Impact of schizophrenia genetic liability on the association between schizophrenia and physical illness: a data linkage study

Kimberley M Kendall\textsuperscript{a}, Ann John\textsuperscript{b}, Sze Chim Lee\textsuperscript{b}, Elliott Rees\textsuperscript{a}, Antonio F Pardiñas\textsuperscript{a}, Marcos Del Pozo Banos\textsuperscript{b}, Michael J Owen\textsuperscript{a}, Michael C O'Donovan\textsuperscript{a}, George Kirov\textsuperscript{a}, Keith Lloyd\textsuperscript{b}, Ian Jones \textsuperscript{a}, Sophie E Legge\textsuperscript{a*}, James TR Walters\textsuperscript{a*}

\textsuperscript{*} authors contributed jointly to this work

\textsuperscript{a} MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, UK

\textsuperscript{b} Health Data Research UK, Swansea University Medical School, Swansea University, UK

Corresponding Author

Prof James Walters
MRC Centre for Neuropsychiatric Genetics and Genomics
Hadyn Ellis Building
Maindy Road
Cardiff
CF24 4HQ
02920 688434
waltersjt@cardiff.ac.uk
Abstract

Background

Individuals with schizophrenia are at higher risk of physical illnesses, which are a major contributor to their 20-year reduced life expectancy. It is currently unknown what causes the increased risk of physical illness in schizophrenia.

Aims

To link genetic data from a clinically ascertained sample of individuals with schizophrenia to anonymised NHS records. To assess (i) rates of physical illness in those with schizophrenia, and (ii) whether physical illness in schizophrenia is associated with genetic liability.

Method

We linked genetic data from a clinically ascertained sample of individuals with schizophrenia (CardiffCOGS, n=896) to anonymised NHS records held in the Secure Anonymised Information Linkage (SAIL) databank. Physical illnesses were defined from the General Practice Database and Patient Episode Database for Wales. Genetic liability for schizophrenia was indexed by (i) rare CNVs, and (ii) polygenic risk scores.

Results

Individuals with schizophrenia in SAIL had increased rates of epilepsy (standardised rate ratio (SRR)=5.34), intellectual disability (SRR=3.11), type 2 diabetes (SRR=2.45), congenital disorders (SRR=1.77), ischaemic heart disease (SRR=1.57) and smoking (SRR=1.44) in comparison to the general SAIL population. In those with schizophrenia, carrier status for schizophrenia-associated CNVs and neurodevelopmental disorder-associated CNVs was associated with height (P=0.015–0.017), with carriers being 7.5–7.7 cm shorter than non-carriers. We did not find evidence that the increased rates of
poor physical health outcomes in schizophrenia are associated with genetic liability for the disorder.

Conclusions

This study demonstrates the value of and potential for linking genetic data from clinically ascertained research studies to anonymised health records. The increased risk for physical illness in schizophrenia is not caused by genetic liability for the disorder.
Introduction

Individuals with schizophrenia have a 20-year reduced life expectancy.\(^{(1)}\) A major contributing factor is the increased rate of poor physical health outcomes in individuals with schizophrenia, related to conditions such as metabolic, cardiovascular and respiratory disease.\(^{(2)}\) Identifying the underlying reasons for these health disparities will provide a first step towards closing this health gap and thus has become a priority in schizophrenia research and clinical care in recent years.\(^{(3)}\) It is unclear whether the poorer physical health outcomes in schizophrenia arise from (i) the pleiotropic action of schizophrenia risk factors, or (ii) are secondary to illness-related factors such as negative symptoms, adverse effects of treatment (particularly antipsychotic medication) or poorer access to healthcare.

Genetic factors make major contributions to schizophrenia risk, arising from both common genetic variation and rare variants such as copy number variants (CNVs), the latter of these also being associated with physical health consequences in population samples.\(^{(4,5)}\) A recent study, in which genetic and electronic health record data were linked for 106,160 individuals, reported associations between polygenic risk scores for schizophrenia and several physical health phenotypes including smoking and reduced rates of obesity.\(^{(6)}\) At present, it is not known whether these genetic risk factors for schizophrenia are also associated with physical ill health in those with the disorder.

One reason for the dearth of research examining genetics and physical health outcomes in those with mental health disorders is likely to be the lack of available physical health data in existing psychiatric cohorts. In contrast, linkage of national registry data to anonymised mental and physical healthcare records, as exemplified in Nordic countries,
has provided a rich resource for mental health research and led to important insights including into physical health outcomes in schizophrenia.(7,8) The use of anonymised National Health Service (NHS) records and linkage of patient data to national registries in the UK is gaining momentum (9,10) and this approach has enabled researchers to investigate healthcare related outcomes for individuals with major mental health disorders such as schizophrenia.(2) The amalgamation and linkage of health record and genomic datasets offers further potential and initial reports describing linkage are emerging.(6,11–13) This type of approach has the potential to facilitate the collection of large-scale phenotypic data without the need to burden patients with extensive assessments. This may have particular advantage for physical health outcomes, which could be tracked longitudinally with updated health record linkage.

In this study, we aimed to link genetic data from a clinically ascertained sample of individuals with schizophrenia to anonymised NHS and administrative datasets in the Secure Anonymised Information Linkage (SAIL) databank in Wales (9), with a focus on the rich physical health outcome data held in primary care electronic health resources. We then aimed to examine the association between physical health outcomes and genetic liability for schizophrenia as indexed by (i) rare (frequency <1%) CNVs, and (ii) polygenic risk scores.

**Methods**

**Participants**

Study individuals (n = 958, aged 17-84 years, 41% female) from the Cardiff Cognition in Schizophrenia (CardiffCOGS) sample were recruited from community, inpatient and voluntary sector mental health services in the UK and underwent detailed phenotypic
assessment including a Schedules for Clinical Assessment in Neuropsychiatry (SCAN) interview. Interview data and clinical case-note vignettes were then used to arrive at a best estimate lifetime diagnosis according to DSM-IV criteria. Ethical approval was obtained from relevant NHS Multisite Research Ethics Committees and written informed consent was obtained for all study participants for genetic research and linkage of their genetic information to the SAIL databank. Further information on the CardiffCOGS sample has been published elsewhere.

**Electronic cohort from SAIL**

**Data source and data linkage**

The SAIL databank (http://www.saildatabank.com) is a national data repository containing anonymised, person-based, linkable data in Wales for over 3 million people. The procedure for linking research study data to SAIL has been described elsewhere. In brief, data from CardiffCOGS study individuals who had consented to linkage were imported into SAIL in line with permissions already granted to SAIL relating to good practice in research governance and privacy protection. We adopted a split file approach to separate individual identifiers from the interview data. Identity matching and creation of pseudonymised linkage keys (Anonymised Linking Fields, ALF) were performed by a trusted third party prior to linkage and further encryption of data sets using deterministic matching based on NHS number or probabilistic matching using available demographics based on the Welsh Demographic Service (WDS) dataset (all person’s registered with a general practice (GP) surgery). We included participants whose data were probabilistically linked with adequate level of matching accuracy (matching score ≥ 0.9).
We used the General Practice Database (GPD), containing diagnoses, symptoms, investigations, prescribed medication, referrals, coded hospital contacts and test results. At time of analysis, 77% (333/432) of GP surgeries in Wales supplied their data to SAIL. We also extracted data from the Patient Episode Database for Wales (PEDW), an NHS Wales hospital admissions dataset consisting of clinical information from all hospital admissions (inpatient and day cases) covering the entire population of Wales.

**Measures for health outcomes**

We used ICD-10 and Read Codes for GDP and PEDW datasets respectively to ascertain all studies health outcomes. For schizophrenia, we adopted the codes that were validated and used in previous studies.(16,18,19) We selected smoking, type 2 diabetes mellitus, ischaemic heart disease and body mass index (BMI) because of their high frequency in clinical samples of individuals with schizophrenia and their potential to contribute to increased mortality either directly, or via phenomena such as metabolic syndrome. We selected congenital disorders, intellectual disability and epilepsy as they are neurodevelopmental phenotypes with direct relevance to CNV carrier status. All of these variables also had either established SAIL algorithms or were considered to have high-quality data available in SAIL. We identified individuals with intellectual disability (20,21), ischaemic heart disease (17), epilepsy (22), diabetes mellitus (23) and determined smoking status (24) based on previously published works. The list of codes used for extracting height, BMI, and for identifying individuals with congenital disorders are given in Supplementary Table 1. For identifying schizophrenia, intellectual disability, ischaemic heart disease, epilepsy, diabetes mellitus and congenital disorders, we combined both the GPD and PEDW datasets. We additionally extracted individuals’ height, body mass index (BMI) and smoking status from the GPD where available.
For comparison, we extracted lifetime diagnoses and estimated crude rates, as well as age- and sex-standardised rates, of health outcomes for the whole population and those diagnosed with schizophrenia between the ages of 17 and 84 years (as at 30/06/2016). For estimating standardised rates from SAIL, crude rates were first computed for 5 age groups (17-34, 35-44, 45-54, 55-64, and 65-84 years) for both sexes. Standardised rates were then estimated based on the age and sex distribution in the clinical cohort. Linked data in SAIL were interrogated using structured query language (SQL DB2).

**Genotyping and CNV calling**

Samples were genotyped on OmniCombo or OmniExpress arrays (Illumina).(15) After standard quality control, imputation was performed using IMPUTE2 (25) and the 1000 genomes (phase 3) and UK10K reference panels.(26) Best-guess genotypes were generated with the following thresholds: minimal genotypic confidence > 90%, INFO-score > 0.9, MAF > 1%, and HWE p value < 1 x 10^{-10}.

Full details of the CNV calling methods and quality controls measures used have been published elsewhere.(27) Illumina Genome Studio (version 2011.1) was used to process raw intensity data into log R ratios (LRR) and B allele frequencies (BAF). PennCNV (version 1.0.3.18) was then used to call CNVs based on 666,868 probes common to all Illumina arrays used.(28) CNVs were joined if separated by less than 50% of their combined length. CNVs were excluded if they were (i) called with fewer than 10 probes, (ii) overlapped low copy repeats by more than 50% of their length, (iii) had a probe density of less than 1 probe/20kb, or (iv) had a frequency of >1%. (29)
CNVs

To examine for enrichment of rare, pathogenic CNVs in physical health comorbidity, we analysed the presence of 12 CNVs robustly associated with the risk of schizophrenia (15,27), and 54 CNVs nominally associated (p < 0.05) with intellectual disability, autism spectrum disorder or schizophrenia (‘neurodevelopmental CNVs’).(30) Following the approach taken in our previous work, 15q11.2 duplications were excluded due to their high frequency.(31) CNV burden analyses were carried out using PLINK on regions of variable copy number at two size thresholds (i) ≥500KB and (ii) ≥1MB and converted into carrier status for the purpose of regression analyses.

Polygenic risk scores

Polygenic risk scores were calculated using the largest published schizophrenia GWAS meta-analysis (39,915 cases, 64,639 controls) as a training set and using the established method described by Wray et al.(32,33) All study individuals were excluded from the training set. Scores were generated using the --score function in PLINK (29) for SNPs with MAF >10%, INFO score >0.9, a low linkage disequilibrium to each other and excluding indels and the extended MHC region. Polygenic risk scores were calculated at nine p-value thresholds; 1 x 10⁻⁸, 1 x 10⁻⁶, 1 x 10⁻⁴, 1 x 10⁻³, 0.01, 0.05, 0.1, 0.2 and 0.5.

Analysis

Rates of physical illness

Linked data in SAIL were interrogated using structured query language (SQL DB2). All crude rates and standardised rates of health outcomes were expressed as a percentage of population affected (lifetime prevalence). All standardised rate ratios (SRRs) and
their 95% confidence intervals (95% CIs) were calculated as previously described.(34–36)

Ascertainment rates of behaviours and diagnoses
We evaluated the agreement on diagnoses between the clinical and electronic cohorts for each health outcome by constructing 2-by-2 contingency tables based on the paired responses from the interview and SAIL. Level of agreement was then assessed by unweighted Cohen's Kappa coefficient (37) and Gwet's AC1. (38) 95% confidence intervals (CIs) of Cohen's Kappa and Gwet's AC1 were estimated as described in Fleiss, Cohen and Everitt (39) and Gwet (38) respectively. Strength of agreement metrics was categorised according to previously described criteria.(40)

CNV analyses
Association analyses were carried out using linear regression for average BMI (normalised using log10 transformation) and average height, and Firth logistic regression for all binary traits (type 2 diabetes mellitus, smoking, ischaemic heart disease, congenital disorders, epilepsy, intellectual disability). All analyses included age and sex as covariates.(41)

Polygenic risk analyses
We regressed a model for each polygenic risk score created from various training p-value thresholds against a base model including age, sex, the first five principal components and any additional principal components from the first 20 that were nominally associated with the phenotype of interest. We repeated these analyses with the addition of covariates reflecting symptom severity, nonresponse to antipsychotics,
antipsychotic exposure, smoking status and genotyping platform (defined in supplementary Table 2). All statistical analyses were carried out in R and results were subject to Bonferroni correction for the 8 phenotypes examined (p value threshold = 0.0063).

Results

A total of 896 (93.5%) study individuals from CardiffCOGS were linked to health records held in the SAIL databank. Linked study individuals had an age range of 17-84 years (mean 44 years), 371 (41%) were female and 724 (81%) had genetic data available.

Physical health outcomes in CardiffCOGS and SAIL

Table 1 and Figure 1 outlines the frequencies of the physical health outcomes in the CardiffCOGS sample and SAIL. Within the SAIL population, individuals with schizophrenia had increased rates of epilepsy (standardised rate ratio (SRR) = 5.34; 95% CI = 5.11, 5.57), intellectual disability (SRR = 3.11; 95% CI = 3.06, 3.11), type 2 diabetes (SRR = 2.45; 95% CI = 2.38, 2.53), congenital disorders (SRR = 1.77; 95% CI = 1.57, 1.99), ischaemic heart disease (SRR = 1.57; 95% CI = 1.51, 1.63) and smoking (SRR = 1.44; 95% CI = 1.42, 1.46). Individuals within CardiffCOGS had higher rates of type II diabetes (SRR = 1.29; 95% CI = 1.10, 1.52) compared to the population with schizophrenia in SAIL. Crude unadjusted population rates are given in Supplementary Table 3.

A comparison of ascertainment rates of behaviours and diagnoses in CardiffCOGS and SAIL
For physical health outcomes that were also reported at interview (smoking, type 2 diabetes mellitus, ischaemic heart disease, and epilepsy), we compared agreement with their health records (Table 2). Both Cohen's $\kappa$ and Gwet's AC1 for all physical conditions ranged from 0.502 to 0.936. These show that the agreement of the rates ascertained from the interview and health records were moderate to high. The highest agreements were observed for ischaemic heart disease and epilepsy. The strength of agreement was lowest for smoking behaviour (Cohen’s $\kappa = 0.380$ and Gwet’s AC1 = 0.621), reflecting only fair to substantial agreement.

**Copy number variation**

A total of 2.1% ($n = 15$) of the CardiffCOGS sample carried a schizophrenia-associated CNV, 4.9% ($n = 32$) carried a neurodevelopmental CNV, 9.5% carried a chromosomal duplication or deletion $\geq 500$KB (19 deletions, 52 duplications, 2 both) and 3.2% carried a duplication or deletion $\geq 1$MB (6 deletions, 17 duplications). We found no evidence that CNV carriers had increased rates of poor physical health outcomes (Table 3). However, average height was nominally associated with carrier status for CNVs associated with schizophrenia (beta = -0.075; 95% CI = -0.14, -0.01; $p = 0.017$) and neurodevelopmental disorders (beta = -0.077; 95% CI = -0.64, -0.07; $p = 0.015$). CNV carriers were on average 7.5-7.7cm shorter than non-carriers (Table 3). There was no evidence for association between rare CNVs of 500KB or greater and any of the phenotypes examined (Supplementary Table 4).

**Polygenic risk for schizophrenia**

We found no evidence for an association between polygenic risk scores for schizophrenia and the physical health outcomes studied (Figure 2, Supplementary Table
5), although there was weak evidence for an association with ischaemic heart disease at the genome-wide p-value threshold (odds ratio (OR) = 1.65; 95% CI = 1.22, 2.24; adjusted $R^2 = 0.035; p = 0.001$). The lack of association between schizophrenia PRS and physical health outcomes remained in sensitivity analyses covarying for symptom severity, non-response to antipsychotics, antipsychotic exposure, smoking status and genotyping array (Supplementary Table 6). However, we did identify significant associations between neuroleptic treatment resistance and type 2 diabetes (OR = 2.94; 95% CI = 1.79, 4.85; $p = < 0.0001$) and an association between symptom severity and intellectual disability (OR = 1.24; 95% CI = 1.05, 1.46; $p = 0.0012$). There were additional nominal associations ($p < 0.05$) between treatment resistance and epilepsy, and smoking and ischaemic heart disease (Supplementary Figure 6).

**Discussion**

In this study, we report the linkage of genetic data from a clinically ascertained sample of individuals with schizophrenia to anonymised NHS health records in the SAIL databank. Consistent with data collected at interview, we found that individuals with schizophrenia in Wales had increased rates of neurodevelopmental disorders (epilepsy, intellectual disability and congenital disorders), smoking, type 2 diabetes mellitus and ischaemic heart disease compared with the general population. However, the results of our genetic analyses suggest that these increased rates are not due to genetic liability to schizophrenia; we found no evidence for an association between genetic risk for schizophrenia indexed by CNVs or PRS and physical health outcomes.

These findings are supported by a recent population-based study that linked genetic and health record data for 106,160 individuals and found an inverse association
between increased schizophrenia PRS and obesity and an inverse association with diabetes when controlling for schizophrenia diagnosis or for antipsychotic medication history. (6) No association was found between schizophrenia PRS and the other physical health outcomes found in this study. (6) Thus, the evidence available indicates that that increased rates of poor physical health observed in patients with schizophrenia is unlikely to be driven by the genetic liability for the disorder.

There may be many other factors contributing to an increased rate of the physical health outcomes examined. For the non-neurodevelopmental outcomes, these include, but are not limited to, medication side effects and lifestyle choices such as smoking and diet. For example, weight gain caused by antipsychotic medication, poor diet and smoking may all contribute to the risk of type 2 diabetes mellitus and ischaemic heart disease. This is supported by our study; we found significant associations between treatment resistance and type 2 diabetes mellitus and nominal associations between treatment resistance and epilepsy, and smoking and ischaemic heart disease. The associations with treatment resistance may reflect the frequent use of clozapine in this patient group, an antipsychotic with propensity to cause weight gain and thus as a result may increase the risk of type 2 diabetes mellitus. Our additional analyses also found that intellectual disability was significantly associated with a greater illness severity. However, the reasons for this association are not yet clear but may reflect a greater burden of deleterious rare variants such as CNVs in these individuals. Importantly, several of these non-genetic factors are amenable to targeted intervention, which may reduce the risk of their consequent physical health outcomes.
Our finding that carriers of rare neurodevelopmental CNVs tend to be shorter than CNV noncarriers is in keeping with previous work on CNV associations with anthropometric traits in a sample of 191,161 adults. Macé et al report associations between both total CNV burden, several individual CNV loci (also examined in our study) and alterations in height. Additional findings from Macé et al of a lack of association between anthropomorphic traits and schizophrenia suggest that the CNVs affect height independent of disease.

The rates of smoking ascertained in our sample (87.0%) and in those with schizophrenia in SAIL (83.9%) appear high but are comparable with several studies also examining ‘ever’ smoking in individuals with schizophrenia. Similarly, the rates of ischaemic heart disease in the CardiffCOGS and SAIL schizophrenia population groups (both 6.9%) are comparable to rates established in other studies. CardiffCOGS had a higher rate of type 2 diabetes mellitus (17.4%) than the schizophrenia population in SAIL (13.5%) and other study samples such as a Swedish cohort study of individuals with schizophrenia (11 – 12.5%). This may reflect the higher than average proportion of individuals with chronic schizophrenia and long-term antipsychotic medication use in the CardiffCOGS sample.

**Study limitations**

The main limitation of this study was related to sample size; it is possible that we were underpowered to detect genetic associations with small effect sizes. However, it can still be concluded that genetic liability to schizophrenia does not have a large or significant impact of the occurrence of physical comorbidity. Nonetheless, our plan for future work is to link genetic data for a far greater number of individuals to their health records.
This study does provide an important exemplar for the value of linking genetic data to routinely collected health related data. Such an approach has great potential to generate a wealth of evidence, which can be translated into improved health outcomes for patients.
Author Details

Kimberley M Kendall, BSc, MBCh, MRCPsych, Wellcome Trust Clinical Research Fellow, MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, Cardiff University, UK

Ann John, BSc, MBBS, MRCGP, MPH, FFPH, PGCHE, Professor of Public Health and Psychiatry, Health Data Research UK, Swansea University Medical School, Swansea University, UK

Sze Chim Lee, BEng, BA, MSc, PhD, Senior Research Data Scientist, Health Data Research UK, Swansea University Medical School, Swansea University, UK

Elliott Rees, BSc, MSc, PhD, Research Fellow, MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, Cardiff University, UK

Antonio F Pardiñas, BSc, MSc, PhD, Lecturer, MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, Cardiff University, UK

Marcos Del Pozo Banos, BSc, MSc, PhD, Senior Research Data Scientist, Health Data Research UK, Swansea University Medical School, Swansea University, UK

Michael J. Owen, BSc, MB, ChB, PhD, FRCPsych, Sir Professor of Psychiatry, MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, Cardiff University, UK

Michael C. O'Donovan, BSc, MB, PhD, FRCPsych, Professor of Psychiatry, MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, Cardiff University, UK
George Kirov, MD, PhD, MRCPsych, Professor of Psychiatry, MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, Cardiff University, UK

Keith Lloyd, MBBS, MSc(econ), MSc, MD, FRCP(Edin) FRCPsych, FRCPsych, Professor of Psychiatry, Health Data Research UK, Swansea University Medical School, Swansea University, UK

Ian Jones, BSc, MBBS, MSc, PhD, MRCPsych, Director and Professor of Psychiatry, National Centre for Mental Health, MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, Cardiff University, UK

Sophie E. Legge, BSc, PhD, Research Associate, MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, Cardiff University, UK

James T.R. Walters, BM, MSc, PhD, MRCPsych, Professor of Psychiatry, MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, Cardiff University, UK

Declaration of Interest
The authors declare no conflicts of interest.

Funding
This project was funded by a Wellcome Trust Institutional Strategic Support Fund (ISSF) Populations Pilot Award. This project was supported also by the following grants: Medical Research Council Centre (MR/L010305/1), Program (G0800509) and Project
(MC_PC_17212) grants to Cardiff University, Medical Research Council Pathfinder (MR_PC_17211) to Swansea University, an MQ grant (MQDS16/36) to Cardiff University and the Health and Care Research Wales National Centre for Mental Health (155838) at Swansea University. KK was funded by a Wellcome Trust Clinical Research Training Fellowship.

**Acknowledgements**

We thank the participants, clinicians, field team and MRC Centre for Neuropsychiatric Genetics and Genomics laboratory staff for their help with the CardiffCOGS study. We thank the staff at deCODE Genetics for sample genotyping.
Figure 1. Rates of physical health phenotypes in CardiffCOGs schizophrenia sample, the schizophrenia population in SAIL, and the general population in SAIL.

Figure 2. A graph of the results from regression models for the association between polygenic risk for schizophrenia and physical health outcomes. Odds ratios are shown for the p value thresholds at which markers were selected.
<table>
<thead>
<tr>
<th>Phenotype</th>
<th>CardiffCOGS rate % (n)</th>
<th>SAIL schizophrenia population rate % (n)</th>
<th>SAIL population rate % (n)</th>
<th>SRR\textsubscript{sam, sch} (95% CI)</th>
<th>SRR\textsubscript{sch, gen} (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Congenital disorder</td>
<td>1.34 (12)</td>
<td>0.93 (326)</td>
<td>0.53 (21,745)</td>
<td>1.44 (0.81-2.57)</td>
<td>1.77 (1.57-1.99)</td>
</tr>
<tr>
<td>Intellectual disability</td>
<td>1.56 (14)</td>
<td>1.80 (623)</td>
<td>0.58 (22,142)</td>
<td>0.87 (0.51-1.48)</td>
<td>3.11 (3.06-3.11)</td>
</tr>
<tr>
<td>Ischaemic heart disease</td>
<td>6.93 (62)</td>
<td>6.99 (3,449)</td>
<td>4.46 (207,197)</td>
<td>0.99 (0.77-1.28)</td>
<td>1.57 (1.51-1.63)</td>
</tr>
<tr>
<td>Epilepsy</td>
<td>4.69 (42)</td>
<td>7.49 (2,628)</td>
<td>1.40 (52,020)</td>
<td>0.63 (0.46-0.85)</td>
<td>5.34 (5.11-5.57)</td>
</tr>
<tr>
<td>T2DM</td>
<td>17.43 (156)</td>
<td>13.50 (5,293)</td>
<td>5.51 (216,187)</td>
<td>1.29 (1.10-1.52)</td>
<td>2.45 (2.38-2.53)</td>
</tr>
<tr>
<td>Smoking (current/ex)</td>
<td>86.99 (689)</td>
<td>83.87 (22,120)</td>
<td>58.38 (1,649,589)</td>
<td>1.04 (0.97-1.11)</td>
<td>1.44 (1.42-1.46)</td>
</tr>
</tbody>
</table>

Table 1. Rates of physical health phenotypes in CardiffCOGS sample compared to the population rates in SAIL. sam – sample (CardiffCOGS), sch - schizophrenia, gen - general population. SRR\textsubscript{sam, sch} represents the standardised rate ratio of the clinical cohort to the schizophrenia population ascertained in SAIL. SRR\textsubscript{sch, gen} represents the standardised rate ratio of the schizophrenia population ascertained in SAIL to the general population ascertained in SAIL. SAIL rates were standardised using the age and sex distribution from the clinical cohort as reference. Crude unadjusted population rates are given in Supplementary Table 2. Numbers stated are out of 895 for the CardiffCOGS sample (except for smoking which was out of 792). Numbers stated for the population rate are out of 3,852,471 (except for smoking which was out of 2,958,064) and for schizophrenia population are out of 35,944 (26,588 for smoking).
<table>
<thead>
<tr>
<th></th>
<th>CardiffCOGS Interview</th>
<th>% Reported Affected at CardiffCOGS Interview</th>
<th>% Reported Affected in SAIL</th>
<th>Cohen's $\kappa$ (95% CI)</th>
<th>Gwet’s AC1 (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>T2DM</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Affected</td>
<td>Yes</td>
<td>75</td>
<td>13.34</td>
<td>0.675 (0.597 - 0.754)</td>
<td>0.884</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>43</td>
<td>17.69</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unaffected</td>
<td></td>
<td>14</td>
<td>535</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Ischaemic Heart Disease</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Affected</td>
<td>Yes</td>
<td>21</td>
<td>5.96</td>
<td>0.502 (0.362 - 0.641)</td>
<td>0.936</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>19</td>
<td>6.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unaffected</td>
<td></td>
<td>18</td>
<td>596</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Epilepsy</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Affected</td>
<td>Yes</td>
<td>24</td>
<td>7.81</td>
<td>0.529 (0.397 - 0.661)</td>
<td>0.935</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>10</td>
<td>5.11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unaffected</td>
<td></td>
<td>28</td>
<td>604</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Smoking</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Affected</td>
<td>Yes</td>
<td>515</td>
<td>71.92</td>
<td>0.380 (0.307 - 0.452)</td>
<td>0.621</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>119</td>
<td>77.41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unaffected</td>
<td></td>
<td>74</td>
<td>111</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Inter-rater agreement of the rates of ascertainment between the CardiffCOGS interview and linkage to NHS records. The smoking variable compares results for current/ex smoking status in SAIL and ever a regular smoker at interview. Cohen’s $\kappa$ refers to Cohen’s Kappa statistic (37), Gwet’s AC1 refers to the first order agreement coefficient (38). Both values range between 0 (chance, or no agreement) and 1 (perfect agreement). Values below 0.4 indicate poor agreement, values of 0.4-0.59 indicate moderate agreement and values above 0.6 indicate substantial agreement.
<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Schizophrenia CNVs</th>
<th>Neurodevelopmental CNVs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Carriers</td>
<td>Non-carriers</td>
</tr>
<tr>
<td>Average height (GP)</td>
<td>13</td>
<td>611</td>
</tr>
<tr>
<td>Average BMI log10 (GP)</td>
<td>13</td>
<td>604</td>
</tr>
<tr>
<td>Type 2 diabetes mellitus (Both)</td>
<td>Aff</td>
<td>&lt;5</td>
</tr>
<tr>
<td>Smoking – current or ex-smoker (GP)</td>
<td>Aff</td>
<td>9</td>
</tr>
<tr>
<td>Ischaemic heart disease (both)</td>
<td>Unaff</td>
<td>&lt;5</td>
</tr>
<tr>
<td>Congenital disorders (Both)</td>
<td>Aff</td>
<td>&lt;5</td>
</tr>
<tr>
<td>Unaff</td>
<td>14</td>
<td>701</td>
</tr>
<tr>
<td>Epilepsy (Both)</td>
<td>Aff</td>
<td>&lt;5</td>
</tr>
<tr>
<td>Unaff</td>
<td>13</td>
<td>673</td>
</tr>
</tbody>
</table>

Association analysis results for physical health phenotypes and two groups of CNVs (12 schizophrenia CNVs and 53 neurodevelopmental CNVs). GP/hospital/both in parentheses in the phenotype column refers to the source of diagnostic information used. Effect size refers to odds ratio in all cases except average height and average BMI for which the effect size is standardised Beta. Aff – affected; Unaff – unaffected; 95% CI – 95% confidence interval; p – uncorrected p value. Counts below 5 are masked to preserve participant anonymity.
References


