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


**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 National Health Service (NHS) records. To assess (a) rates of physical illness in those with schizophrenia, and (b) whether physical illness in schizophrenia is associated with genetic liability.

**Method**

We linked genetic data from a clinically ascertained sample of individuals with schizophrenia (Cardiff Cognition in Schizophrenia participants, n = 896) to anonymised NHS records held in the Secure Anonymised Information Linkage (SAIL) database. Physical illnesses were defined from the General Practice Database and Patient Episode Database for Wales. Genetic liability for schizophrenia was indexed by (a) rare copy number variants (CNVs), and (b) 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 = 2.43), congenital disorders (SRR = 1.77), ischaemic heart disease (SRR = 1.57) and smoking (SRR = 1.44) in comparison with 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 were 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.

**Keywords**

Schizophrenia; psychotic disorders; genetics; physical health.

**Copyright and usage**

© The Author(s) 2020. Published by Cambridge University Press on behalf of the Royal College of Psychiatrists. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

---

* Joint last authors.
Study aims
In this study, we aimed to link genetic data from a clinically ascertained sample of individuals with schizophrenia to anonymised NHS and administrative data-sets in the Secure Anonymised Information Linkage (SAIL) databank in Wales,7 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 (a) rare (frequency <1%) CNVs, and (b) polygenic risk scores.

Participants
Study individuals (n = 958, aged 17–84 years, 41% female) from the Cardiff Cognition in Schizophrenia (CardiffCOGS) sample were recruited from community, in-patient and voluntary sector mental health services in the UK and underwent detailed phenotypic assessment including a Schedules for Clinical Assessment in Neuropsychiatry (SCAN) interview.14 Interview data and clinical case-note vignettes were then used to arrive at a best estimate lifetime diagnosis according to DSM-IV criteria.15 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.16

Electronic cohort from SAIL
Data source and data linkage
The SAIL databank (http://www.saildatabase.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.16,17,18 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) 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 data-set (all individuals registered with a general practice (GP) surgery). We included participants whose data were probabilistically linked with an adequate level of matching accuracy (matching score ≥0.9).18

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 data-set consisting of clinical information from all hospital admissions (in-patient and day cases) covering the entire population of Wales.

Measures for health outcomes
We used ICD-10 and Read codes for GDP and PEDW data-sets, respectively, to ascertain health outcomes.19 For schizophrenia, we adopted the codes that were validated and used in previous studies.18,20,21 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, ischaemic heart disease, epilepsy, diabetes mellitus and congenital disorders, we combined both the GPD and PEDW data-sets. We additionally extracted individuals’ height, BMI and smoking status from the GPD where available.

For comparison, we extracted lifetime diagnoses and estimated crude rates, as well as age- and gender-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 June 2016). For estimating standardised rates from SAIL, crude rates were first computed for five age groups (17–34, 35–44, 45–54, 55–64 and 65–84 years) for both genders. Standardised rates were then estimated based on the age and gender distribution in the clinical cohort. Linked data in SAIL were interrogated using structured query language (SQL DB2).

Genetic data
Genotyping and CNV calling
Samples were genotyped on OmniCombo or OmniExpress arrays (Illumina).16 After standard quality control, imputation was performed using IMPUTE227 and the 1000 genomes (phase 3) and UK10 K reference panels.28 Best-guess genotypes were generated with the following thresholds: minimal genotypic confidence >90%, INFO-score >0.9, minor allele frequency (MAF) >1%, and Hardy-Weinberg equilibrium P-value <1 × 10−10.

Full details of the CNV calling methods and quality controls measures used have been published elsewhere.29 Illumina Genome Studio (version 2011.1) was used to process raw intensity data into log R ratios (LRR) and B allele frequencies. PennCNV (version 1.0.3.18) was then used to call CNVs based on 666 868 probes common to all Illumina arrays used.30 CNVs were joined if separated by less than 50% of their combined length. CNVs were excluded if they were (a) called with fewer than ten probes, (b) overlapped low copy repeats by more than 50% of their length, (c) had a probe density of less than 1 probe/20 kb, or (d) had a frequency of >1%.31

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,16,29 and 54 CNVs nominally associated (P < 0.05) with intellectual disability, autism spectrum disorder or schizophrenia (‘neurodevelopmental CNVs’).32 Following the approach taken in our previous work, 15q11.2 duplications were excluded because of their high frequency.33 CNV burden analyses were carried out using PLINK on regions of variable copy number at two size thresholds (a) ≥500 KB and (b) ≥1 MB and converted into carrier status for the purpose of regression analyses.
Polygenic risk scores

Polygenic risk scores were calculated using the largest published schizophrenia genome-wide association study meta-analysis (39 915 individuals with schizophrenia, 64 639 controls) as a training set and using the established method described by Wray et al.34,35 All study individuals were excluded from the training set. scores were generated using the –score function in PLINK33 for Single nucleotide polymorphisms with MAF >10%, INFO score >0.9, a low linkage disequilibrium to each other and excluding indels and the extended major histocompatibility complex region. Polygenic risk scores were calculated at nine P-value thresholds; 1 × 10^{-9}, 1 × 10^{-8}, 1 × 10^{-7}, 1 × 10^{-6}, 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 were calculated as previously described.36–38

Ascertainment rates of behaviours and diagnoses

We evaluated the agreement on diagnoses between the clinical and electronic cohorts for each health outcome by constructing two × two contingency tables based on the paired responses from the interview and SAIL. Level of agreement was then assessed by unweighted Cohen’s kappa coefficient39 and Gwet’s AC1.40 The 95% confidence intervals of Cohen’s kappa and Gwet’s AC1 were estimated as described in Fleiss, Cohen & Everitt41 and Gwet,40 respectively. Strength of agreement metrics was categorised according to previously described criteria.42

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, gender, 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, non-response 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 eight 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 Fig. 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 (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 2 diabetes (SRR = 1.29, 95% CI 1.10–1.52) compared with the population with schizophrenia in SAIL. Crude unadjusted population rates are given in supplementary Table 3.

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 k 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.43 The highest agreements were observed for ischaemic heart disease and epilepsy. The strength of agreement was lowest for smoking behaviour (Cohen’s k = 0.380 and Gwet’s AC1 = 0.621), reflecting only fair to substantial agreement.

CNV

A total of 2.1% (n = 15) of the CardiffCOGS sample carried a schizophrenia-associated CNV, 4.9% (n = 32) carried a

---

Table 1: Rates of physical health phenotypes in Cardiff Cognition in Schizophrenia (CardiffCOGS) sample compared with the population rates in Secure Anonymised Information Linkage (SAIL).

<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_CardiffCOGS_schizophrenia</th>
<th>SRR_CardiffCOGS_gen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Congenital disorder</td>
<td>1.34 (12)</td>
<td>0.93 (326)</td>
<td>0.53 (27145)</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>0.38 (22142)</td>
<td>0.87 (0.51–1.48)</td>
<td>3.11 (2.06–3.11)</td>
</tr>
<tr>
<td>Ischaemic heart disease</td>
<td>6.93 (62)</td>
<td>6.99 (3349)</td>
<td>4.46 (207197)</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 (2626)</td>
<td>1.40 (620203)</td>
<td>0.63 (0.46–0.85)</td>
<td>5.34 (5.11–5.57)</td>
</tr>
<tr>
<td>Type 2 diabetes mellitus</td>
<td>17.43 (156)</td>
<td>13.50 (5290)</td>
<td>5.51 (216187)</td>
<td>1.29 (1.01–1.52)</td>
<td>2.45 (2.38–2.53)</td>
</tr>
<tr>
<td>Smoking (current/ex)</td>
<td>86.99 (689)</td>
<td>83.87 (22120)</td>
<td>58.38 (1649589)</td>
<td>1.04 (0.97–1.11)</td>
<td>1.44 (1.42–1.46)</td>
</tr>
</tbody>
</table>

sum – sample (CardiffCOGS), sch – schizophrenia, gen – general population.

a. SRR_CardiffCOGS_gen represents the standardised rate ratio (SRR) of the clinical cohort to the schizophrenia population ascertained in SAIL. SRR_CardiffCOGS_sch represents the standardised rate ratio of the schizophrenia population ascertained in SAIL to the general population ascertained in SAIL. SRRs were standardised using the age and gender distribution from the clinical cohort as reference. crude unadjusted population rates are given in supplementary Table 3. 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). Rates given in parentheses in SAIL columns are standardised to account for differences in age and sex distribution between the SAIL and CardiffCOGS cohorts.
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 $-0.14$ to $-0.01$, $P = 0.017$) and neurodevelopmental disorders (beta = $-0.077$, 95% CI $-0.64$ to $-0.07$, $P = 0.015$) (Table 3). CNV carriers were on average 7.5–7.7 cm shorter than non-carriers.

There was no evidence for association between rare CNVs of 500 KB 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 (Fig. 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

**Fig. 1** Rates of physical health phenotypes in the Cardiff Cognition in Schizophrenia (CardiffCOG) schizophrenia sample, the schizophrenia population in Secure Anonymised Information Linkage (SAIL) and the general population in SAIL.

T2DM, type 2 diabetes mellitus.

**Table 2** Interrater agreement of the rates of ascertainment between the Cardiff Cognition in Schizophrenia (CardiffCOG) interview and linkage to National Health Service records.

<table>
<thead>
<tr>
<th>SAIL</th>
<th>CardiffCOG interview</th>
<th>% Reported affected at CardiffCOG 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>T2DM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Affected</td>
<td>75</td>
<td>13.34</td>
<td>17.69</td>
<td>0.675 (0.597–0.754)</td>
<td>0.884 (0.853–0.915)</td>
</tr>
<tr>
<td>Unaffected</td>
<td>14</td>
<td>536</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ischaemic heart disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Affected</td>
<td>21</td>
<td>5.96</td>
<td>6.12</td>
<td>0.502 (0.362–0.641)</td>
<td>0.936 (0.915–0.957)</td>
</tr>
<tr>
<td>Unaffected</td>
<td>18</td>
<td>596</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Epilepsy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Affected</td>
<td>24</td>
<td>7.81</td>
<td>5.11</td>
<td>0.529 (0.397–0.661)</td>
<td>0.935 (0.914–0.956)</td>
</tr>
<tr>
<td>Unaffected</td>
<td>28</td>
<td>604</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoking</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Affected</td>
<td>515</td>
<td>71.92</td>
<td>77.41</td>
<td>0.380 (0.307–0.452)</td>
<td>0.621 (0.567–0.675)</td>
</tr>
<tr>
<td>Unaffected</td>
<td>74</td>
<td>111</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SAIL, Secure Anonymised Information Linkage.

a 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; Gwet’s AC1 refers to the first order agreement coefficient. 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.
association between schizophrenia polygenic risk scores 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 non-response to antipsychotics and type 2 diabetes (OR = 2.94, 95% CI 1.79–4.85, \( P \leq 0.0001 \)) 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 non-response to antipsychotics and epilepsy, and smoking and ischaemic heart disease (supplementary Table 6).

**Discussion**

In this study, we report the linkage of genetic data from a clinically ascertained sample of individuals with schizophrenia to

### Table 3  Association analysis results for physical health phenotypes and two groups of copy number variants (CNVs, 12 schizophrenia CNVs and 53 neurodevelopmental CNVs)a

<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>----------------------------------</td>
<td>----------</td>
<td>--------------</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></td>
<td></td>
</tr>
<tr>
<td>Affected</td>
<td>&lt;5</td>
<td>126</td>
</tr>
<tr>
<td>Unaffected</td>
<td>11</td>
<td>576</td>
</tr>
<tr>
<td>Smoking – current or ex-smoker (GP)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Affected</td>
<td>9</td>
<td>552</td>
</tr>
<tr>
<td>Unaffected</td>
<td>15</td>
<td>656</td>
</tr>
<tr>
<td>Congenital disorders (both)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Affected</td>
<td>&lt;5</td>
<td>8</td>
</tr>
<tr>
<td>Unaffected</td>
<td>14</td>
<td>701</td>
</tr>
<tr>
<td>Epilepsy (both)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Affected</td>
<td>&lt;5</td>
<td>8</td>
</tr>
<tr>
<td>Unaffected</td>
<td>14</td>
<td>701</td>
</tr>
</tbody>
</table>

\( P \), uncorrected P-value; GP, general practice.

a. General practice/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 body mass index for which the effect size is standardised Beta. Counts below five are masked to preserve participant anonymity.

![Fig. 2](https://www.cambridge.org/core)  
**Fig. 2** Graphs of the results from regression models for the association between polygenic risk for schizophrenia and physical health outcomes.

(a) Odds ratio; (b) Beta. Odds ratios are shown for the \( P \)-value thresholds at which markers were selected.
with small effect sizes. However, it can still be concluded that genetic liability to schizophrenia does not have a large or significant impact on 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.

Study limitations

The main limitation of this study was related to sample size; it is possible that we were underpowered to detect genetic associations

Supplementary material

Supplementary material is available online at https://doi.org/10.1192/bjp.2020.42

Data availability

All data requests should be submitted to the corresponding authors for consideration.

Acknowledgements

We thank the participants, clinicians, field team and MRC Centre for Neuropsychiatric Genetics and Genomics laboratory staff for their help with the CARDIFCOGS study. We thank the staff at deCODE Genetics for sample genotyping.

Author contributions

K.K., A.J., S.C.L., S.E.L. and J.T.R.W. drafted the manuscript. K.K., J.T.R.W., S.E.L. and S.C.L. carried out the data analysis. A.J., S.C.L., S.E.L. and J.T.R.W. drafted the manuscript. All authors contributed to the statistical revision of the manuscript for important intellectual content and gave final approval of the version to be published.

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/LC_17211) to Swansea University, an MQ grant (M035/163a) to Cardiff University and the Health and Care Research Wales National Centre for Mental Health (150838) at Swansea University. K.K. was funded by a Wellcome Trust Clinical Research Training Fellowship.

Declaration of interest

The authors declare no conflicts of interest.
References


14 Wing JK, Babor T, Brugha T, Burke J, Cooper JE, Giel R, et al. SCAN. Schedules for Clinical Assessment in Neuropsychiatry. Arch Gen Psychiatry 1990; 47: 589–93.


30 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007; 81: 559–75.


