Development and validation of polygenic risk profiles of schizophrenia

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

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, Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore, MD, USA

9, International Max Planck Research School for Translational Psychiatry (IMPRS-TP), Max-Planck-Institute of Psychiatry, Munich, Germany

10, Department of Immunology, Faculty of Medicine, Ahvaz, Iran

11, Department of Psychiatry and Psychotherapy, LMU University Hospital, Munich, Germany

12, Department Clinical Translation, Max Planck Institute of Psychiatry, Munich, Germany, Munich, Germany

13, Department of Psychiatry and Psychotherapy, Faculty of Medicine and University Hospital Bonn, Bonn, Germany

14, Institute of Human Genetics, Faculty of Medicine and University Hospital Bonn, Bonn, Germany

15, German Center for Mental Health (DZPG), partner site Munich/Augsburg, Germany

16, Centres for Psychiatry Suedwuerttemberg, Centres for Psychiatry Suedwuerttemberg, Ravensburg, Germany

17, Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, USA

18, Department of Genetic Medicine, 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

Joint senior authors

Manuscript word count : 3016

Corresponding author

Prof. Emanuel Schwarz, PhD

Phone: +4962117032390

Email: emanuel.schwarz@zi-mannheim.de

Address: Hector Institute for Artificial Intelligence in Psychiatry (HITKIP), M7 68161 Mannheim, Germany

Keypoints

1. Question

Can polygenic scores derived from many psychiatry-relevant GWAS be used to stratify schizophrenia patients in a biologically and clinically meaningful way?

2. Findings

We identified two clusters of schizophrenia patients and characterized them as distinct genetic risk profiles based on their polygenic risk contributions. While both profiles showed comparable risk for schizophrenia, one profile was associated with higher liability for neuroticism and depression, reduced cognitive performance, and more complex clinical manifestations compared with patients in the second profile showing the opposite characteristics.

3. Meaning

Our results provide replicable insights into the genetic architecture of schizophrenia and have the potential to inform future personalized treatments.

Abstract

1. Importance

Schizophrenia is clinically and biologically heterogeneous, with marked variability in course and treatment response. Stratification of patients may advance individualized therapy and clarify underlying mechanisms.

2. Objective

To identify, validate, and replicate genetic risk factors distinct to each other in schizophrenia patients using polygenic risk scores of large numbers of psychiatry-relevant phenotypes.

3. Design

We analyzed genetic and phenotypic data from FinnGen (discovery dataset) and two independent cohorts for validation (PsyCourse, Bari). Using PRScope, a framework for standardizing polygenic score calculation and for patient stratification, we calculated 413 psychiatry-related polygenic scores for individuals with schizophrenia and controls. The resulting multi-PGS matrix was used to stratify patients through a data-driven approach. The findings were validated via cross-validation in FinnGen, and replicated in PsyCourse and Bari.

4. Setting

We used available large-scale datasets containing genetic and phenotypic information and publicly available GWAS summary statistics.

5. Participants

Data from FinnGen (7,486 schizophrenia cases; 27,288 controls), PsyCourse (419 cases; 299 controls), and Bari (531 cases; 738 controls) were included in this study.

6. Main outcome and measures

We examined the validity of the genetic differences among patients, predictive accuracy across datasets, contributing genetic domains, and phenotypic differences, including clinical severity proxies.

7. Results

Two data-driven clusters of patients with differing genetic risk profiles were identified. Both showed comparable genetic liability for schizophrenia, but diverged strongly in genetic risk for multiple psychiatry-related traits. The profile of the first cluster was characterized by a higher risk of depression, neuroticism, and low cognitive performance, whereas the second cluster showed a profile similar to that of healthy controls in these dimensions. Despite equal genetic liability for schizophrenia, the first profile showed higher predictability for schizophrenia in a case–control prediction model and was associated with significant differences in clinical severity indicators, such as higher clozapine use.

8. Conclusions and Relevance

A data-driven, polygenic risk–based approach revealed two biologically and clinically distinct genetic risk profiles. Genetic liability for depression traits, neuroticism, and low cognition beyond schizophrenia liability itself, appears to shape disease penetrance and severity of schizophrenia. These findings highlight the potential of multivariate polygenic risk stratification for refining schizophrenia nosology and tailoring interventions.

Introduction

Schizophrenia (SCZ) is a severe, usually chronic mental disorder with a lifetime prevalence estimate of up to 1% and an incidence peak in early adulthood[1]. Its severity is underscored by a life expectancy approximately 15-years shorter than the general population[2] and a mere 10-20% competitive employment rate[1,3,4].

SCZ is a phenotypically complex disease, and its individually variable manifestations span multiple domains like perception, affect, cognition, behavior, speech, motor skills, and autonomic regulation. Furthermore, the complexity and heterogeneity of SCZ extend to treatment response with approximately one-third of patients responding poorly[5,6]. Flattening of affect, avolition and cognitive decline (collectively negative symptoms), are especially persistent[7,8]. Formulations of sub-classification of SCZ have been abandoned in the latest editions of the commonly used diagnostic manuals (ICD-11 and DSM-V) due to a lack of reliability and clinical utility[9].

Evidence suggests that SCZ encompasses a multitude of overlapping neurobiological processes[10],[11]. Identifying more homogenous subgroups in terms of these processes would advance the etiological understanding of the disorder and personalized medicine approaches[12]. In family and twin studies, heritability for SCZ has been estimated at 60%-80% overall[13–15], among the highest for psychiatric disorders, and 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 basis for understanding the source of that heterogeneity through stratification[1,19,20]. Genome-wide association studies (GWAS) of SCZ indicate that it is a highly polygenic disorder with thousands of common variants (i.e., single-nucleotide polymorphisms (SNPs)) contributing to SCZ liability, each exerting a small effect on overall disorder risk[21,22]. It is estimated that around 25% of liability to SCZ is attributable to the total contribution of common genetic variation (i.e., SNP-heritability). 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 called polygenic risk scores (PRS)) for diseases and a wide variety of traits. These indices of polygenic predisposition show robust predictive performance in independent target samples, but they currently fall short of explaining disease liability to the degree suggested by SNP-heritability estimates[19,23].

Therefore, considered together with moderate-to-high genetic correlations with other psychiatric disorders (e.g., r_g=0.7 with bipolar disorder (BIP) and r_g =0.35 with depression)[15,22] and related traits (e.g., r_g=0.21 with Neuroticism)[23,24], an approach based on a broad spectrum of PGS seems promising. Furthermore, SCZ PGS are primarily designed to distinguish cases from controls, a limitation potentially remedied by a multi-PGS strategy in case-case discrimination. While the PGS of SCZ combined with additional psychiatric PGS with known associations with SCZ have been used for subtyping[25], and a data-driven multi-PGS approach has demonstrated superior performance in predicting SCZ case-control status over single-PGS models[26], the latter strategy has not been applied to characterize patients based on distinct genetic risk profiles (GRPs) before.

Building on this recent research, we aimed to: (i) identify distinct GRPs in SCZ by integrating large numbers of PGS of psychiatry-relevant phenotypes using robust, reproducible machine learning (ML) methods; (ii) validate these distinct GRPs in independent datasets; (iii) characterize the SCZ patients showing these GRPs in terms of clinical manifestations.

To achieve this, we developed *PRS*cope, a framework that generates PGS from a broad set of GWAS and then uses the resulting multi-PGS matrix to identify patient clusters through a data-driven approach and characterise them as GRPs based on their PGS compositions. Our analysis included data from a total of 8,436 patients with SCZ, and 28,325 healthy controls (HC) from three independent datasets, to identify and validate GRPs. In addition, we used data from 7,555 patients with BIP, and from 27,187 patients with Major Depressive Disorder (MDD), and from 37,539 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 disorders.

Methods

Overview

The methods used can be separated into the following steps: I) Automated GWAS search, II) Quality control (QC) of the GWAS, III) QC of the genotype data in the target datasets, IV) Calculation of the corresponding PGS in the target datasets V) Clustering of patients based on multi-PGS similarities in the training dataset, VI) Predicting the cluster assignments in the validation datasets, VII) Phenotypic profiling, VIII) Genetic profiling. Steps I) - IV) were done using PRScope, and steps V) - VIII) were done by the ML framework implemented for phenotype or disease stratification (for a schematic overview of the Methods, see Figure 1; for more details, see eMethods in the Supplement).

PRScope

PRScope extracts GWAS from the GWAS Catalog[27] based on a predefined selection of psychiatry-relevant categories specified by a list of Experimental Factor Ontology[28] (EFO) ids, which were provided as input to the pipeline. The harmonized GWAS fulfilling the filtering criteria and annotated with the provided EFOs were downloaded from GWAS Catalog and quality controlled, resulting in 348 GWAS. Furthermore, 65 GWAS, not listed in the GWAS Catalog but accessible from Psychiatric Genomics Consortium (PGC), Complex Trait Genetics (CTG), and other public sources, were identified and included in the analysis, using the same harmonization pipeline[29] as applied by the GWAS Catalog. The corresponding PGS were then calculated for each of the subjects in the target genotype datasets using the Bayesian-based PGS method LDPred-2[30]. The application of PRScope resulted in quality-controlled PGS for 413 psychiatry-relevant phenotypes that were computed in three cohorts ($Train_{FinnGen}$, $Val_{PsyCourse}$, Val_{Bari}) (Figure 1a,1b).

Machine learning framework

ML was applied to identify and validate subgroups of patients with SCZ characterized by distinct multi-PGS profiles. The analytical pipeline involved: (a) ranking of PGS with regard to their absolute correlation with SCZ; (b) clustering of SCZ patients based on multi-PGS profile into two clusters via k-means clustering using the most relevant PGS; (c) training a separate random forest[31] model for each cluster to distinguish SCZ cases from HC; and (d) predicting cluster membership across cross-validation folds on $Train_{FinnGen}$ and on validation datasets (models trained on $Train_{FinnGen}$, and validated on $Val_{PsyCourse}$ and Val_{Bari}), resulting in one prediction from Cluster 1 model and one from Cluster 2 model for each dataset (Figure 1c).

These predictions were then used to evaluate the predictive accuracy of each model. SCZ patients predicted as patients by exactly one of the models were assigned to the corresponding group (hereafter referred to as GRP) and the GRPs were characterized based on their polygenic risk contributions, as well as psychopathological measures (Val_{Bari}), comorbid disorders and medication information ($Train_{FinnGen}$), and selected phenotypes ($Val_{PsyCourse}$). The whole procedure was applied for each selected feature size separately, starting with the highest ranked PGS, adding PGS based on their ranking, resulting in a gradually increasing feature size (Figure 1d). For comparison, an alternative scheme where we group patients according to their predicted probabilities of having schizophrenia, as derived from prediction models trained on the same PGS as in our primary approach, was applied, hereafter referred to as *composite PGS*.

Genetic correlations among GWAS were used to categorize the 45 most SCZ-relevant PGS into meaningful PGS categories. Finally, this analysis was complemented by conjunctional False Discovery Rate analysis[32,33] to investigate whether genetic susceptibility for such phenotypes was mediated by shared genetics between PGS categories.

Cohorts

Data from FinnGen release 12 ($Train_{FinnGen}$) (7,486/27,288 cases/controls for SCZ, 7,555/11,198 for BIP, 27,187/26,993 for MDD. Controls for BIP and MDD are partly shared, but not with SCZ controls) were used for the identification of multi-PGS profiles, and to test associations. Data from the PsyCourse Study[34] ($Val_{PsyCourse}$) (419/299 cases/controls for SCZ) and Bari cohort[21] (Val_{Bari}) (531/738 cases/controls for SCZ) were used for GRP validation. All cohorts were independent of the original GWAS the PGS were derived from.

Results

The PRScope procedure was applied to data from the $Train_{FinnGen}$[35] to identify two data-driven clusters of SCZ. Separate ML models were then trained to characterize each of these clusters as GRPs and enable their prediction in independent test data. The predictive value of these cluster models, and their dependence on the number of included PGS, was first tested in $Train_{FinnGen}$ using a cross-validation approach (Figure 2a, for uncorrected p-values see eTable 1). Subsequently, these models were applied to the independent test data, illustrating that GRPs could be successfully validated at the selected feature sizes (Figure 2a). Similarly to $Train_{FinnGen}$ cross-validation, in both validation datasets ($Val_{PsyCourse}$ and Val_{Bari}), we observed that both GRP 1 and 2 were significantly positively associated with SCZ from feature size 5 onwards ($p <$ rejection threshold of the Benjamini-Hochberg (BH) procedure, see for untransformed p-values and predicted GRP proportions eTable 2-6), although for $Val_{PsyCourse}$ the p-value exceeded 0.05 for the GRP 2 model at feature size 24 (Figure 2a; shown as a dip due to the negative log transformation). The GRPs did not differ significantly in terms of age or sex in any of the three datasets.

With increasing feature size, the composition of the GRPs changed (Figure 2b), and we observed a substantial reallocation of patients, which started leveling off with approximately feature size 12 (Figure 3). Thus, in order to test clinical associations with the GRPs, we decided to predict the GRPs at feature sizes (1-6, 12, 24, 33, 45) that represent the points of a high rate of reallocation (see eMethods) into two validation datasets (Cohorts' demographics; eTable 7-9).

To assess the merits of our clustering approach, we additionally trained models simply predicting SCZ in the FinnGen data with the same PGS features we use in PRScope and used the median split of the predicted probabilities in CV test folds as groups that correspond to the GRP grouping. This alternative approach will be referred to as the *composite PGS* (see eMethods).

The contribution of neuroticism, depression and cognitive performance-related predisposition to SCZ susceptibility is mediated by shared genetic variation

The top 45 selected features were grouped using pairwise genetic correlation analysis of the corresponding GWAS, resulting in 4 categories of PGS which can be characterized as relating to SCZ, cognition, depression and neuroticism (see eFigure 1,2, and eMethods *Formation of the PGS categories*). Figure 2 illustrates that the two GRPs were (at higher feature sizes) defined by a divergence in these PGS categories while their overall SCZ risk was very similar. Specifically, GRP 1 showed markedly elevated neuroticism and depression related PGS and relatively low cognition related PGS, whereas GRP 2 displayed an opposite pattern for neuroticism and depression related PGS and intermediate levels (between GRP 1 and HC) for the cognition related PGS.

The composite PGS approach shows a somewhat similar pattern of genetic risk, with the predicted high risk group showing higher levels of SCZ, depression and neuroticism, and lower cognition PGS. However, the differences are markedly smaller except for SCZ PGS, for which it is much larger (eFigure 3).

Since SCZ has well-established genetic correlations with neuroticism, depression, and cognition[21,36], we investigated whether these aspects were collectively reflected as disproportional risk contributions of pleiotropic SNPs. For this, we constructed a subset PGS based on SNPs that were associated with SCZ and each of the PGS categories, using conjunctional FDR[32,33] (conjFDR). All SNP weights were taken from the SCZ summary statistic. We could show that all three of these subset PGS predicted SCZ better than PGS constructed of p-value matched (as indicated in the SCZ GWAS) SNPs outside their respective pleiotropy SNP group (Figure 4). eTable 10-12 show the pleiotropic variants as identified by conjFDR.

Genetic risk profiles have transdiagnostic relevance and differ in clinical measures

To clinically characterize the identified GRPs, differences in clinical features were assessed across the investigated datasets. All phenotypic results are shown for feature size 24 only for simplicity of

illustration (for other feature sizes see eFigure 4-19). In $Train_{FinnGen}$, GRP 1 was associated with a more severe and complex clinical manifestation in patients with SCZ as indicated by higher rates of antidepressant prescriptions and comorbid psychiatric diagnoses; including depression, neurotic (F40-F48 in ICD10), personality, and eating disorders (Figure 5). Importantly, the composite PGS comparison showed no significant differences in comorbidity rates (eFigure 20-30, and eTable 13 for group sizes).

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. 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 GRPs as represented by the PGS categories defining them. We found a significantly (FDR corrected p-values < 0.05) higher score for GRP 1 in the Inventory of depressive symptomatology (IDS-C_{30}) (Figure 5). In $Train_{FinnGen}$, GRP 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 GRP 2 model predictions followed a slightly different pattern, where the associations with both BIP and DEP were negative, with no consistent difference in magnitude across feature sizes (eFigure 15).

Discussion

Here, we aimed to capture distinct GRPs within the SCZ patients through the simultaneous investigation of patients' susceptibility to psychiatry-relevant phenotypes using 413 publicly available GWAS, leveraging a massive training dataset ($Train_{FinnGen}$). We asked whether finding two genetically distinctive clusters of SCZ patients might help to identify phenotypically distinct and clinically relevant subgroups and lead to a better understanding of the genetic risk architecture of SCZ. This resulted in two distinct GRPs with similar SCZ risk; one GRP was characterized by high PGS of phenotypes relating to neuroticism and depression and low PGS for cognitive traits (GRP 1), whereas the other GRP (GRP 2) had a profile similar to HC in those dimensions. While the GRPs exhibited approximately the same elevated levels of genetic liability for SCZ, a random forest ML model based on GRP 1 clearly outperformed a model based on GRP 2 in predicting SCZ in $Train_{FinnGen}$ and in independent validation datasets ($Val_{PsyCourse}$ and Val_{Bari}). Furthermore, patients with a GRP 1 tended to have clinical manifestations that indicate higher rates of disease complexity (psychiatric comorbidity), a clinical profile which does not show up when grouping the patients along their SCZ PGS alone. Interestingly, the two groups did not differ in regards to phenotypic measures which are commonly used to stratify SCZ patients (e.g., PANSS score / subscores), indicating their potential value beyond a phenotypical or clinical assessment of patients. While previous studies have demonstrated that disease prediction can be improved with a multi-PGS approach[26], this is the first time it has been applied to data-driven clustering of patients. The observation that the model based on GRP 1 is a considerably better predictor of SCZ than the one based on GRP 2, despite the similar SCZ risk, has several potential explanations that are not mutually exclusive. First, the genetic risk contributions for neuroticism, depression-related traits, and low cognitive performance may reflect biological mechanisms facilitating the actualization of disease risk. Second, the corresponding genetic risk may be only partially captured by SCZ PGS. Third, SCZ genetic risk estimation may also be more accurate in GRP 1 than in GRP 2 patients. In this scenario, the higher levels of polygenic risk of non-SCZ traits/disorders in GRP 1 would suggest a higher prevalence of alleles that confer risk to both SCZ and other related traits. 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. 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. Finally, phenotypic mediation, such as neuroticism being associated with higher stress levels, which in turn increase the risk of developing SCZ, could play a role[37,38].

The current study has several limitations. First, 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 GRPs. This will likely not capture the entire complexity of SCZ risk mediated by common variants. It should also be noted that the observed divergence in GRPs can arise from the correlation structure (a result of the univariate feature selection) of the input variables to the clustering algorithm, even without the existence of truly separate clusters. Second, our method only considered genetic predisposition to psychiatry-related traits. It is possible that the including of genetic risk in somatic domains would improve the classification and provide more insight into underlying biological mechanisms[39–41]. Third, 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 may not generalize to individuals with non-European ancestry[42,43].

In summary, our data-driven ML scheme successfully leveraged a large number of psychiatry-related PGS, for the first time, to derive two GRPs of SCZ patients. A prediction model based on GRP 1, 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. Especially due to the GRPs not being linked to common clinical measures (e.g., clinical questionnaires like the PANSS), they may also prove useful in personalizing treatments, a goal which a stratification based on clinical measures so far failed to accomplish despite multiple

attempts[44,45]. A prospect that could be evaluated in a randomized and controlled clinical trial. Taken together, this work establishes a framework for integrating multi-phenotype PGS with ML approaches for the genetic stratification of complex disorders. While demonstrated here in SCZ, the framework can be applied to other psychiatric and somatic conditions.

Acknowledgments

We thank all research participants and all researchers and clinicians who collected, generated, or processed the data used in this study.

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, awarded to L.A.A. and G.P.; Project CUP : H97G22000210007, and the Apulian regional government for the project: Early Identification of Psychosis Risk, awarded to A.B.

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. and L.A.A have 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 SR, Cannon TD. Schizophrenia. *N Engl J Med*. 2019;381(18):1753-1761. doi:10.1056/NEJMra1808803

2. Chang CK, Hayes RD, Perera G, 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*. 2011;6(5):e19590. doi:10.1371/journal.pone.0019590

3. Fleischhacker WW, Arango C, Arteel P, et al. Schizophrenia--time to commit to policy change. *Schizophr Bull*. 2014;40 Suppl 3(Suppl 3):S165-S194. doi:10.1093/schbul/sbu006

4. Kahn RS, Sommer IE, Murray RM, et al. Schizophrenia. *Nat Rev Dis Primers*. 2015;1:15067. doi:10.1038/nrdp.2015.67

5. Fusar-Poli P, Papanastasiou E, Stahl D, et al. Treatments of negative symptoms in schizophrenia: Meta-analysis of 168 randomized placebo-controlled trials. *Schizophr Bull*. 2015;41(4):892-899. doi:10.1093/schbul/sbu170

6. Potkin SG, Kane JM, Correll CU, et al. The neurobiology of treatment-resistant schizophrenia: paths to antipsychotic resistance and a roadmap for future research. *NPJ Schizophr*. 2020;6(1):1. doi:10.1038/s41537-019-0090-z

7. Lincoln TM, Dollfus S, Lyne J. Current developments and challenges in the assessment of negative symptoms. *Schizophr Res*. 2017;186:8-18. doi:10.1016/j.schres.2016.02.035

8. Dollfus S, Lyne J. Negative symptoms: History of the concept and their position in diagnosis of schizophrenia. *Schizophr Res*. 2017;186:3-7. doi:10.1016/j.schres.2016.06.024

9. Braff DL, Ryan J, Rissling AJ, Carpenter WT. Lack of use in the literature from the last 20 years supports dropping traditional schizophrenia subtypes from DSM-5 and ICD-11. *Schizophr Bull*. 2013;39(4):751-753. doi:10.1093/schbul/sbt068

10. Owen MJ. New approaches to psychiatric diagnostic classification. *Neuron*. 2014;84(3):564-571. doi:10.1016/j.neuron.2014.10.028

11. Clementz BA, Sweeney JA, Hamm JP, et al. Identification of distinct psychosis biotypes using brain-based biomarkers. *Am J Psychiatry*. 2016;173(4):373-384. doi:10.1176/appi.ajp.2015.14091200

12. Millan MJ, Andrieux A, Bartzokis G, et al. Altering the course of schizophrenia: progress and perspectives. *Nat Rev Drug Discov*. 2016;15(7):485-515. doi:10.1038/nrd.2016.28

13. Hilker R, Helenius D, Fagerlund B, et al. Heritability of schizophrenia and schizophrenia spectrum based on the nationwide Danish Twin Register. *Biol Psychiatry*. 2018;83(6):492-498. doi:10.1016/j.biopsych.2017.08.017

14. Cardno AG, Marshall EJ, Coid B, et al. Heritability estimates for psychotic disorders: the Maudsley twin psychosis series. *Arch Gen Psychiatry*. 1999;56(2):162-168. doi:10.1001/archpsyc.56.2.162

15. Lichtenstein P, Yip BH, Björk C, et al. Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet*. 2009;373(9659):234-239. doi:10.1016/S0140-6736(09)60072-6

16. Kendler K. What psychiatric genetics has taught us about the nature of psychiatric illness and what is left to learn. *Mol Psychiatry*. 2013;18:1058-1066. doi:10.1038/mp.2013.50

17. Alnæs D, Kaufmann T, van der Meer D, et al. Brain heterogeneity in schizophrenia and its association with polygenic risk. *JAMA Psychiatry*. 2019;76(7):739-748. doi:10.1001/jamapsychiatry.2019.0257

18. Pergola G, Penzel N, Sportelli L, Bertolino A. Lessons learned from parsing genetic risk for schizophrenia into biological pathways. *Biol Psychiatry*. 2023;94(2):121-130. doi:10.1016/j.biopsych.2022.10.009

19. Owen MJ, Legge SE, Rees E, Walters JTR, O'Donovan MC. Genomic findings in schizophrenia and their implications. *Mol Psychiatry*. 2023;28(9):3638-3647. doi:10.1038/s41380-023-02293-8

20. Jablensky A. Subtyping schizophrenia: implications for genetic research. *Mol Psychiatry*. 2006;11(9):815-836. doi:10.1038/sj.mp.4001857

21. Trubetskoy V, Pardiñas AF, Qi T, et al. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature*. 2022;604(7906):502-508. doi:10.1038/s41586-022-04434-5

22. Legge SE, Santoro ML, Periyasamy S, Okewole A, Arsalan A, Kowalec K. Genetic architecture of schizophrenia: a review of major advancements. *Psychol Med*. 2021;51(13):2168-2177. doi:10.1017/S0033291720005334

23. O'Donovan MC, Owen MJ. The implications of the shared genetics of psychiatric disorders. *Nat Med*. 2016;22(11):1214-1219. doi:10.1038/nm.4196

24. Gupta P, Galimberti M, Liu Y, et al. A genome-wide investigation into the underlying genetic architecture of personality traits and overlap with psychopathology. *Nat Hum Behav*. 2024;8(11):2235-2249. doi:10.1038/s41562-024-01951-3

25. Chen J, Mize T, Wu JS, et al. Polygenic risk scores for subtyping of schizophrenia. *Schizophr Res Treatment*. 2020;2020(1):1638403. doi:10.1155/2020/1638403

26. Albiñana C, Zhu Z, Schork AJ, et al. Multi-PGS enhances polygenic prediction by combining 937 polygenic scores. *Nat Commun*. 2023;14(1):4702. doi:10.1038/s41467-023-40330-w

27. Sollis E, Mosaku A, Abid A, et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res*. 2023;51(D1):D977-D985. doi:10.1093/nar/gkac1010

28. Malone J, Holloway E, Adamusiak T, et al. Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*. 2010;26(8):1112-1118. doi:10.1093/bioinformatics/btq099

29. Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published

genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. 2019;47(D1):D1005-D1012. doi:10.1093/nar/gky1120

30. Privé F, Arbel J, Vilhjálmsson BJ. LDpred2: better, faster, stronger. *Bioinformatics*. 2021;36(22-23):5424-5431. doi:10.1093/bioinformatics/btaa1029

31. Breiman L. Random Forests. *Mach Learn*. 2001;45(1):5-32. doi:10.1023/a:1010933404324

32. Andreassen OA, Thompson WK, Schork AJ, et al. Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. *PLoS Genet*. 2013;9(4):e1003455. doi:10.1371/journal.pgen.1003455

33. Smeland OB, Frei O, Shadrin A, et al. Discovery of shared genomic loci using the conditional false discovery rate approach. *Hum Genet*. 2020;139(1):85-94. doi:10.1007/s00439-019-02060-2

34. Budde M, Anderson-Schmidt H, Gade K, et al. A longitudinal approach to biological psychiatric research: The PsyCourse study. *Am J Med Genet B Neuropsychiatr Genet*. 2019;180(2):89-102. doi:10.1002/ajmg.b.32639

35. Kurki MI, Karjalainen J, Palta P, et al. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature*. 2023;613(7944):508-518. doi:10.1038/s41586-022-05473-8

36. Wootton O, Shadrin AA, Bjella T, et al. Genomic insights into the shared and distinct genetic architecture of cognitive function and schizophrenia. *Sci Rep*. 2024;14(1):15356. doi:10.1038/s41598-024-66085-y

37. Schneider TR. The role of neuroticism on psychological and physiological stress responses. *J Exp Soc Psychol*. 2004;40(6):795-804. doi:10.1016/j.jesp.2004.04.005

38. Almuqrin A, Georgiades A, Mouhitzadeh K, Rubinic P, Mechelli A, Tognin S. The association between psychosocial stress, interpersonal sensitivity, social withdrawal and psychosis relapse: a systematic review. *Schizophrenia (Heidelb)*. 2023;9(1):22. doi:10.1038/s41537-023-00349-w

39. Uellendahl-Werth F, Maj C, Borisov O, et al. Cross-tissue transcriptome-wide association studies identify susceptibility genes shared between schizophrenia and inflammatory bowel disease. *Commun Biol*. 2022;5(1):80. doi:10.1038/s42003-022-03031-6

40. Gong W, Guo P, Li Y, 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*. 2023;80(4):360-370. doi:10.1001/jamapsychiatry.2022.4974

41. Tylee DS, Lee YK, Wendt FR, et al. An atlas of genetic correlations and genetically informed associations linking psychiatric and immune-related phenotypes. *JAMA Psychiatry*. 2022;79(7):667-676. doi:10.1001/jamapsychiatry.2022.0914

42. Peterson RE, Kuchenbaecker K, Walters RK, et al. Genome-wide association studies in ancestrally diverse populations: Opportunities, methods, pitfalls, and recommendations. *Cell*. 2019;179(3):589-603. doi:10.1016/j.cell.2019.08.051

43. Fatumo S, Chikowore T, Choudhury A, Ayub M. A roadmap to increase diversity in genomic studies. *Nat Med*. Published online 2022. https://www.nature.com/articles/s41591-021-01672-4

44. Leucht S, Chaimani A, Krause M, et al. The response of subgroups of patients with schizophrenia to different antipsychotic drugs: a systematic review and meta-analysis. *Lancet Psychiatry*. 2022;9(11):884-893. doi:10.1016/S2215-0366(22)00304-2

45. McCutcheon RA, Pillinger T, Mizuno Y, et al. The efficacy and heterogeneity of antipsychotic response in schizophrenia: A meta-analysis. *Mol Psychiatry*. 2021;26(4):1310-1320. doi:10.1038/s41380-019-0502-5

Figure Legends

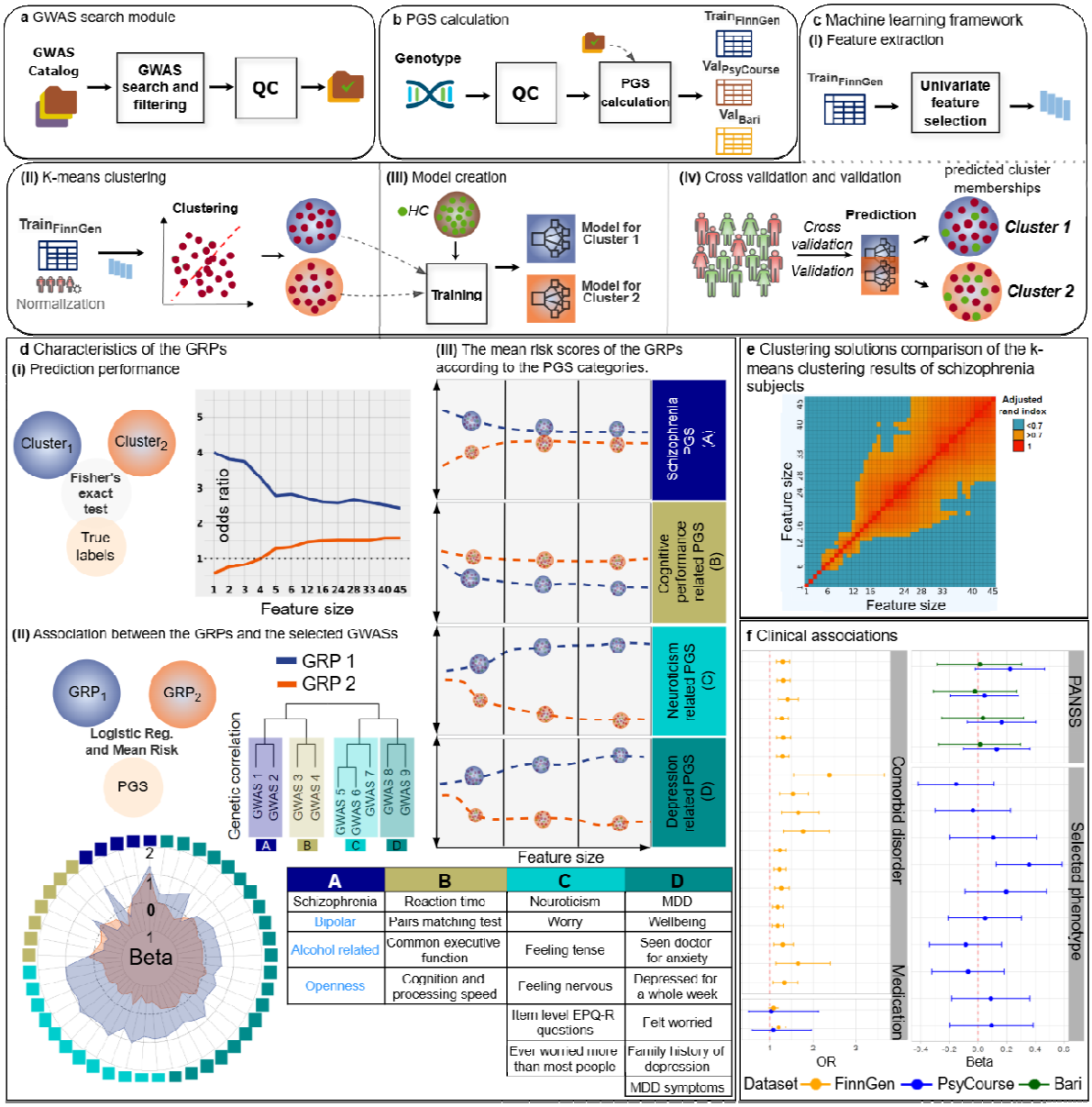


Figure 1. Overview of the PRScope framework and analysis pipeline. a) Psychiatry-related GWAS, annotated with relevant ontologies and fulfilling our filtering criteria, were downloaded from the GWAS Catalog and pre-processed for the subsequent PGS calculation. **b)** Genotype data from the three cohorts were pre-processed. PGS calculations for 413 pre-processed GWAS summary statistics were performed for each GWAS summary statistic using Bayesian-based PGS method

LDPred2. **c)** ML was applied to identify two data-driven clusters with differing PGS profiles in large-scale training dataset. These GRPs were then validated in two independent datasets and through cross-validation in the discovery dataset. Definition of *Cluster*: Predicted cluster memberships derived from cluster-specific models include both HC and SCZ subjects. Definition of *GRP*: SCZ patients uniquely correctly predicted by these models. (i) Ranking of PGS with regard to their absolute correlation with SCZ. The top 45 PGS that showed the strongest correlation with SCZ were included in ML analysis (from 1 to 45, stepwise). 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) Clustering of SCZ patients based on a multi-PGS profile into two clusters via k-means clustering method using the selected features. The PGS of SCZ patients for a given GWAS were adjusted by subtracting the median of the HC and dividing by the standard deviation of the HC. (iii) Random forest ML was applied to train models for each cluster separately. (iv) Cluster memberships were predicted by the model in independent data and cross-validation test folds. **d)** The characterization of the multi-PGS signature of each GRP. (i) Associations between the predicted cluster memberships and diagnoses were quantified as odds ratios and significance was determined using two-sided Fisher's exact test. (ii) The beta coefficients obtained from logistic regression represent the associations between each of the 45 most SCZ-relevant PGS and GRP relative to HC. The table shows the PGS categories identified based on genetic correlation among GWAS. SCZ category included only SCZ PGS. Bipolar disorder, personality trait and alcohol related traits were not included in GRP characterization. (iii) PCs were computed separately for each GRP using all PGS within a PGS category, and the mean values of the first PCs were plotted for each GRP and feature size. **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 GRPs, differences in clinical features were assessed across the investigated datasets.

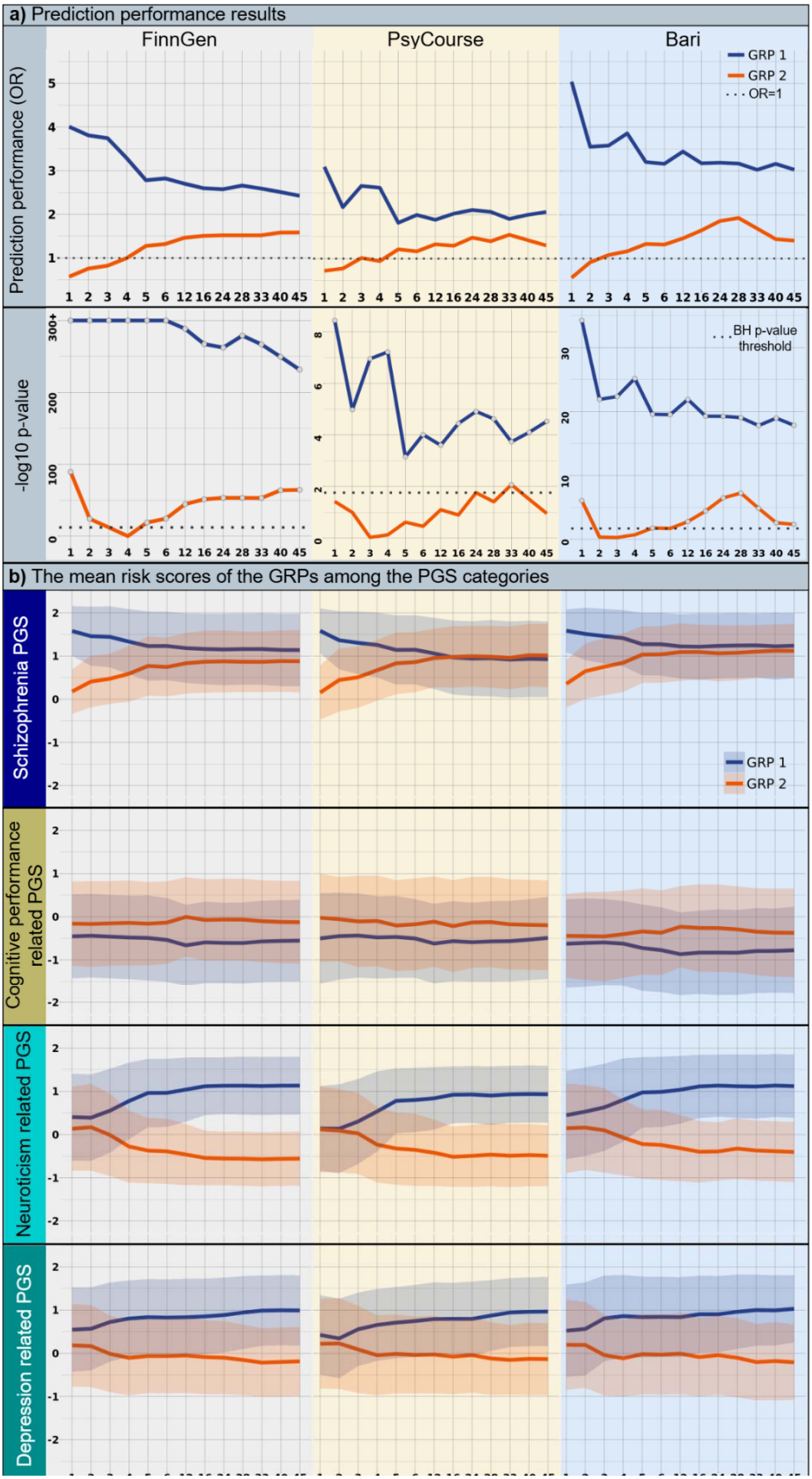
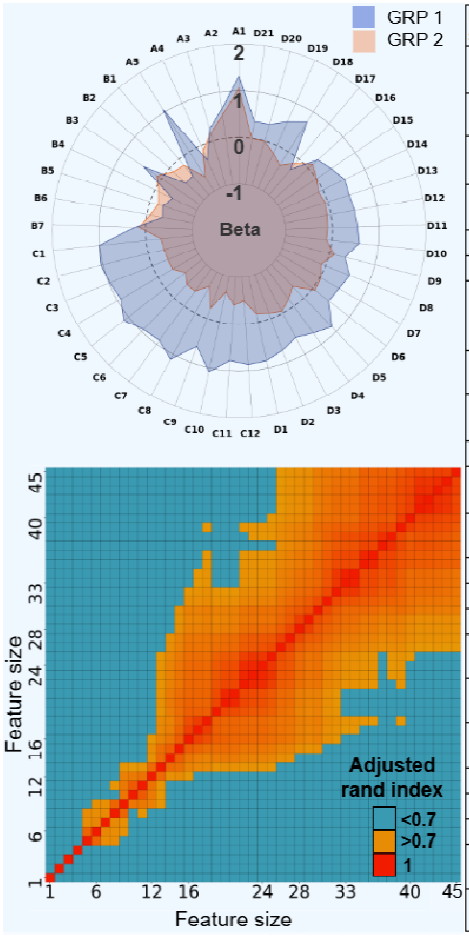

a) Prediction performance results
FinnGen
PsyCourse
Bari
GRP 1
GRP 2
OR=1
Prediction performance (OR)
-log10 p-value
BH p-value threshold
b) The mean risk scores of the GRPs among the PGS categories
Schizophrenia PGS
Cognitive performance related PGS
Neuroticism related PGS
Depression related PGS
GRP 1
GRP 2

Figure 2. Results of GRP validation. a) Associations between the predicted cluster memberships and diagnoses were quantified as odds ratios and significance were determined using two-sided Fisher's exact test where the input was a cross table of either GRP 1 or GRP 2 model and SCZ. The raw p-values were -log10 transformed to increase the readability of the plot. Feature sizes from 1 to 45 (x-axis) were plotted to demonstrate the formation with increasing feature size. 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 dashed lines on the p-value plots represent the rejection threshold of the BH procedure (number of tests = 26). The point symbols on the p-value plots indicate the significance of the odds ratios based on a BH p-value threshold. **b)** The PGS compositions of the GRPs were illustrated for the three PGS categories (neuroticism, depression, and cognition-related PGS categories) and for SCZ PGS. The mean of the first PC (to ensure comparability across units of analysis, all measures were scaled) of each PGS category was plotted across the feature sizes from 1 to 45 (x-axis) for GRP 1, GRP 2. The mean values were adjusted such that HC were centered at zero.



Feature size	GWAS	Phenotype	Feature size	GWAS	Phenotype
1	A1	Schizophrenia (PGC-2022, Bari excluded)	24	D9	Trauma
2	A2	Bipolar Disorder (PGC-2019)	25	A4	Reason for reducing amount of alcohol drunk: Other reason
3	D1	Seen psychiatrist for nerves anxiety tension or depression	26	D10	MDD symptoms anhedonia (community)(mdd2024symptoms)
4	C1	Would you call yourself tense or	27	D11	Ever unenthusiastic/disinterested for a whole week
5	C2	Tense / highly strung	28	D12	Ever depressed for a whole week
6	B1	Cognition and processing speed	29	D13	MDD (PGC-2018)
7	D2	Exploring the genetic overlap between 12 psychiatric disorders	30	C8	12-item EPQ-R neuroticism questionnaire items with yes endorsement
8	C3	Feeling tense	31	D14	Ever had prolonged feelings of sadness or depression
9	B2	Common executive function	32	C9	Ever worried more than most people would in similar situation
10	B3	Pairs matching task	33	D15	Trauma exposure
11	D3	MDD symptoms depressed (community)(mdd2024symptom	34	D16	Autism
12	C4	Would you call yourself a nervous person?	35	C10	Neuroticism
13	C5	Worry	36	D17	Social and economic stability
14	C6	Nervous feelings	37	D18	Depression (broad)
15	D4	Depression/anxiety	38	B6	Common executive function
16	D5	Seen doctor for nerves anxiety tension or depression	39	C11	Do you suffer from nerves?
17	D6	Illnesses of mother: Severe depression	40	A5	Openness (Big five 2024)
18	D7	Illnesses of siblings: Severe depression	41	D19	Ever suffered mental distress preventing usual activities
19	C7	Feeling nervous	42	B7	Reaction time
20	B4	Common executive function	43	D20	Illnesses of father: Severe depression
21	B5	Reaction time	44	D21	Depression
22	D8	Ever felt worried tense or anxious	45	C12	Worrier / anxious feelings
23	A3	Reason for reducing amount of alcohol drunk: Health			

Figure 3. PGS composition of the GRPs and stability across the feature sizes. The polar plot was generated using the FinnGen cohort's association results for the features size 45. To enable comparison across PGS categories, the PGS categories were plotted sequentially. The legend (in a form of a table) of the polar plot shows the ranked order of the 45 most SCZ-relevant PGS. The beta coefficients obtained from logistic regression represent the associations between each PGS and GRP relative to HC. Adjusted rand index (ARI) comparisons were generated using the FinnGen cohort's k-means clustering results and plotted as heatmap. ARI was calculated directly using the whole sample prediction results. To enable comparison across features, 0.7 is selected as the threshold for high similarity. Detailed information about the 45 GWAS listed in the table can be found in eTable 14. The values used to create the polar plot can be found in eTable 15. The ARI values used to generate the heatmap can be found in eTable 16.

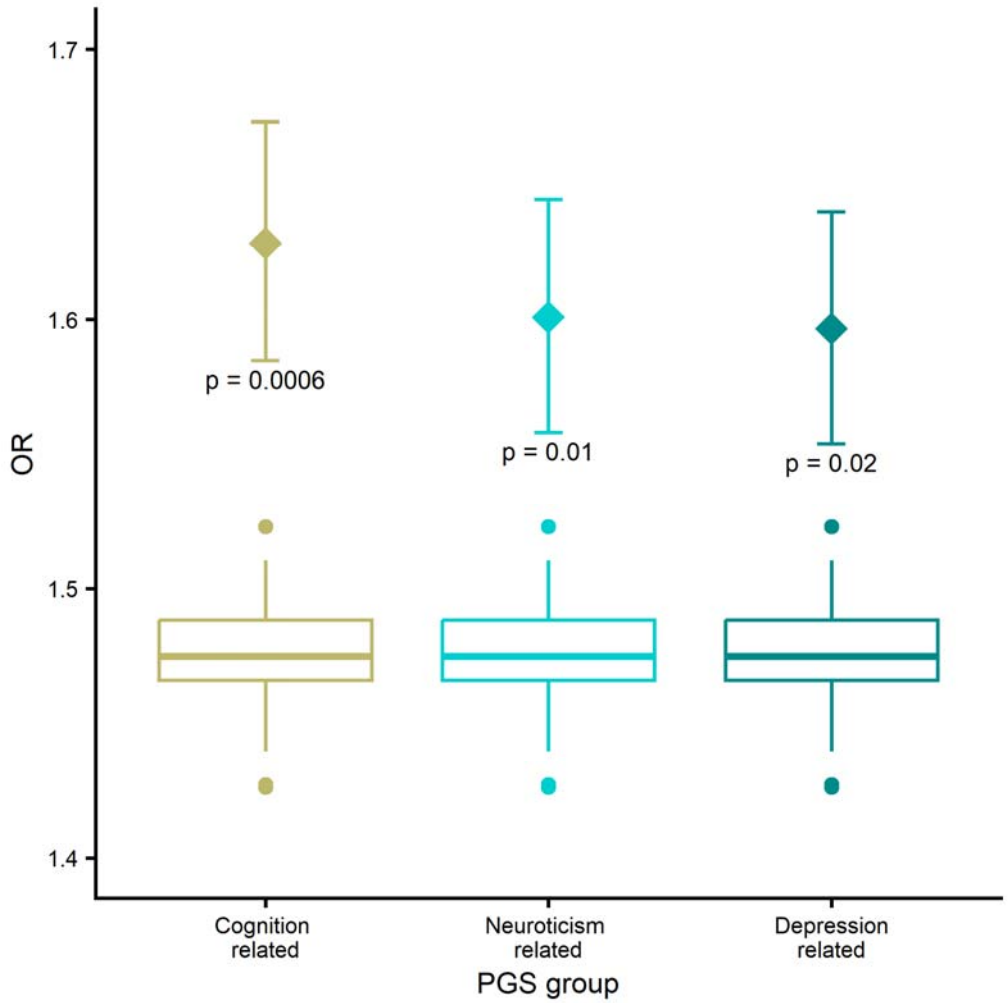


Figure 4. Predictive performance of the conjFDR-based PGS in Train$_{FinnGen}$. The diamonds indicate the odds ratios for a PGS constructed with conjunctional FDR with a given PGS category (see Methods) in a logistic regression where true SCZ status was the dependent variable. The boxplots represent estimates from similar models with different PGSs 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 PGSs are contrasted with the conjFDR PGS.

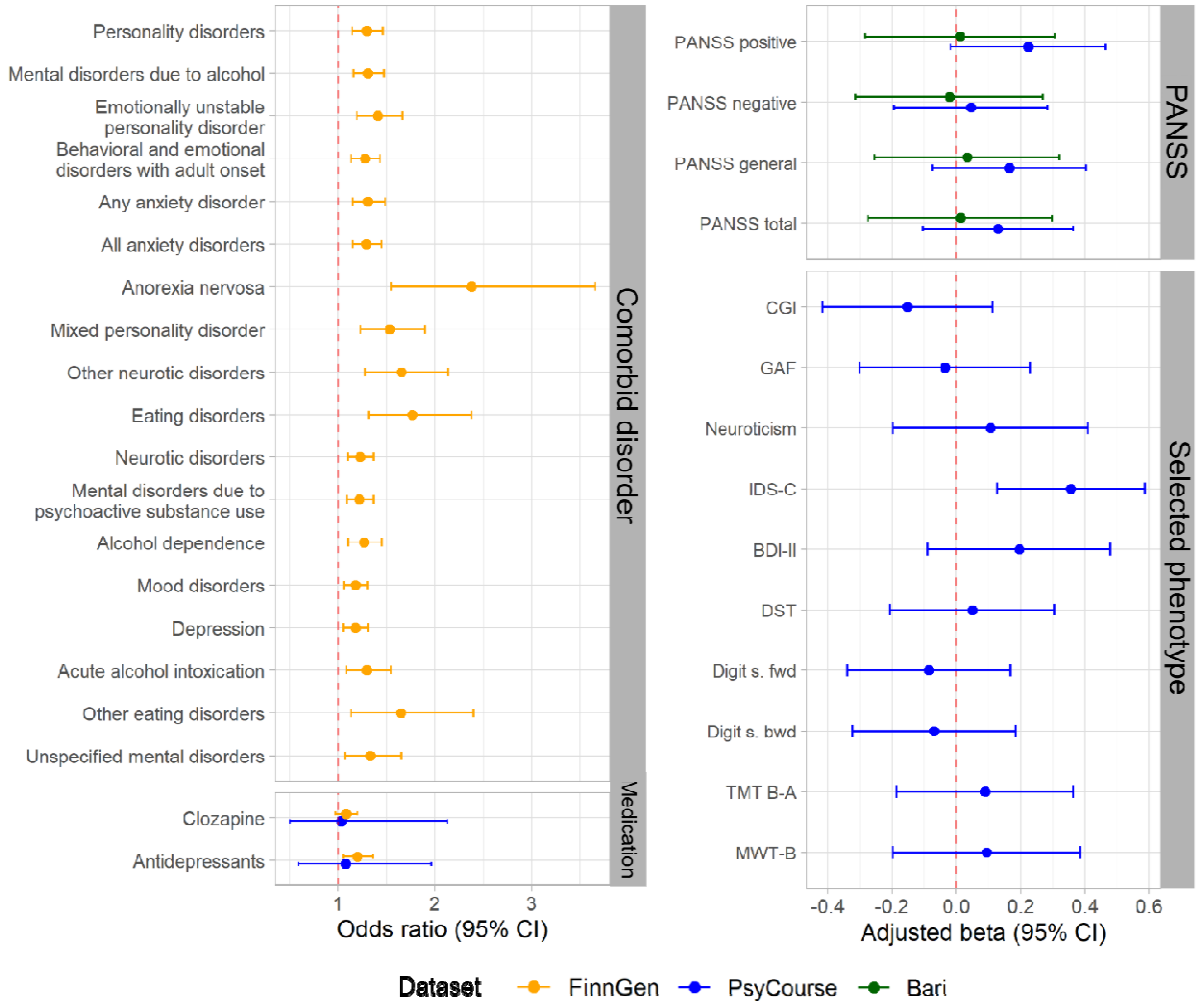


Figure 5. Phenotypic profiling of GRP 1 and GRP 2. Feature size = 24. Baseline: GRP 2. All comparisons are adjusted for the SCZ PGS, sex, and age. In $Val_{PsyCourse}$ and Val_{Bari} additionally adjusted for center/study; all results with the 95% CI (displayed) not overlapping the null effect are also significant with an FDR of 0.05 applied within each category (for comorbid disorders only the significant ones are shown); analyzed case numbers of GRP 1 and GRP 2 are listed in brackets ($n_{GRP\ 2}$ / $n_{GRP\ 1}$): Clozapine (105/88), Antidepressants (105/88), PANSS = Positive and Negative Syndrome Scale, positive, negative, general and total score ($Val_{PsyCourse}$ 108/108, Val_{Bari} 89/95). CGI = Clinical global impression (109/108), GAF = Global Assessment of Functioning (108/108), Neuroticism = Dimension from Personality (87/76), IDS-C_{30} = Inventory of depressive symptomatology (109/103), BDI-II = Beck Depression Inventory-II (98/93), DST = Digit-Symbol-Test

(105/103), Digit s. fwd = Verbal digit span forward (108/105), Digit s. bwd = Verbal digit span backward (108/105), TMT B-A = Trail making test part B - part A (105/197), MWT-B = Multiple-Choice Vocabulary Intelligence Test (74/74). For all FinnGen analyses, the corresponding group sizes are (2696/3077).