

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository:<https://orca.cardiff.ac.uk/id/eprint/165636/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Legge, Sophie, Pardinias, Antonio , Woolway, Grace, Rees, Elliott , Cardno, Alastair, Escott-Price, Valentina , Holmans, Peter , Kirov, George , Owen, Michael , O'Donovan, Michael and Walters, James 2024. An assessment of the genetic and phenotypic features of Schizophrenia in the UK biobank. JAMA Psychiatry

Publishers page:

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



An Assessment of the Genetic and Phenotypic Features of Schizophrenia in the UK Biobank

Sophie E Legge¹ PhD; Antonio F Pardiñas¹ PhD; Grace Woolway¹ PhD; Elliott Rees¹ PhD; Alastair G Cardno² PhD; Valentina Escott-Price¹; Peter Holmans¹ PhD; PhD; George Kirov¹ PhD; Michael J Owen¹ PhD; Michael C O'Donovan¹ PhD; James TR Walters¹ PhD

¹ Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, UK

² Leeds Institute of Health Sciences, Division of Psychological and Social Medicine, Faculty of Medicine and Health, University of Leeds, Leeds, United Kingdom

Corresponding authors

Professor James TR Walters

Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, CF24 4HQ, UK

Email: WaltersJT@cardiff.ac.uk

Dr Sophie E Legge

Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, CF24 4HQ, UK

Email: LeggeSE8@cardiff.ac.uk

Word count of manuscript: 3408

Key Points

Question: How do individuals with a diagnosis of schizophrenia recruited in a large volunteer-based research resource (UK Biobank) differ from those in the Psychiatric Genomics Consortium (PGC) or those recruited from clinical settings?

Findings: This cross-sectional study found that liability to schizophrenia in UK Biobank had a high genetic correlation with the PGC. Compared to 4 clinically-ascertained schizophrenia samples, UK Biobank schizophrenia cases had significantly lower schizophrenia genetic liability as indexed by polygenic risk, lower rates of CNVs, and fewer phenotypic features of poor outcome.

Meaning: Individuals with schizophrenia in UK Biobank have fewer features of illness severity. Studies such as UK Biobank can help to capture the full range of heterogeneity in schizophrenia research.

Abstract

Importance: Large-scale biobanks provide important opportunities for mental health research, but selection biases raise questions regarding the comparability of individuals to those in clinical research settings.

Objective: Compare i) genetic liability to psychiatric disorders in individuals with schizophrenia in UK Biobank with the Psychiatric Genomics Consortium (PGC), and ii) genetic liability and phenotypic features with participants recruited from clinical settings.

Design: Cross-sectional. Data collected between 1993 and 2021. Analysis conducted between July 2021 and June 2023.

Setting: UK Biobank is population-based. Other schizophrenia samples are cross-sectional cohorts recruited from clinical settings.

Participants: Participants included UK Biobank schizophrenia cases ($n=1438$) and controls ($n=499,971$), a UK sample of participants with a clinical diagnosis of treatment-resistant schizophrenia (CLOZUK, $n=14,000$), and three cross-sectional clinical research samples; CardiffCOGS ($n=767$), Cardiff F-Series ($n=648$), and Cardiff Affected Sib-Pairs ($n=381$).

Exposure: None

Main Outcomes and Measures: A genome-wide association study (GWAS) of UK Biobank schizophrenia case-control status was conducted and the results were compared with those from the PGC via genetic correlations. To test for differences with the clinical samples, polygenic risk scores (PRS) were calculated for schizophrenia,

bipolar disorder, depression, and intelligence using PRS-CS. PRS and phenotypic comparisons were conducted using pairwise logistic regressions. The proportions of individuals with copy number variation (CNV) associated with schizophrenia were compared using Firth's logistic regression.

Results: Liability to schizophrenia in UK Biobank was highly correlated with the latest GWAS from the PGC ($r_g=0.98$, standard error=0.18), and showed the expected patterns of correlations with other psychiatric disorders. Schizophrenia PRS explained 6.8% of the variance in liability for schizophrenia case status in UK Biobank. UK Biobank schizophrenia cases had significantly lower schizophrenia PRS than three of the clinically-ascertained samples and significantly lower rates of schizophrenia-associated CNVs than CLOZUK. The UK Biobank schizophrenia cases had higher educational attainment and employment rates than the clinically-ascertained schizophrenia samples, lower rates of smoking, and a later age of onset of psychosis.

Conclusions and relevance: Individuals with schizophrenia in UK Biobank, and likely other volunteer-based biobanks, represent those less severely affected. Their inclusion in wider studies should enhance the representation of the full spectrum of illness severity.

Introduction

Large population-based volunteer biobanks are increasingly being used to study human disease. Millions of participants across the world from newly available biobanks will be made available for research within the next 5 years. However, these samples are known to be subject to ascertainment biases¹ in particular ‘healthy volunteer bias’. For example, the 5.5% of the 9.2 million people invited to UK Biobank that participated are disproportionately female, socioeconomically advantaged, and of white ethnicity. They are also less likely to be obese, to smoke, report fewer health conditions and have lower mortality rates². While ascertainment bias clearly affects prevalence estimates, it has been argued that it does not affect exposure-disease associations or scientific inference^{3,4}. However, studies have shown that these biases can change effect sizes in genetic association studies⁵, impact downstream analyses¹, and new methods are being developed to detect biases and offset them⁶.

It is unclear how these selection biases coupled with differing methods of identifying and defining affected status, such as the use of self-report and electronic health records, influence the features of schizophrenia cohorts identified through large population-based samples, and how such cohorts will differ from clinically-ascertained samples. Non-random participation does not just affect population-based cohorts. Clinically-ascertained studies of serious mental illness, typically through secondary care, can be underrepresented for those who have difficulty obtaining such care due to socioeconomic and other causes of health care disparities^{7,8}. Moreover, they are also likely to be underrepresented for people with mild forms of the disorder who may not be referred to secondary care much less hospitalized, while those with excellent clinical

outcomes are likely to be discharged early from secondary care, biasing against secondary care or hospital-based recruitment.

UK Biobank offers the opportunity to learn lessons of general relevance for large-scale volunteer-based studies⁹. While the UK Biobank population as a whole has been well characterised, the genetic and phenotypic features of those with serious mental illness have not. Here, we investigate the extent to which schizophrenia as diagnosed in UK Biobank resembles schizophrenia in large genetic studies, as represented by those included in the Psychiatric Genomics Consortium (PGC) or as diagnosed in clinically-ascertained samples. We compared genetic correlations with the PGC, and polygenic risk scores, rates of copy number variation, and phenotypic features of individuals with schizophrenia in UK Biobank with four independent UK based samples. These findings are of general relevance to studies from other human biobanks, to mental health cohorts defined from electronic health records, and other alternative sources.

Methods

Participants

Participants were included from UK Biobank¹⁰ ($\sim n=500,000$) and 4 schizophrenia sample collections (Table 1, CLOZUK¹¹ [$\sim n=14,000$], CardiffCOGS¹² [$n=767$], Cardiff F-Series¹² [$n=648$], and Cardiff Affected Sib-Pairs¹² [$n=381$]). Genetic analyses included all samples and phenotypic analyses included all samples apart from CLOZUK.

UK Biobank is a biomedical database and research resource of approximately 500,000 individuals from across the UK aged between 40 and 69 years at recruitment (between 2006-2010)¹⁰. There are 4 sources from which a schizophrenia diagnosis can be detected

in UK Biobank; self-report (field IDs 20002 and 20544), ICD-10 F20 medical record diagnosis from hospital admissions (field IDs 41202 and 41204) or death records (field IDs 40001 and 40002), or an equivalent read code from primary care records (field ID 130875). eAppendix 1 further describes these sources. We defined schizophrenia in UK Biobank as a schizophrenia diagnosis reported from at least one of these sources. 1438 participants met one or more of these criteria at the time of analysis (eTable 1), which was based on data extracted in July 2021. Controls were defined as participants who had no indication of a psychotic disorder from the above sources (ICD-10 F21-29 codes inclusive). The North-West Multi-Centre Ethics Committee granted ethical approval to UK Biobank and all participants provided written informed consent. This study was conducted under UK Biobank project numbers 13310 and 14421.

CLOZUK is an anonymised sample of approximately 14,000 individuals taking clozapine in the UK with a diagnosis of treatment-resistant schizophrenia, previously described¹¹. CardiffCOGS ($n=767$), Cardiff F-Series ($n=648$), and Cardiff Affected Sib-Pairs ($n=381$) participants were recruited from community, inpatient and voluntary mental health services in the UK¹². The Cardiff Affected Sib-Pairs sample includes families with 2 or more siblings diagnosed with schizophrenia (or schizoaffective disorder, provided one of the siblings had schizophrenia). ICD-10 or DSM-IV F20 schizophrenia diagnoses in CardiffCOGS, Cardiff F-Series, and Cardiff Affected Sib-Pairs were based on Schedules for Clinical Assessment in Neuropsychiatry¹³ (SCAN) interviews and lifetime psychiatric clinical case notes. All schizophrenia sample collections received UK National Research Ethics Service approval and study participants provided written informed consent.

Comparison to PGC

We conducted a GWAS of schizophrenia in the UK Biobank and used the results to calculate genetic correlations with the PGC samples. The GWAS compared participants in UK Biobank with schizophrenia to participants without any mental or behavioural disorder (defined as ICD-10 F00-F99 in field category 1712) to circumvent artificial enrichments in the genetic correlations with other psychiatric conditions. UK Biobank participants were genotyped on either the UK Biobank Axiom or the UK BiLEVE Axiom purpose-built arrays. Standard quality control procedures were applied prior to imputation using the Haplotype Reference Consortium (HRC) panel, as previously described^{14,15}. SNPs were excluded using PLINK¹⁶ in line with thresholds used by the PGC¹⁷: minor allele frequency (MAF) < 0.01, Hardy-Weinberg equilibrium (HWE) p-value < 1×10^{-6} using the 'midp' and 'keep-fewhet' options for multi-population datasets, imputation INFO score < 0.9, SNP call rate < 0.95. Individuals with SNP missingness > 0.05 were excluded.

Association testing was based on the Scalable and Accurate Implementation of GEneralized mixed model (SAIGE)¹⁸ method. SAIGE is appropriate when case-control numbers are unbalanced and/or in the context of population structure. The null logistic model was conducted on a reduced dataset of relatively independent SNPs (n=90,684), created using PLINK's¹⁶ pruning procedure ($r^2 < .05$ and 500-kilobase window). Covariates included in the null logistic model were the first five principal components, plus any principal components from the first 20 that were associated with schizophrenia, genotyping array, self-reported ethnicity, sex and age at interview (schizophrenia cases were younger than unaffected controls). The leave-one-chromosome-out option was implemented to account for related individuals. Post GWAS processing was conducted using FUMA¹⁹ to annotate and visualise the results.

Genetic correlations were calculated using LDSC^{20,21} between the schizophrenia GWAS in UK Biobank and GWAS for schizophrenia¹⁷, bipolar disorder²², major depressive disorder (MDD)²³, attention deficit/hyperactivity disorder (ADHD)²⁴, autism spectrum disorder (ASD)²⁵, anorexia nervosa²⁶, cannabis use disorder²⁷, alcohol use disorder²⁸ and intelligence²⁹. Corresponding genetic correlations were also calculated for the PGC GWAS for schizophrenia and differences with UK Biobank schizophrenia results assessed via chi-square tests.

A schizophrenia PRS was calculated in UK Biobank using a method consistent with the PGC¹⁷ to allow comparison of the variance explained in schizophrenia case-control with the PGC and UK Biobank. The PRS was calculated via a clumping and thresholding approach in PRSicev2³⁰ for those of European genetic ancestry, as previously described³¹.

Comparison to clinically-ascertained cohorts

We compared PRSs, rates of CNVs, and phenotypic features between cohorts. Unless otherwise stated, statistical analyses were conducted in R.

Polygenic risk scores

CardiffCOGS, Cardiff F-Series, and Cardiff Affected Sib-Pairs were genotyped on the Illumina HumanOmniExpress (version 8 or 12). CLOZUK samples were genotyped on either the Illumina HumanOmniExpress or Combo¹¹. For Cardiff University samples, quality control and imputation using the HRC was conducted as part of the DRAGON-Data protocol³². The steps taken to combine the genetic data from UK Biobank and our clinically-ascertained cohorts to calculate PRS are described in eAppendix 2. A subset of SNPs from this combined dataset with low levels of linkage disequilibrium ($r^2 < 0.2$ at 500kb window) were used to identify unrelated individuals and to calculate principal

components. The randomly selected unrelated individuals were identified using KING-robust kinship estimator in PLINK. A kinship cut-off of 0.044 was used, equivalent to removing third-degree relatives. Principal components were calculated using PC-Air³³ from the GENESIS package. Plots comparing principal components by study showed no evidence of differences by study/genotyping array (eFigure 1).

PRS were calculated for schizophrenia¹⁷, bipolar disorder²², MDD³⁴, and intelligence²⁹ based on GWAS summary statistics that did not overlap with those in the present study. In collaboration with Cardiff University after permission from UK Biobank under project number 13310, the Schizophrenia Working Group of the PGC generated a custom GWAS that excluded the UK Biobank participants (based on checksums derived from the genomic data) and the Cardiff University samples. Intelligence summary statistics were derived as part of a related project^{14,29}. Bipolar disorder²² and MDD³⁴ summary statistics were obtained from the PGC. Summary statistics were cleaned using `summaRygwasqc`³⁵. Using all SNPs in the combined dataset, we used PRS-CS³⁶ and PLINK to calculate the PRS, using the EUR UK Biobank reference dataset, 10,000 burnin iterations, 25,000 MCMC iterations and a phi value of 1 for schizophrenia, and the default phi value for intelligence, bipolar disorder and MDD.

We scaled the PRS in all samples using principal components³⁷ to allow comparisons regardless of ancestry. This approach was effective as demonstrated by eFigure 2, which displays the adjusted and unadjusted PRS in biogeographical genetic ancestry groups³⁸ (groups defined in eAppendix 3 and eFigure 3). Pairwise comparisons for the PRS were made between schizophrenia cases in UK Biobank and other samples using logistic regression controlling for sex. We repeated analyses in individuals of European genetic ancestry as defined by biogeographical grouping to ensure results were consistent.

Schizophrenia-associated CNVs

Details of CNV calling have been described for CLOZUK^{39,40}, UK Biobank⁴¹, and CardiffCOGS⁴⁰. The Cardiff F-Series and Cardiff Affected Sib-Pairs samples were called as part of the DRAGON-Data protocol³². One member from each third-degree (or more closely) related pair within each dataset was removed at random. As the CNVs of interest are rare, we combined the participants from CardiffCOGS, Cardiff F-Series and Cardiff Sib-pairs. Analyses were restricted to individuals of European genetic ancestry, as defined in eAppendix 4, due to the low numbers of observations and because the majority of individuals in the clinically-ascertained schizophrenia samples were of European genetic ancestry. We compared the number of individuals in UK Biobank with schizophrenia that carried any of 12 schizophrenia-associated CNVs³⁹ (listed in eTable 2) to the other samples using pairwise Firth's logistic regressions covarying for sex.

Schizophrenia-related phenotypes

UK Biobank cases were compared to those in CardiffCOGS, Cardiff F-Series and Cardiff Affected Sib-Pairs and controls in UK Biobank for phenotypes known to be related to schizophrenia including demographics, education attainment, cognitive ability and known psychiatric and physical co-morbidities of schizophrenia. CLOZUK was not included due to the absence of relevant phenotypic data. It was not possible to include phenotypes from UK Biobank's mental health questionnaire due to the low completion rate in individuals with schizophrenia (14.5%), a return rate much lower than for the UK Biobank as a whole (31.5%). Comparisons were made only when equivalent definitions were available across samples, i.e., were assessed using similar wording on their respective questionnaires and/or where responses could be harmonised into comparable categories. eTable 3 details each phenotype and its definition in each

sample. Pairwise comparisons were calculated between schizophrenia cases in UK Biobank and the other samples using logistic regression controlling for sex and age at recruitment. Year of birth was also included for the education variables. Secondary analyses were conducted restricted to those of European genetic ancestry (as defined in Supplementary Methods 3).

Results

1438 individuals (0.3% of UK Biobank total sample; 38.2% female and 61.8% male; mean [SD] age, 54.7 [8.3] years at recruitment) in UK Biobank were identified with a schizophrenia diagnosis from at least one of the available sources; hospital records (n=1102, 76.7%), self-report (708, 49.2%), primary care records (n=75, 5.2%) and death records (n=23, 1.6%). 462 individuals had more than one source of diagnosis (all combinations listed in eTable 1). A maximum of 499,475 controls without a psychotic disorder were selected as a comparison group (99.4%; 54.4% female and 45.6% male; mean [SD] age, 56.5 [8.1] years at recruitment).

Comparison to PGC

After quality control, a GWAS including 1363 schizophrenia cases and 358774 controls from UK Biobank did not identify any genome-wide significant associated loci (threshold $p < 5 \times 10^{-8}$) as expected for a case sample of this size (eFigures 4 and 5; $\lambda_{GC}=1.03$).

Schizophrenia in the UK Biobank had a genetic correlation with the latest PGC schizophrenia GWAS¹⁷ that was close to 1 ($r_g=0.98$, standard error (se)=0.18). The genetic correlations between UK Biobank schizophrenia and bipolar disorder ($r_g=0.73$, se=0.14), MDD ($r_g=0.34$, se=0.08), intelligence ($r_g=-0.14$, se=0.06), or between any of the

other neuropsychiatric disorders were not significantly different from the genetic correlations between those traits and the latest PGC schizophrenia GWAS study (Figure 1, eTable 4).

Schizophrenia PRS calculated from the PGC GWAS was associated with schizophrenia case-control status within those of European genetic ancestry in UK Biobank (liability $r^2=6.8\%$; OR=2.04; 95% CI=1.92,2.17; $P=6.05 \times 10^{-110}$). A liability r^2 of 6.8% would be the 54th highest value out of 76 comparable samples in the latest PGC GWAS¹⁷.

Comparisons to clinically-ascertained cohorts

Polygenic risk scores

Schizophrenia cases in UK Biobank in comparison to the clinically-ascertained cohorts had on average a lower schizophrenia PRS, significantly so in comparison to CLOZUK, Cardiff F-Series and Cardiff Affected Sib-Pairs (Figure 2, eTable 5). The intelligence PRS for individuals with schizophrenia in UK Biobank was higher in comparison to CLOZUK and CardiffCOGS but not compared to Cardiff F-Series or Cardiff Affected Sib-Pairs (eTable 5, Figure 2). These results were consistent when restricting analyses to individuals of European genetic ancestry (eTable 6).

Schizophrenia case-control status in UK Biobank was also associated with the bipolar disorder, MDD and intelligence PRS (Figure 2, eTable 5).

Copy number variation

16 individuals (1.6%) with schizophrenia in UK Biobank had a schizophrenia-associated CNV compared to 0.8% of controls (OR=2.07; 95% CI=1.22,3.25; $P=8.98 \times 10^{-3}$). eTable 2 lists the number of carriers per CNV and cohort. The CNV rate for schizophrenia cases in UK Biobank was lower than for the schizophrenia cases in CLOZUK (1.6% vs. 2.7%,

OR=0.60; 95% CI= 0.35,0.95; P=0.03). A similar pattern was observed for the comparison between the UK Biobank and the combined sample of CardiffCOGS, Cardiff F-Series, and Cardiff Affected Sib-Pairs (1.6% vs. 2.4%, OR=0.63; 95% CI=0.33,1.17; P=0.14); although not significant, the sample size in that analysis means power to demonstrate a true difference is low.

Schizophrenia-related phenotypes

Phenotypic features of the samples are displayed in Figure 3 and eTable 7. Rates of comorbid affective diagnoses for the UK Biobank schizophrenia cases are described in eAppendix 5. Compared to the clinically-ascertained schizophrenia samples, cases in UK Biobank had patterns consistent with lower severity of illness (Table 2); they were less likely to be male, and males were more likely to have children (there was no difference in females). All cognitive indices including educational attainment and cognitive ability were higher in UK Biobank schizophrenia cases. Cases in UK Biobank had higher rates of current employment and an older self-reported age of onset of psychosis compared to all the clinical samples. Cases in UK Biobank had lower rates of smoking, but equivalent rates of comorbid physical illness once age was adjusted for.

Compared with controls, UK Biobank cases were more likely to be male, less likely to have been married, or to have had children, had lower educational outcomes indexed by a high school qualification (GCSEs), a degree, and had lower cognitive ability as measured by fluid intelligence (Figure 3, Table 2). They had higher rates of depression, tobacco use, epilepsy, heart disease and type 2 diabetes. Individuals with schizophrenia of working age had lower levels of current employment.

There was no evidence of an underrepresentation of schizophrenia cases from ethnic minorities compared to CLOZUK or controls (eTables 8-9). All phenotypic analyses were consistent in analyses restricted to those of European genetic ancestry.

Discussion

We compared individuals with schizophrenia from UK Biobank with those with schizophrenia in the PGC and with four clinically-ascertained schizophrenia research samples. Schizophrenia in UK Biobank had the genomic and phenotypic features expected from previous research but consistent with them being less severely affected.

Schizophrenia in UK Biobank had a genetic correlation of $r_g=0.98$ with the latest PGC schizophrenia GWAS¹⁷. Schizophrenia PRS explained 6.8% of the variance in liability for schizophrenia case-control status in those of European genetic ancestry in UK Biobank, which, while lower than the variance explained across the PGC samples as a whole (7.3% in all samples, 8.1% in those of European genetic ancestry)¹⁷, is within the range of other schizophrenia PGC samples. The association between schizophrenia PRS and schizophrenia case-control status in UK Biobank (OR=1.69 in all samples and OR=2.06 in those of European genetic ancestry) was also comparable to estimates from the PsycheMERGE consortium⁴² (OR=1.55) and US Veterans Affairs Health Care System⁴³ (OR=1.56). A recent study found the average schizophrenia PRS did not differ between individuals with schizophrenia identified via different diagnostic sources in UK Biobank⁴⁴. In addition, we observed phenotypic associations expected of schizophrenia such as an excess of males, lower cognitive outcomes, low rates of current employment, and rates of physical health comorbidities in UK Biobank schizophrenia cases comparable to epidemiological samples of schizophrenia⁴⁵.

After PGC schizophrenia, the next highest genetic correlation for UK Biobank schizophrenia was with bipolar disorder²² ($r_g=0.73$), the psychiatric disorder most genetically correlated with schizophrenia, and correlations with other psychiatric disorders were consistent with those from the PGC schizophrenia GWAS¹⁷ indicating that the genetics of the schizophrenia diagnosis in UK Biobank is compatible with others typically used in genomic studies. This is further supported by the strength of the schizophrenia PRS (OR=1.69) association with schizophrenia case-control status in contrast to bipolar disorder (OR=1.20) or MDD (OR=1.06).

Comparisons with clinically-ascertained schizophrenia cohorts indicated that those with schizophrenia in UK Biobank likely represent less severely affected cases. Compared to the other schizophrenia samples, UK Biobank cases had lower rates of males, higher cognitive ability and educational attainment, lower rates of smoking, older age of onset of psychosis, and higher current employment. Further, the rate of schizophrenia-associated CNVs and the schizophrenia PRS was lower in UK Biobank cases compared to Cardiff schizophrenia samples, although the latter is well within the range of values for individual studies included in the PGC¹⁷.

These findings reported here almost certainly reflect in part ascertainment differences. It is likely that focussing on clinically-ascertained samples in research may bias estimates towards more severe outcomes and that UK Biobank could offer an opportunity to study those with better outcomes. In addition, we found ethnic minorities to be equally represented in UK Biobank schizophrenia cases compared to CLOZUK or controls. While biobanks have advantages, they also have biases and tend to under sample individuals with serious mental illness and hence are an inefficient way to recruit large numbers of representative schizophrenia cases. Further, many phenotypes routinely collected in

clinical schizophrenia cohorts were not available in UK Biobank and the majority of people with schizophrenia did not complete online follow-up questionnaires such as the mental health questionnaire. Given future studies of the genetic basis of heterogeneity in schizophrenia will require both large numbers and high-quality assessments, targeted cohorts will still be needed, but these could be enhanced by the use of linked electronic medical records⁴⁶. There is an inevitable added cost for studies recruiting individuals with serious forms of mental illness, but this is essential if we are to base our research on representative samples and be able to generalise our findings.

Our findings have important implications for schizophrenia research conducted within, and outside of, UK Biobank. They indicate the need to integrate both cases recruited from secondary mental health services, which will be weighted towards more severe outcomes and those from biobank resources which will capture a higher proportion of less severely affected cases in order to encapsulate the full spectrum of schizophrenia.

Study limitations

In this paper we selected a pragmatic definition of schizophrenia that will be applicable to other biobank studies. More sophisticated definitions based on diagnostic algorithms or machine learning approaches are being developed and could offer further advantages to the field in the future. This study was conducted within the UK and many of the individuals were identified from linked medical records, and so results will need replication to ensure generalisability to other biobanks, countries and healthcare settings. The small sample size for CNV analyses meant that power to demonstrate significant differences was low.

Conclusions

Individuals with schizophrenia in UK Biobank have genomic and phenotypic features consistent with expectations for those with a diagnosis of schizophrenia but represent those less severely affected. The inclusion of such cases in wider schizophrenia studies has the potential to enhance representation of the full spectrum of illness severity.

Conflict of Interest Disclosures

Drs Rees, Walters, O'Donovan and Owen reported receiving grants from Akkrivia Health outside the submitted work. Drs Walters, O'Donovan and Owen reported receiving grants from Takeda Pharmaceuticals Ltd outside the submitted work. No other disclosures were submitted. Drs Walters and O'Donovan are the PGC Schizophrenia Working Group co-chairs. Dr Legge is funded by PGC as Schizophrenia Data Intake Representative.

Funding/Support

The work was supported by a grant from NIMH (Award R01MH124873), a Medical Research Council Centre grant (MR/L010305/1) and programme grant (MR/P005748/1). AFP and JTRW were supported by the EU-funded projects "REALMENT" (964874) and "PsychSTRATA" (101057454). MCOB was supported by the EU-funded project "REALMENT" (964874). ER was supported by a UKRI Future Leaders Fellowship Grant MR/T018712/1. The content is the responsibility of the authors and does not necessarily represent the official views of the funding bodies.

Additional contributions

We acknowledge the support of the Supercomputing Wales project, which is part-funded by the European Regional Development Fund (ERDF) via Welsh Government.

Data access, responsibility, and analysis

SL had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Data sharing statement

UK Biobank data can be obtained upon application from <https://www.ukbiobank.ac.uk/enable-your-research>. UK Biobank schizophrenia GWAS summary statistics can be downloaded from <https://walters.psychm.cf.ac.uk/>. The de-duplicated PGC summary stats for schizophrenia are available on the PGC website <https://figshare.com/articles/dataset/scz2022/19426775>.

References

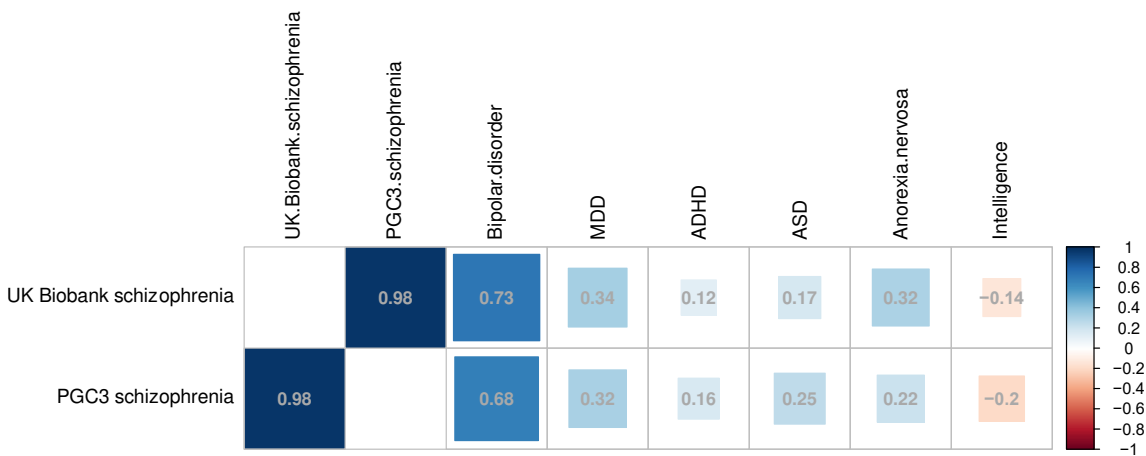
1. Pirastu N, Cordioli M, Nandakumar P, et al. Genetic analyses identify widespread sex-differential participation bias. *Nat Genet.* 2021;53(5):663-671.
2. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol.* 2017;186(9):1026-1034.
3. Rothman KJ, Gallacher JE, Hatch EE. Why representativeness should be avoided. *Int J Epidemiol.* 2013;42(4):1012-1014.
4. Pizzi C, De Stavola B, Merletti F, et al. Sample selection and validity of exposure-disease association estimates in cohort studies. *J Epidemiol Community Health.* 2011;65(5):407-411.
5. Schoeler T, Speed D, Porcu E, Pirastu N, Pingault JB, Kutalik Z. Participation bias in the UK Biobank distorts genetic associations and downstream analyses. *Nat Hum Behav.* 2023;7(7):1216-1227.
6. van Alten S, Domingue BW, Galama T, Marees AT. Reweighting the UK Biobank to reflect its underlying sampling population substantially reduces pervasive selection bias due to volunteering. *medRxiv.* 2022:2022.2005.2016.22275048.

7. Krantz MF, Hjorthøj C, Ellersgaard D, et al. Examining selection bias in a population-based cohort study of 522 children with familial high risk of schizophrenia or bipolar disorder, and controls: The Danish High Risk and Resilience Study VIA 7. *Soc Psychiatry Psychiatr Epidemiol.* 2023;58(1):113-140.
8. Taipale H, Schneider-Thoma J, Pinzón-Espinosa J, et al. Representation and Outcomes of Individuals With Schizophrenia Seen in Everyday Practice Who Are Ineligible for Randomized Clinical Trials. *JAMA Psychiatry.* 2022;79(3):210-218.
9. Cai N, Revez JA, Adams MJ, et al. Minimal phenotyping yields genome-wide association signals of low specificity for major depression. *Nat Genet.* 2020;52(4):437-447.
10. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* 2015;12(3):e1001779.
11. Pardiñas AF, Holmans P, Pocklington AJ, et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat Genet.* 2018;50(3):381-389.
12. Legge SE, Cardno AG, Allardyce J, et al. Associations Between Schizophrenia Polygenic Liability, Symptom Dimensions, and Cognitive Ability in Schizophrenia. *JAMA Psychiatry.* 2021;78(10):1143-1151.
13. World Health Organisation. *Schedules for Clinical Assessment in Neuropsychiatry: Version 2: Manual.* World Health Organization, Division of Mental Health; 1994.
14. Legge SE, Jones HJ, Kendall KM, et al. Association of Genetic Liability to Psychotic Experiences With Neuropsychotic Disorders and Traits. *JAMA Psychiatry.* 2019.
15. Leonenko G, Baker E, Stevenson-Hoare J, et al. Identifying individuals with high risk of Alzheimer's disease using polygenic risk scores. *Nat Commun.* 2021;12(1):4506.
16. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience.* 2015;4:7.
17. Trubetskoy V, Pardiñas AF, Qi T, et al. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature.* 2022;604(7906):502-508.
18. Zhou W, Nielsen JB, Fritsche LG, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet.* 2018;50(9):1335-1341.
19. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun.* 2017;8(1):1826.
20. Bulik-Sullivan BK, Loh PR, Finucane HK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet.* 2015;47(3):291-+.
21. Bulik-Sullivan B, Finucane HK, Anttila V, et al. An atlas of genetic correlations across human diseases and traits. *Nat Genet.* 2015;47(11):1236-1241.
22. Mullins N, Forstner AJ, O'Connell KS, et al. Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. *Nat Genet.* 2021;53(6):817-829.
23. Howard DM, Adams MJ, Clarke TK, et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat Neurosci.* 2019;22(3):343-352.
24. Demontis D, Walters RK, Martin J, et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat Genet.* 2019;51(1):63-75.
25. Grove J, Ripke S, Als TD, et al. Identification of common genetic risk variants for autism spectrum disorder. *Nat Genet.* 2019;51(3):431-444.

26. Watson HJ, Yilmaz Z, Thornton LM, et al. Genome-wide association study identifies eight risk loci and implicates metabo-psychiatric origins for anorexia nervosa. *Nat Genet.* 2019;51(8):1207-1214.
27. Johnson EC, Demontis D, Thorgeirsson TE, et al. A large-scale genome-wide association study meta-analysis of cannabis use disorder. *Lancet Psychiatry.* 2020;7(12):1032-1045.
28. Walters RK, Polimanti R, Johnson EC, et al. Transancestral GWAS of alcohol dependence reveals common genetic underpinnings with psychiatric disorders. *Nat Neurosci.* 2018;21(12):1656-1669.
29. Savage JE, Jansen PR, Stringer S, et al. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat Genet.* 2018;50(7):912-919.
30. Euesden J, Lewis CM, O'Reilly PF. PRSice: Polygenic Risk Score software. *Bioinformatics.* 2015;31(9):1466-1468.
31. Legge SE, Jones HJ, Kendall KM, et al. Association of Genetic Liability to Psychotic Experiences With Neuropsychotic Disorders and Traits. *JAMA Psychiatry.* 2019;76(12):1256-1265.
32. Lynham AJ, Knott S, Underwood JFG, et al. DRAGON-Data: a platform and protocol for integrating genomic and phenotypic data across large psychiatric cohorts. *BJPsych Open.* 2023;9(2):e32.
33. Conomos MP, Miller MB, Thornton TA. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet Epidemiol.* 2015;39(4):276-293.
34. Wray NR, Ripke S, Mattheisen M, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat Genet.* 2018;50(5):668-681.
35. Hubbard L. summaRygwasqc. 2020; <https://github.com/CardiffMRCPathfinder/summaRygwasqc>.
36. Ge T, Chen CY, Ni Y, Feng YA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat Commun.* 2019;10(1):1776.
37. Khan A, Turchin MC, Patki A, et al. Genome-wide polygenic score to predict chronic kidney disease across ancestries. *Nat Med.* 2022;28(7):1412-1420.
38. Huddart R, Fohner AE, Whirl-Carrillo M, et al. Standardized Biogeographic Grouping System for Annotating Populations in Pharmacogenetic Research. *Clin Pharmacol Ther.* 2019;105(5):1256-1262.
39. Rees E, Kendall K, Pardinas AF, et al. Analysis of Intellectual Disability Copy Number Variants for Association With Schizophrenia. *JAMA Psychiatry.* 2016;73(9):963-969.
40. Rees E, Walters JT, Georgieva L, et al. Analysis of copy number variations at 15 schizophrenia-associated loci. *Br J Psychiatry.* 2014;204(2):108-114.
41. Kendall KM, Bracher-Smith M, Fitzpatrick H, et al. Cognitive performance and functional outcomes of carriers of pathogenic copy number variants: analysis of the UK Biobank. *Br J Psychiatry.* 2019:1-8.
42. Zheutlin AB, Dennis J, Karlsson Linnér R, et al. Penetrance and Pleiotropy of Polygenic Risk Scores for Schizophrenia in 106,160 Patients Across Four Health Care Systems. *Am J Psychiatry.* 2019;176(10):846-855.
43. Bigdeli TB, Voloudakis G, Barr PB, et al. Penetrance and Pleiotropy of Polygenic Risk Scores for Schizophrenia, Bipolar Disorder, and Depression Among Adults in the US Veterans Affairs Health Care System. *JAMA Psychiatry.* 2022;79(11):1092-1101.
44. Grace EW, Sophie EL, Amy L, et al. Assessing the validity of a self-reported clinical diagnosis of schizophrenia. *medRxiv.* 2023:2023.2012.2006.23299622.

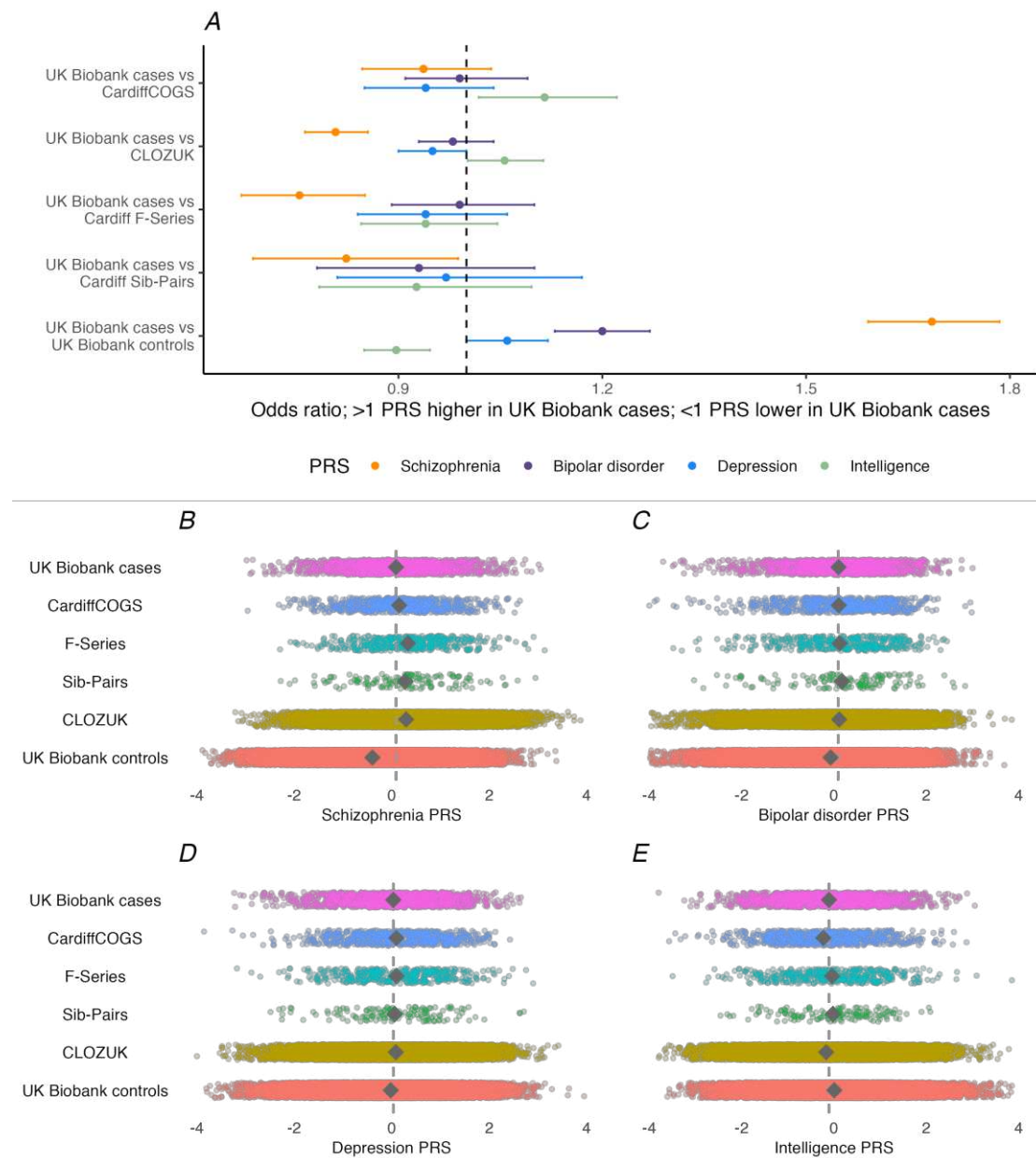
45. Kendall KM, John A, Lee SC, et al. Impact of schizophrenia genetic liability on the association between schizophrenia and physical illness: data-linkage study. *BJPsych Open*. 2020;6(6):e139.
46. Owen MJ, O'Donovan MC. Large-Scale Genomics: A Paradigm Shift in Psychiatry? *Biol Psychiatry*. 2021;89(1):5-7.

Figure 1: Genetic correlations



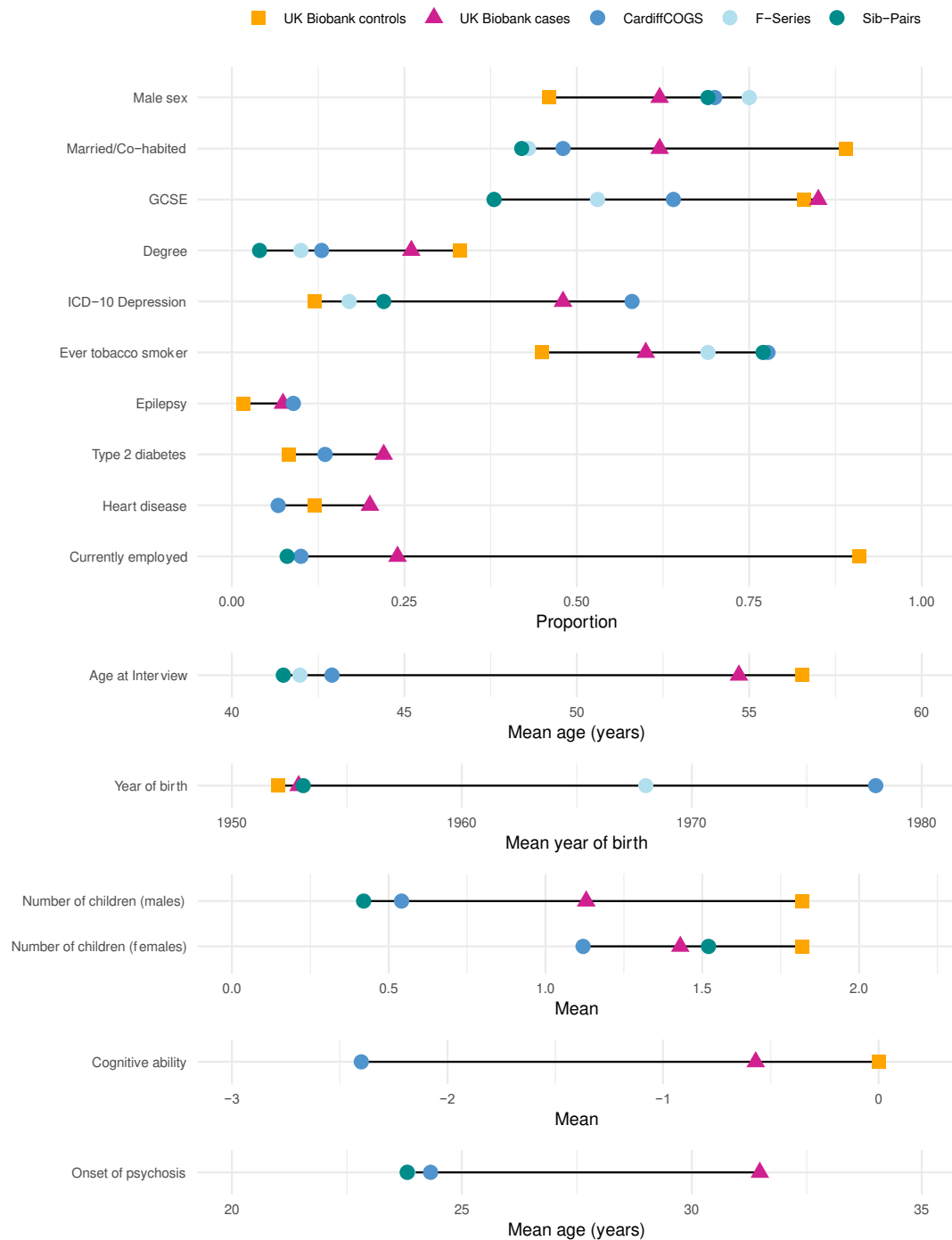
Genetic correlations between schizophrenia GWAS in UK Biobank and PGC3 schizophrenia¹⁷, bipolar disorder²², major depressive disorder²³, attention deficit/hyperactivity disorder (ADHD)²⁴, autism spectrum disorder (ASD)²⁵, anorexia nervosa²⁶ and intelligence²⁹. Comparison correlations with PGC3 schizophrenia are given in second row. Statistics and statistical comparison between the correlations are provided in eTable 5.

Figure 2: Polygenic risk comparisons between cohorts



Plot A displays the odds ratios (OR) from the polygenic risk score comparisons between UK Biobank schizophrenia cases and the other cohorts (eTable 5). The x-axis displays the OR for each comparison listed on y-axis. Colours correspond to polygenic risk score (schizophrenia in orange, bipolar disorder in purple, major depressive disorder in blue and intelligence in green). Points display the OR and bars the 95% CIs. The dotted line corresponds to 1 (null association). Plots B-E display the distribution of each polygenic risk score for each cohort. The dotted line represents the mean polygenic risk score in UK Biobank cases. Diamonds represent the mean polygenic risk score for each sample.

Figure 3: Phenotypic comparisons between cohorts



Cleveland plot of proportion (binary variables) or mean (continuous variables) values for each phenotype for each study as represented by different coloured dots and shapes.

Table 1: Cohort descriptions

	UK Biobank schizophrenia cases	UK Biobank unaffected controls	CLOZUK	CardiffCOGS	Cardiff F-Series	Cardiff Affected Sib-Pairs	PGC3 Schizophrenia
N	1438	499,475	14,666	767	648	381	67,390 cases and 94,015 controls
Female %	38.2%	54.4%	28.8%	29.6%	30.0%	31.2%	34.9%
Male %	61.8%	45.6%	71.2%	70.4%	70.0%	68.8%	65.1%
Age at recruitment (sd)	54.7 (8.3)	56.5 (8.1)	37.7* (11.9)	42.9 (12.3)	42.0 (12.0)	41.5 (12.6)	Unknown
Ancestry description	Multiancestry	Multiancestry	Multiancestry	European ancestry	European ancestry	European ancestry	Multiancestry
Treatment resistance	Unknown	N/A	100% TRS	56.2% TRS	27.5% TRS	40.7% TRS	Estimated minimum of 19%
Ascertainment	Volunteer-based biobank	Volunteer-based biobank	Anonymously ascertained from routine clozapine monitoring services	Clinically- ascertained	Clinically- ascertained	Clinically- ascertained affected sibling pairs	Primarily clinically- ascertained

Descriptions for each cohort. *CLOZUK age at recruitment is estimated from CLOZUK2 only and from the age at registration with Leyden Delta's monitoring system.

Table 2: Phenotypic comparisons between cohorts

Phenotype	UK Biobank cases vs. UK Biobank controls		UK Biobank cases vs. CardiffCOGS		UK Biobank cases vs. Cardiff F-Series		UK Biobank cases vs. Cardiff Affected Sib-Pairs	
	OR (95% CI)	P	OR (95% CI)	P	OR (95% CI)	P	OR (95% CI)	P
Male sex	1.95 (1.75,2.17)	1.30x10 ⁻³⁴	0.81 (0.66,1.01)	0.062	0.86 (0.69,1.09)	0.221	0.91 (0.66,1.26)	0.579
Age at interview	0.80 (0.76,0.84)	1.22x10 ⁻¹⁸	2.51 (2.30,2.75)	2.29x10 ⁻⁸⁹	2.38 (2.17,2.60)	2.21x10 ⁻⁸¹	3.31 (2.85,3.85)	2.53x10 ⁻⁵⁴
Year of birth	1.26 (1.20,1.33)	2.17x10 ⁻¹⁹	0.33 (0.30,0.36)	2.02x10 ⁻¹⁰⁴	0.80 (0.74,0.86)	4.44x10 ⁻⁹	1.01 (0.91,1.13)	0.783
Married/Co-habited	0.19 (0.16,0.23)	1.41x10 ⁻⁷⁸	0.89 (0.68,1.16)	0.390	1.19 (0.91,1.56)	0.212	1.24 (0.88,1.77)	1.24
Number of children (males)	0.64 (0.61,0.69)	4.39x10 ⁻⁴³	1.18 (1.04,1.33)	0.30x10 ⁻³	-	-	1.30 (1.08,1.56)	5.54x10 ⁻³
Number of children (females)	0.75 (0.70,0.81)	7.44x10 ⁻¹⁴	0.96 (0.83,1.11)	0.597	-	-	0.88 (0.75,1.05)	0.164
Currently employed	0.03 (0.03,0.03)	<1x10 ⁻²¹⁶	3.33 (2.38,4.64)	1.78x10 ⁻¹²	-	-	2.98 (1.71,5.17)	1.11x10 ⁻⁴
GCSEs	1.20 (1.03,1.38)	0.016	4.01 (3.09,5.21)	2.59x10 ⁻²⁵	5.30 (4.23,6.65)	1.76x10 ⁻⁴⁷	8.78 (6.56,11.76)	2.52x10 ⁻⁴⁸
Degree	0.66 (0.59,0.75)	2.61x10 ⁻¹¹	2.48 (1.85,3.32)	1.23x10 ⁻⁹	3.23 (2.39,4.36)	2.53x10 ⁻¹⁴	9.74 (4.95,19.16)	4.24x10 ⁻¹¹
Cognitive ability	0.54 (0.49,0.60)	6.17x10 ⁻³⁵	5.74 (4.59,7.18)	8.85x10 ⁻⁵³	-	-	-	-
Ever tobacco smoker	1.78 (1.60,1.99)	7.36x10 ⁻²⁶	0.48 (0.38,0.60)	3.49x10 ⁻¹⁰	0.75 (0.59,0.95)	0.015	0.48 (0.32,0.74)	7.21x10 ⁻⁴
ICD-10 Depression	7.57 (6.82,8.41)	<1x10 ⁻²¹⁶	0.68 (0.55,0.84)	3.27x10 ⁻⁴	4.63 (3.55,6.03)	1.01x10 ⁻²⁹	3.35 (2.37,4.74)	8.63x10 ⁻¹²
Epilepsy	4.77 (3.91,5.81)	1.08x10 ⁻⁵³	0.93 (0.63-1.37)	0.727	-	-	-	-
Type 2 diabetes	3.31 (2.92,3.77)	1.93x10 ⁻⁷⁵	1.30 (0.89-1.73)	0.067	-	-	-	-
Heart disease	1.93 (1.68,2.21)	3.71x10 ⁻²¹	2.49 (1.65-3.75)	1.23x10 ⁻⁵	-	-	-	-
Onset of psychosis	-	-	1.53 (1.34,1.74)	3.83x10 ⁻¹⁰	1.74 (1.50,2.01)	3.86x10 ⁻¹³	1.92 (1.55,2.37)	2.52x10 ⁻⁹

Results from pairwise regressions between UK Biobank schizophrenia cases and each of the other samples for each phenotype assessed. Odds ratio (OR) and 95% confidence intervals (CI) provided with corresponding p-value (P). Odds ratio refers to risk in UK Biobank cases; if greater than 1 this indicates higher rates (or higher values for continuous phenotypes) of said phenotype in UK Biobank cases compared to the other sample and if odds ratio is below 1 this indicates a lower rate (or lower values for continuous phenotypes) of said phenotype. Corresponding proportions and means are presented in eTable 7. Onset of psychosis (n=638) and cognitive ability (n=451) were only available for a subset of UK Biobank schizophrenia cases.