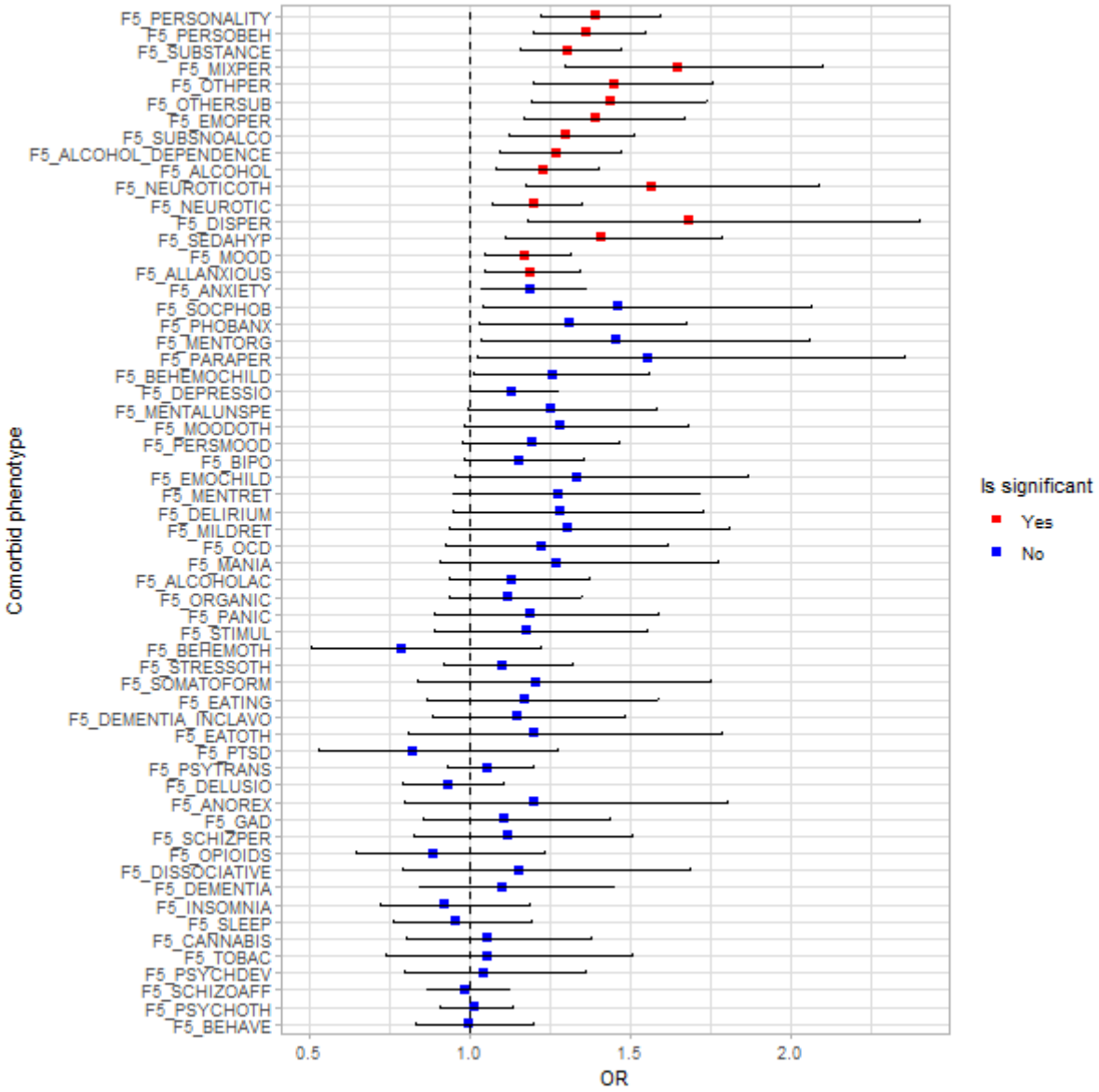
Supplementary Figure 5. Feature size 3: Differences in comorbidity rates of mental disorders and disorder categories between the biotypes. Red color indicates statistical significance at FDR of 0.05. The error bars show 95 % confidence intervals.



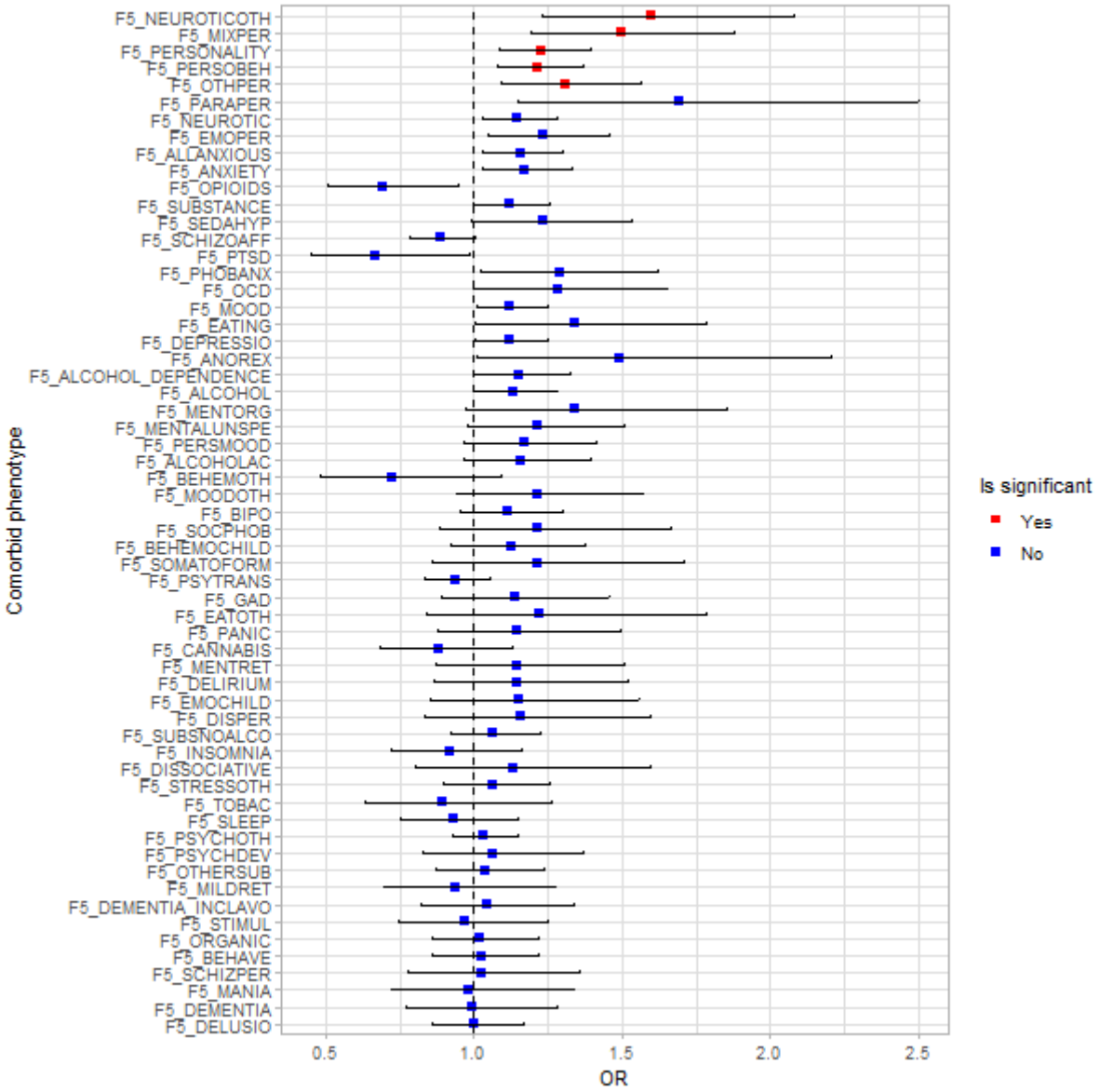
Supplementary Figure 6 . Feature size 4: Differences in comorbidity rates of mental disorders and disorder categories between the biotypes. Red color indicates statistical significance at FDR of 0.05. The error bars show 95 % confidence intervals.



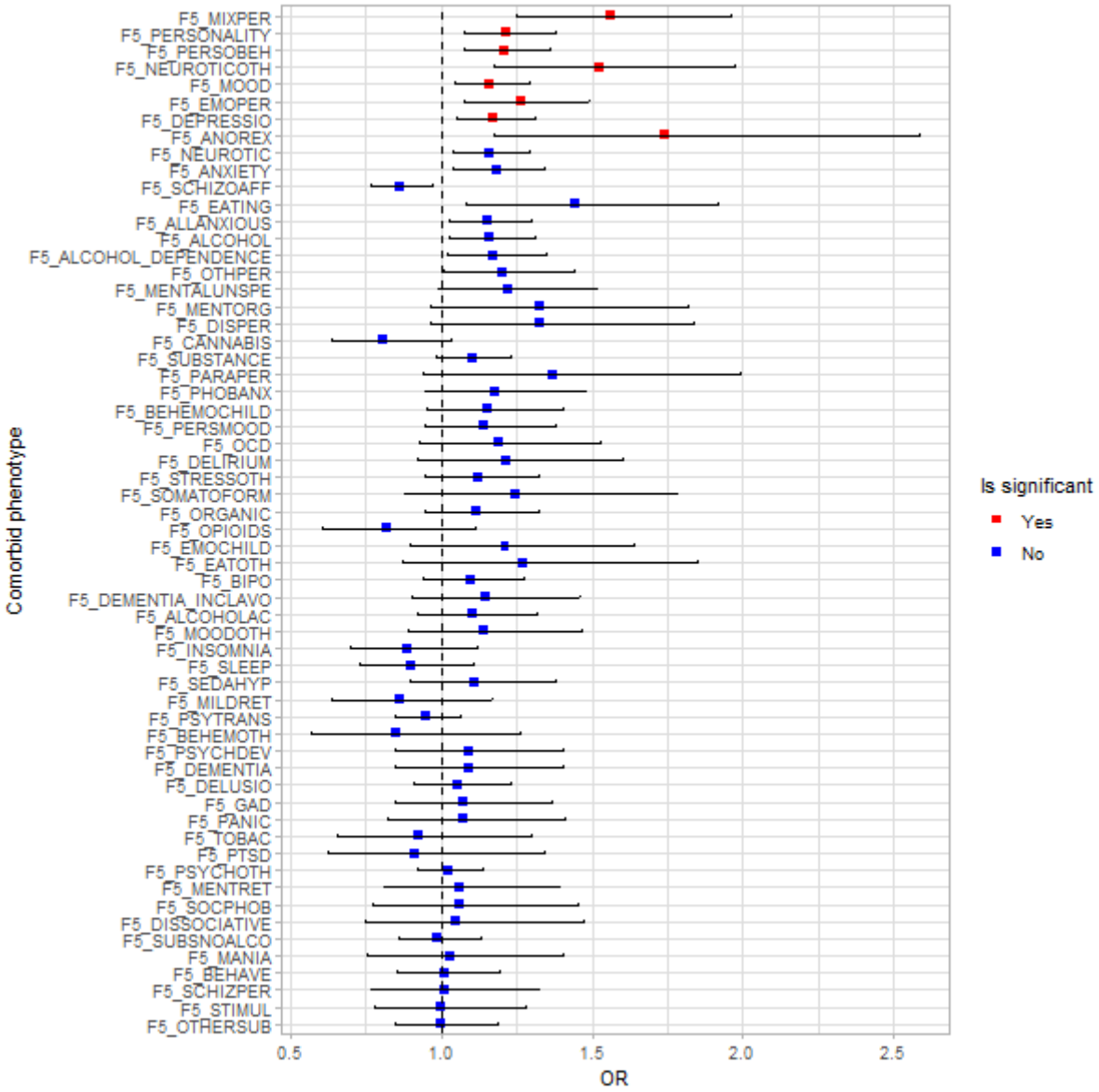
Supplementary Figure 7. Feature size 5: Differences in comorbidity rates of mental disorders and disorder categories between the biotypes. Red color indicates statistical significance at FDR of 0.05. The error bars show 95 % confidence intervals.



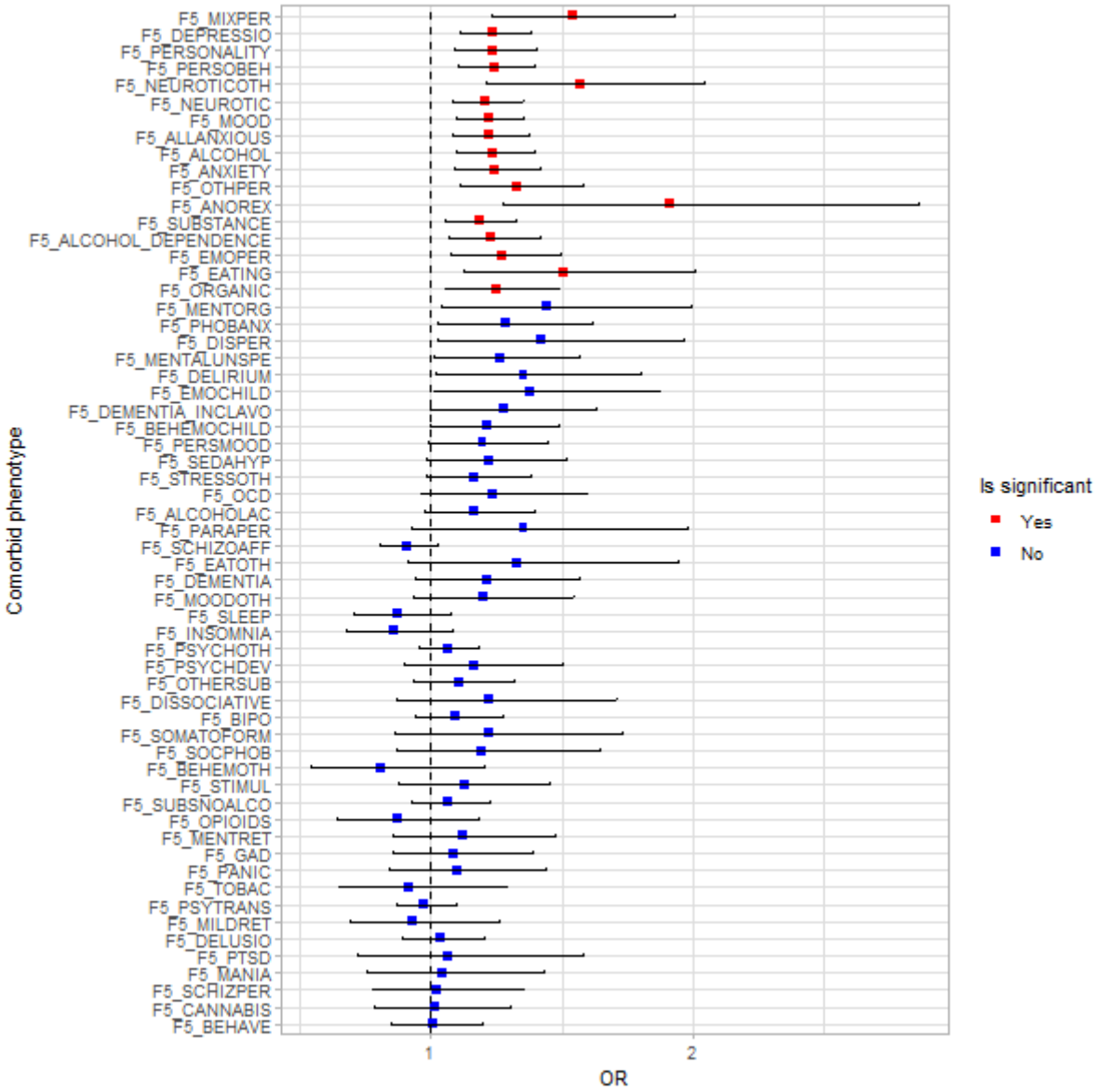
Supplementary Figure 8. Feature size 6: Differences in comorbidity rates of mental disorders and disorder categories between the biotypes. Red color indicates statistical significance at FDR of 0.05. The error bars show 95 % confidence intervals.



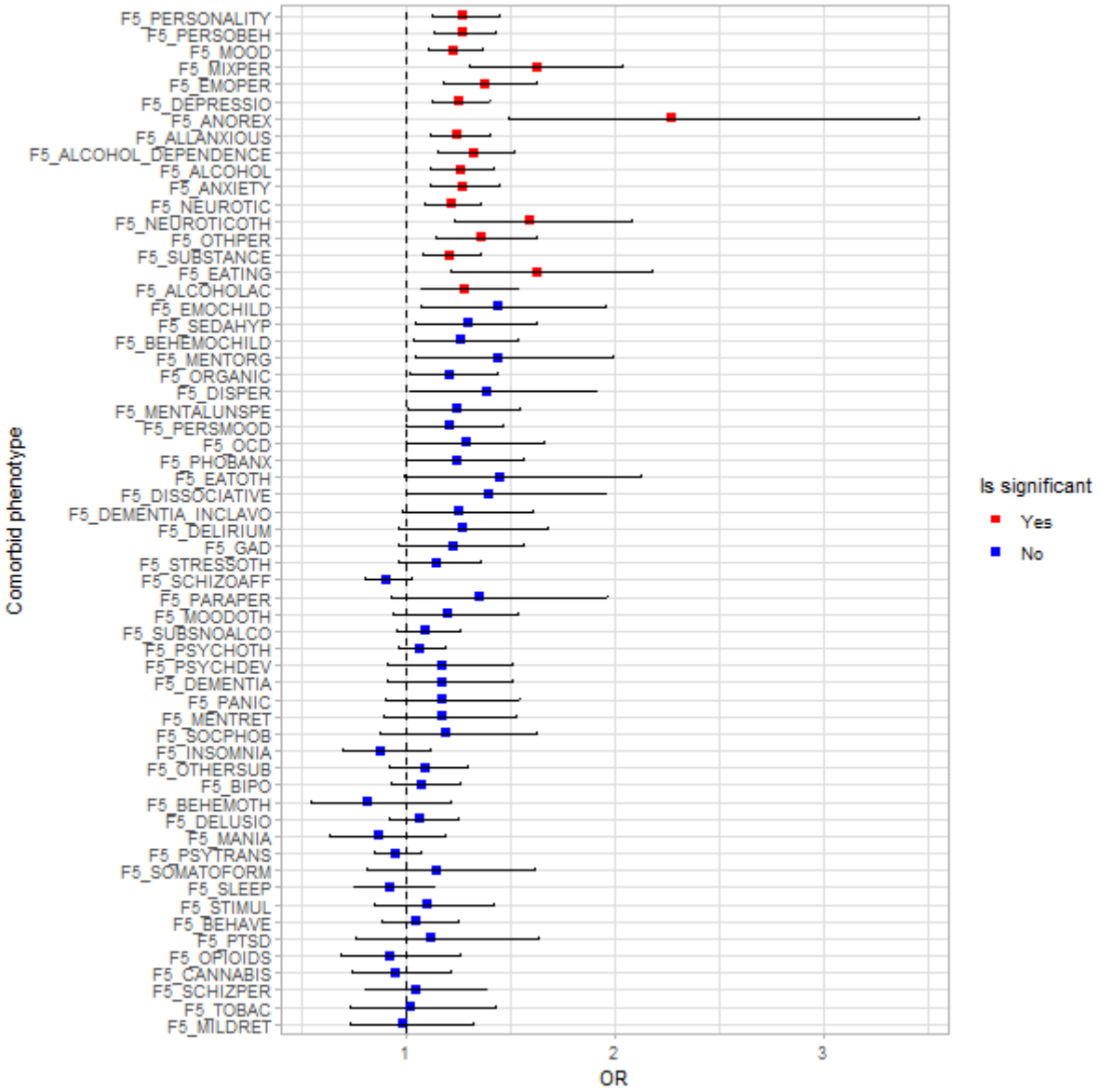
Supplementary Figure 9. Feature size 12: Differences in comorbidity rates of mental disorders and disorder categories between the biotypes. Red color indicates statistical significance at FDR of 0.05. The error bars show 95 % confidence intervals.



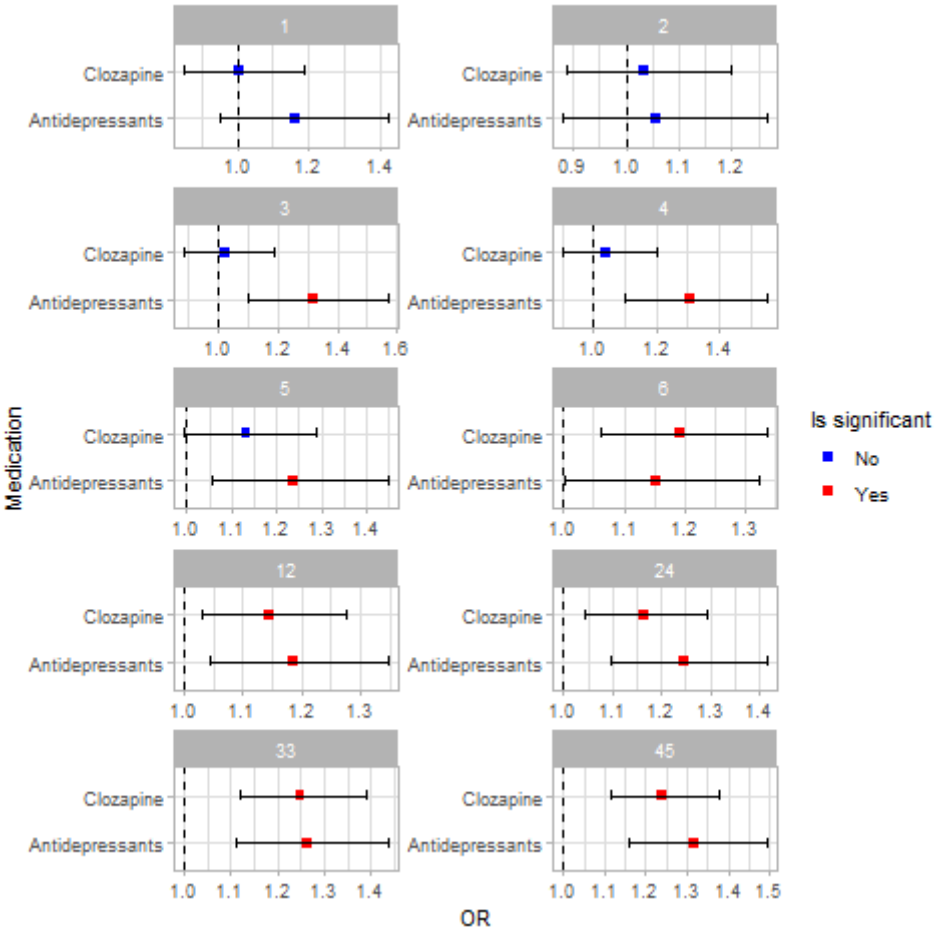
Supplementary Figure 10. Feature size 24: Differences in comorbidity rates of mental disorders and disorder categories between the biotypes. Red color indicates statistical significance at FDR of 0.05. The error bars show 95 % confidence intervals.



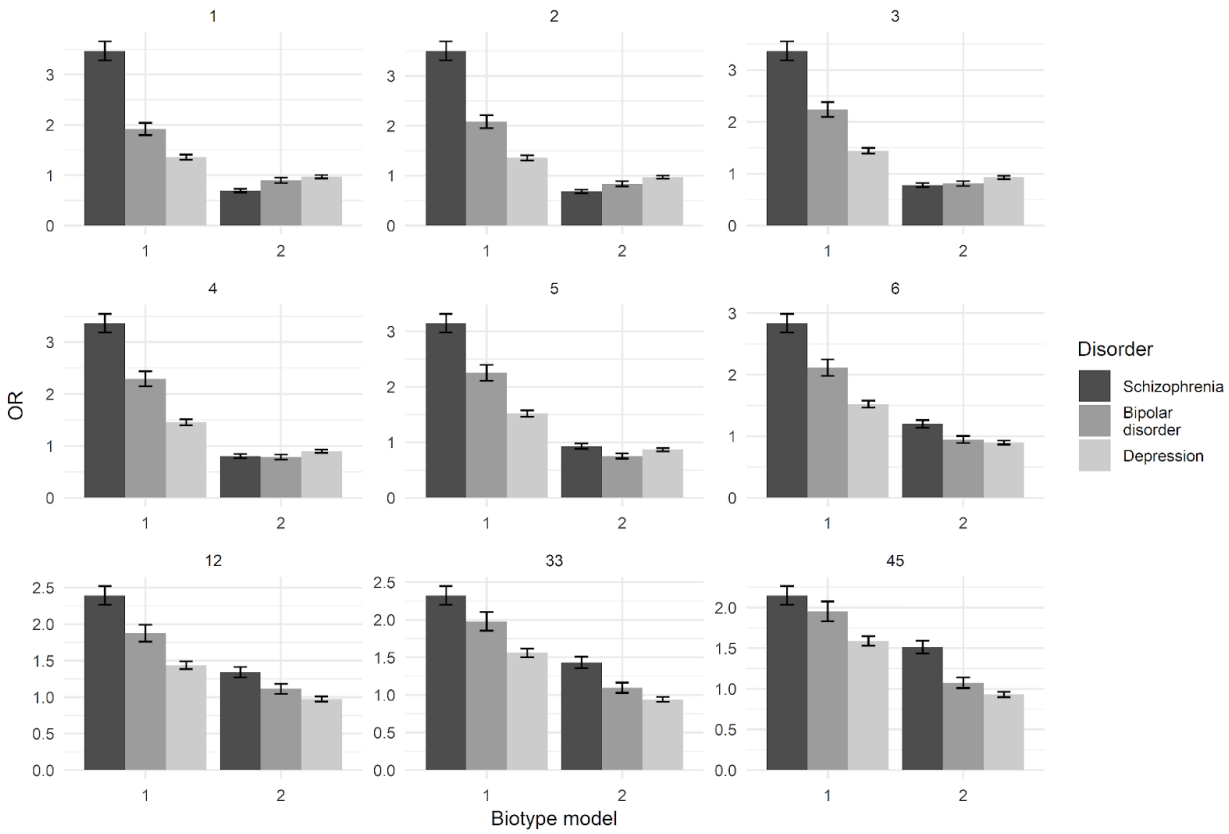
Supplementary Figure 11. Feature size 33: Differences in comorbidity rates of mental disorders and disorder categories between the biotypes. Red color indicates statistical significance at FDR of 0.05. The error bars show 95 % confidence intervals.



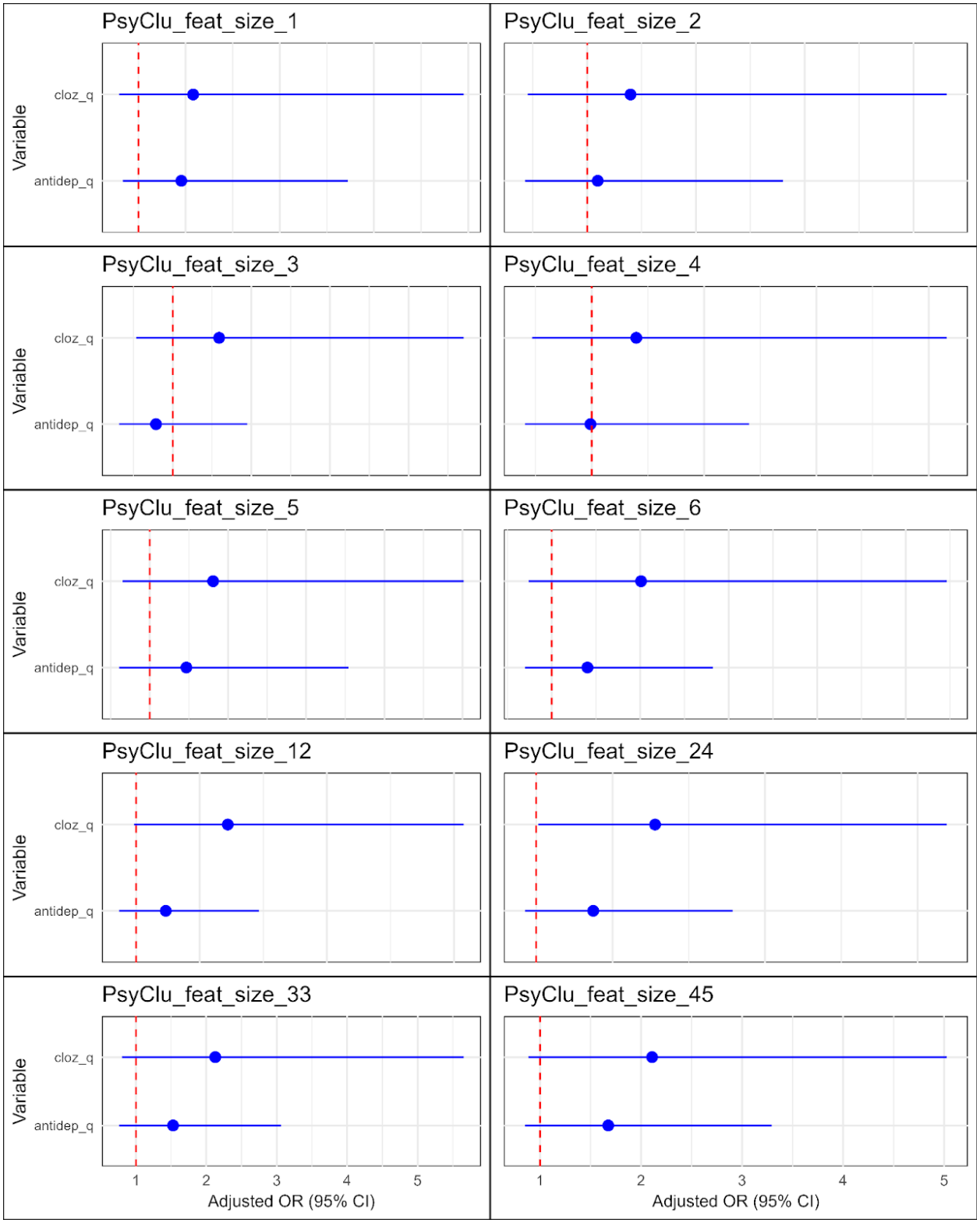
Supplementary Figure 12. Feature size 45: Differences in comorbidity rates of mental disorders and disorder categories between the biotypes. Red color indicates statistical significance at FDR of 0.05. The error bars show 95 % confidence intervals.



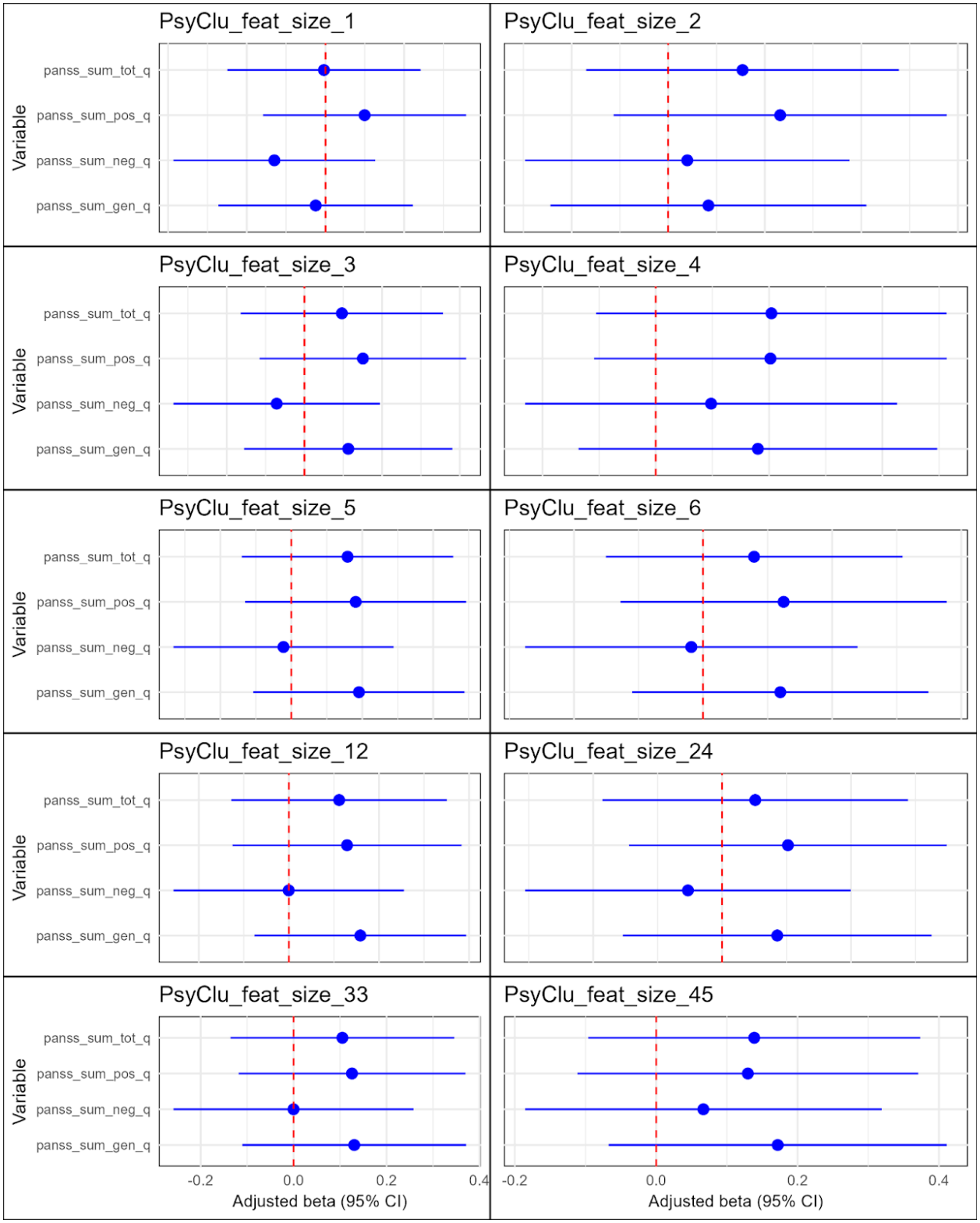
Supplementary Figure 13. Difference in Clozapine and antidepressants prescription rates between the biotypes across feature sizes.



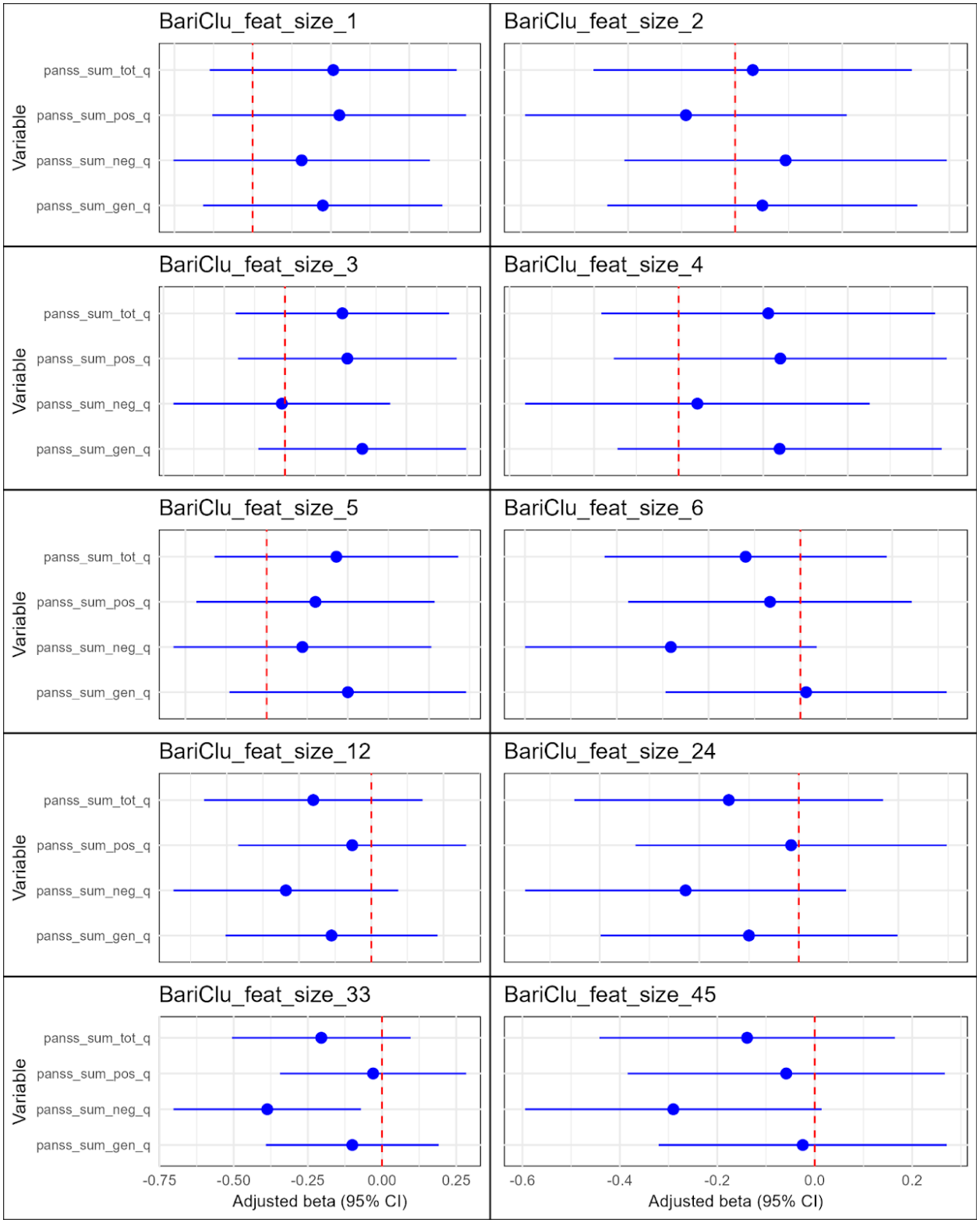
Supplementary Figure 14. Differential diagnostic associations of SCZ biotypes across feature sizes. Note that feature size 12 is depicted in Figure 5.



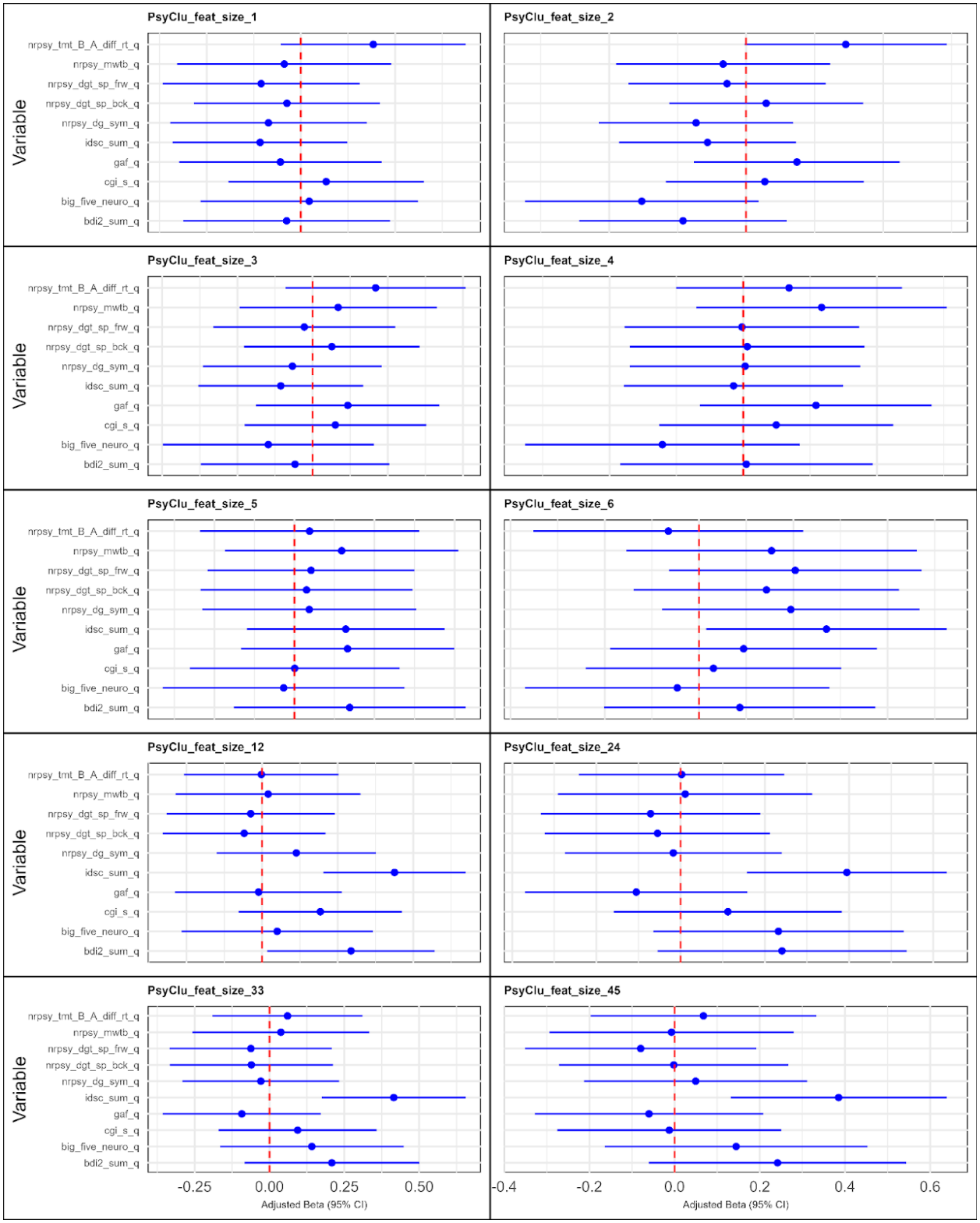
Supplementary Figure 15. Clozapine and antidepressant use association in $Val_{PsyCourse}$ across all feature sizes. The error bars show 95 % confidence intervals.



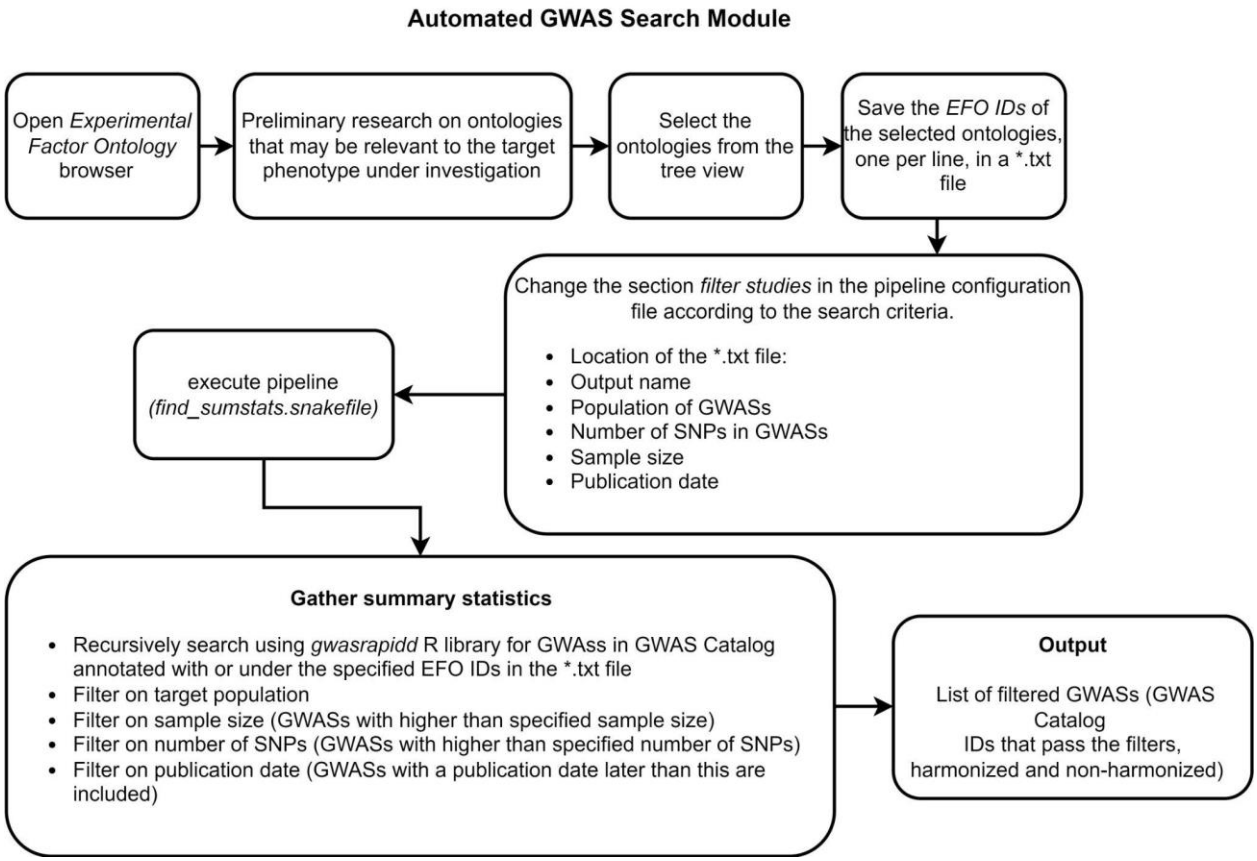
Supplementary Figure 16. PANSS cluster associations in $Val_{PsyCourse}$ across all feature sizes. The error bars show 95 % confidence intervals.



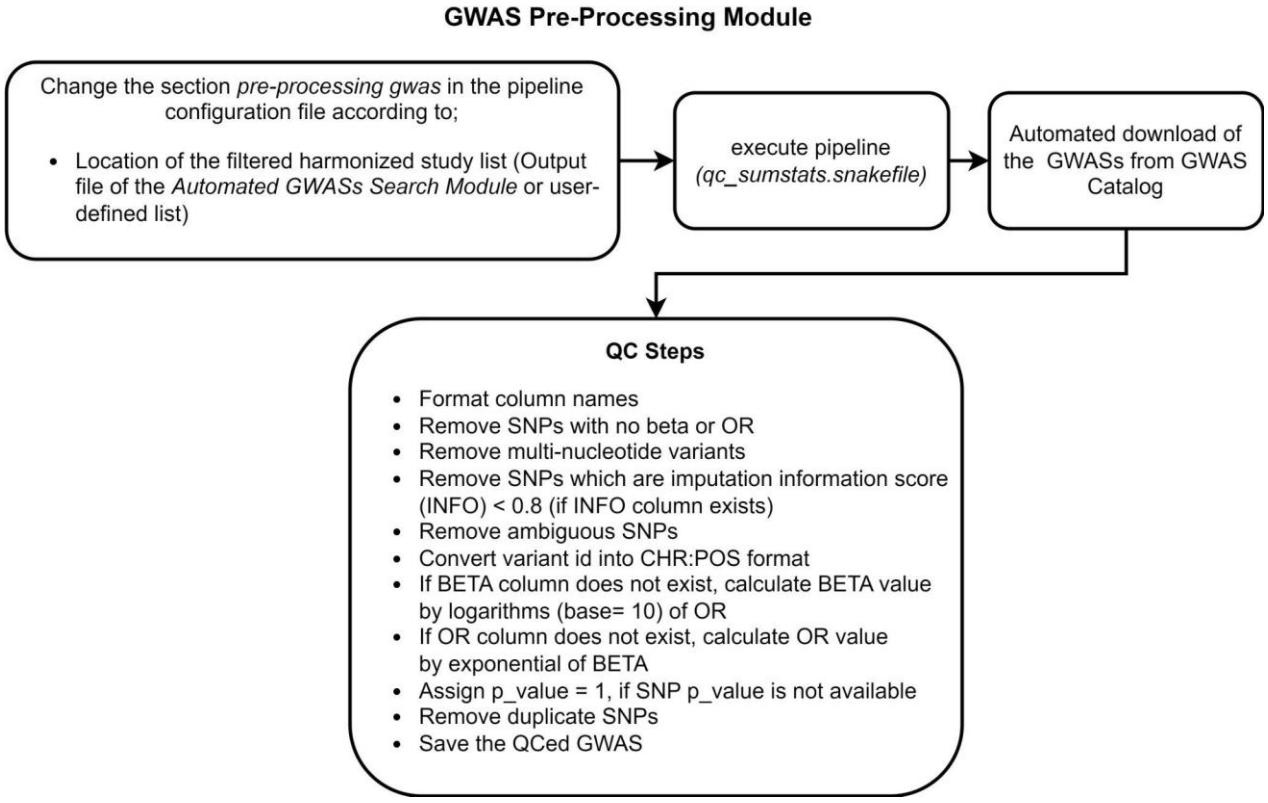
Supplementary Figure 17. PANSS cluster associations in Val_{Bari} across all feature sizes. The error bars show 95 % confidence intervals.



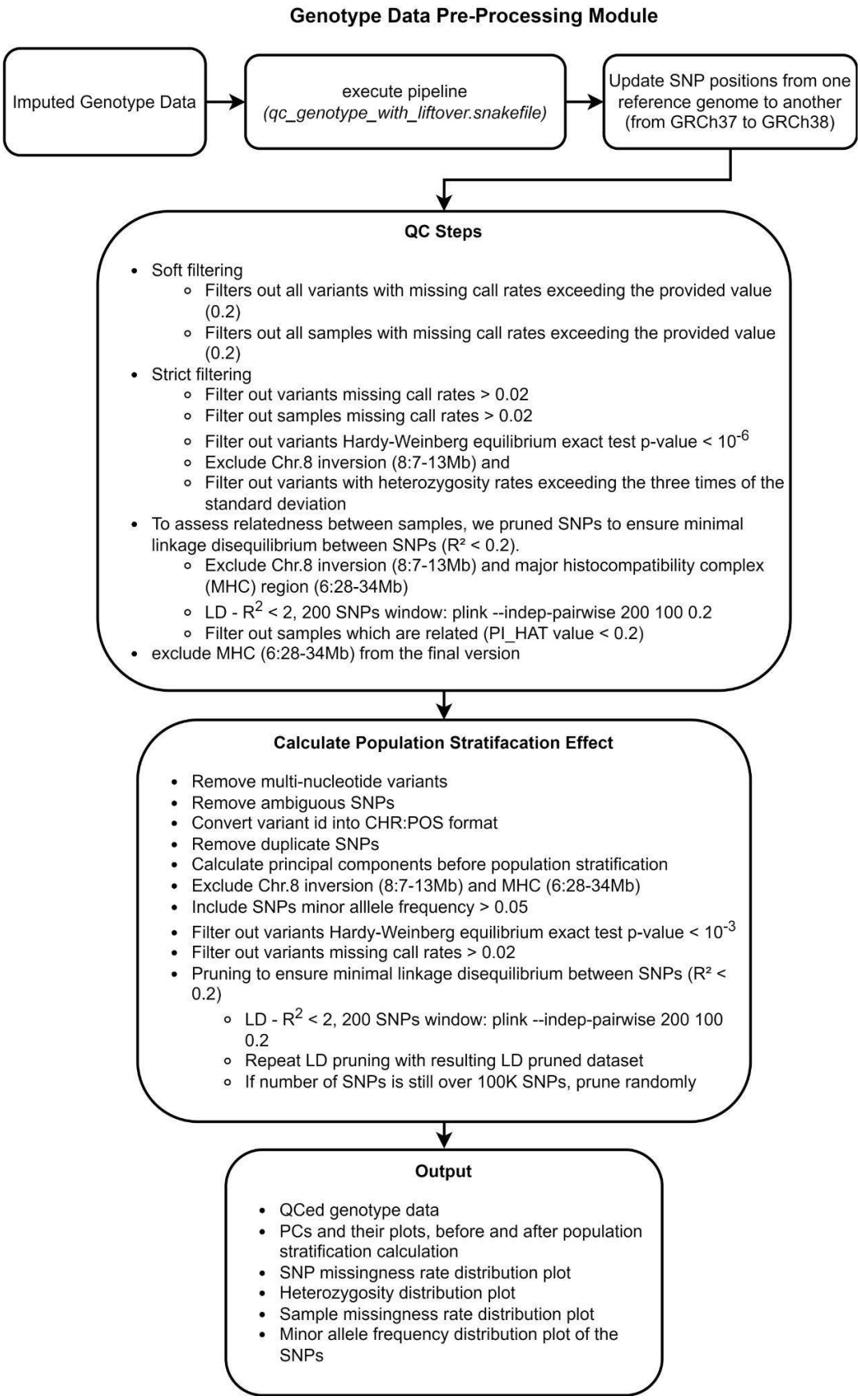
Supplementary Figure 18. Cluster associations of selected phenotypes in $Val_{PsyCourse}$ across all feature sizes. The error bars show 95 % confidence intervals.



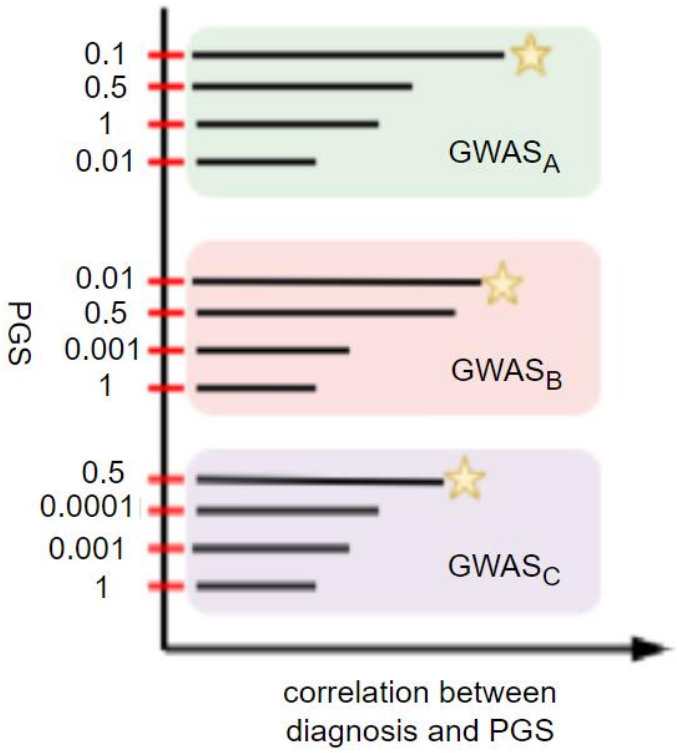
Supplementary Figure 19. **Automated GWAS search module workflow.**



Supplementary Figure 20. **Summary statistics QC module workflow.**



Supplementary Figure 21. **Genotype QC module workflow.**



Supplementary Figure 22. **Feature selection by ranking GWASs according to best PGS.**