Data Resource Profile: Generation Scotland Electronic Health Record

Shona M. Kerr[1*], Archie Campbell[2], Jonathan Marten[1], Veronique Vitart[1], Andrew McIntosh[4,5], David J. Porteous[2,3] and Caroline Hayward[1]

[1] MRC Human Genetics Unit, University of Edinburgh, Institute of Genetics and Molecular Medicine, Western General Hospital, Edinburgh, U.K.

[2] Generation Scotland, Centre for Genomic and Experimental Medicine, University of Edinburgh, Institute of Genetics and Molecular Medicine, Western General Hospital, Edinburgh, U.K.

[3] Medical Genetics Section, Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK

[4] Division of Psychiatry, University of Edinburgh, Royal Edinburgh Hospital, Edinburgh, U.K.

[5] Centre for Cognitive Ageing and Cognitive Epidemiology, Department of Psychology, University of Edinburgh, Edinburgh, UK.

Key Words: EHR; Data; Biobank; Genotype; Generation Scotland

*Corresponding author: shona.kerr@igmm.ed.ac.uk

Data Resource Basics

The Generation Scotland Electronic Health Record data resource is a key component of the Generation Scotland Scottish Family Health Study (GS:SFHS)[1], a biobank conceived in 1999 (www.generationscotland.org) for the purpose of studying the genetics of health areas of current and projected public health importance.[2] GS:SFHS is a large, family-based, intensively-phenotyped cohort of volunteers from the general population across Scotland, UK. The median age at recruitment was 47 for males and 48 for females, and the cohort has 99% white ethnicity.[1] Over 24 000 adults were recruited from 2006 to 2011, with broad and enduring written informed consent for biomedical research. Specific consent was obtained from 23 603 participants for GS:SFHS study data to be linked to their Scottish National Health Service (NHS) records, using their Community Health Index (CHI) number. This identifying number is used for NHS Scotland procedures (registrations, attendances, samples, prescribing and investigations) and allows healthcare records for individuals to be linked across time and location.[3] Here, we describe the NHS electronic health record (EHR) dataset on the sub-cohort of 20 032 GS:SFHS participants with consent and mechanism for record linkage plus extensive genetic (both directly measured and imputed genotype) data. Together with existing study phenotypes, including family history and environmental exposures such as smoking, the EHR is a rich resource of real world data that can be used in research to characterise the health trajectory of participants, available at low cost and a high degree of timeliness, matched to DNA, urine and serum samples and detailed genetic information.

Data resources and population coverage

The numbers of participants with full phenotype data, genome-wide genotype and consent for record linkage are shown in Figure 1. Research data (baseline and derived subsequent to recruitment) are stored by the Generation Scotland biobank in a study database and EHR data by the NHS. EHR data can be extracted by the NHS National Services Scotland electronic

Data Research and Innovation Service (eDRIS) on GS:SFHS participants using their CHI number and put into a secure data centre, such as a safe haven. This data can be analysed through secure access protocols then de-identified individual-level study and routine medical data brought together for specific approved research projects, in line with the expectations of the participants.

GS:SFHS probands were first approached through their General Practitioner (GP) using the CHI database, which has a unique number for each individual in the >96% of the Scottish population registered with a GP.[3] Those who indicated that they and one or more of their relatives were considering participation were sent an information leaflet, a consent form and a questionnaire. The details of this recruitment of 24 090 participants to GS:SFHS have been described previously.[1] DNA was extracted from the blood or saliva of participants[4], and samples were genotyped using a genome-wide SNP array (Illumina OmniExpressExome).[5] A subset of participants was selected for genotyping, consisting of those individuals who were born in the UK, had Caucasian ethnicity, had full baseline phenotype data available from a visit to a GS:SFHS research clinic in Aberdeen, Dundee, Glasgow or Perth, and had consented for their data to be linked to their NHS records.[5] The total number of individuals genotyped was 20 128, of which 20 032 passed additional genetic quality control filtering. Coverage dates for biochemistry EHRs vary regionally across Scotland due to different dates of implementing storage of records in electronic format across the NHS Area Health Boards. Records in NHS Greater Glasgow and Clyde are available from May 2006 onwards, while some NHS Tayside records (covering the Dundee and Perth recruitment areas) go back as far as 1988. Prescribing data are available for participants in the Tayside area dating back to January 1989, in Fife from March 2008, and from the Glasgow area from November 2008. Data has been collected for the Scottish Morbidity Record (SMR01) General / Acute Inpatient and Day Case database from 1960 onwards, computerised from 1968 (http://www.adls.ac.uk/nhs-scotland/general-acute-inpatient-day-case-smr01/?detail). Records are therefore available for most participants from well before the period of recruitment to the GS:SFHS cohort (2006-2011) and subsequent to participation in the study, up to within a few weeks of the date of a data access release. The data resource includes contemporary measures that reflect current tests and treatments.

Data Collected

The study phenotype (cohort profile)[1] and subsequent genotype data (both directly typed and imputed to the Haplotype Reference Consortium release 1.1 panel)[5] have both been described previously and an open access phenotype data dictionary is available (http://dx.doi.org/10.7488/ds/2057). This paper provides the first detailed description of the electronic health record datasets available in GS:SFHS through record linkage and the process to access this data. The structured, coded variables in the biochemistry, prescribing and morbidity datasets in particular are highly valuable for a genetics biobank such as GS:SFHS. Access to a wealth of other more specialized datasets including cancer, maternity inpatient and mental health (http://www.ndc.scot.nhs.uk/National-Datasets/index.asp) is also possible through the same straightforward and transparent application process.[6] A table listing examples of routine and research data sets to which GS:SFHS data and samples can be linked anonymously, using the CHI number, has been published.[1] Once a proposal has been approved, researchers are provided with pseudo-anonymised data extracts. CHI numbers are replaced by unique study numbers and personal identifying information is removed. Numbers of participant records currently available for each of the main EHR data categories, and the proportion with genome-wide genotype data, are illustrated in Figure 2.

Ethical Clearance

GS:SFHS has Research Tissue Bank status from the Tayside Committee on Medical Research Ethics (REC Reference Number: 15/ES/0040). This provides a favourable opinion for a wide range of data uses within medical research, including genetic analyses. Permission for use of NHS EHR data in record linkage projects is obtained from the NHS Public Benefit and Privacy Panel for Health and Social Care, and (for biochemistry data only) from NHS Greater Glasgow and Clyde Health Research Informatics Unit and the Health Informatics Centre, University of Dundee. Only data from those GS:SFHS participants who gave written consent for record linkage of their GS:SFHS study data to their medical records are used.

Funding Sources

This work was supported by the Medical Research Council UK, which provides core funding to the QTL in Health and Disease research programme at the MRC HGU, IGMM, University of Edinburgh. SCI-Store (for laboratory investigation reports, http://www.sci.scot.nhs.uk/products/store/store_main.htm), the Scottish National Prescribing Information System (PIS)[7] and the Scottish Morbidity Record (http://www.ndc.scot.nhs.uk/Data-Dictionary/SMR-Datasets/) are publicly funded NHS data systems. Genotyping of the GS:SFHS DNA samples was carried out by the Genetics Core Laboratory at the Edinburgh Clinical Research Facility, University of Edinburgh and was funded by the Medical Research Council UK and the Wellcome Trust (Wellcome Trust Strategic Award “Stratifying Resilience and Depression Longitudinally” (STRADL) [Reference 104036/Z/14/Z]. Generation Scotland received core support from the Scottish Executive Health Department, Chief Scientist Office [grant number CZD/16/6] and the Scottish Funding Council [HR03006].

Data Resource Use

The genome-wide genetic data in GS:SFHS has been used in a large number of research projects across a wide range of study phenotypes that have generated over 100 publications to date (http://www.ed.ac.uk/generation-scotland/news-events/publications). The first research paper to use record linkage in GS:SFHS, on the impact of parental diabetes on offspring health, was published in 2015.[8] The first successful example of NHS record linkage for genetic research in GS:SFHS involved the identification of over 200 cases with atrial fibrillation and matched controls by linkage to hospital episode (Scottish Morbidity Record, SMR01) data (based on International Classification of Diseases (ICD-10) codes), as part of the AFGen Consortium.[9] An illustrative example of how the genetic and EHR data in GS:SFHS can be used is examining the psychiatric history of cases of major depressive disorder (MDD) and controls using record linkage to the Scottish Morbidity Record (Outpatient and Mental Health Inpatient datasets) and prescription data (for history of antidepressants). This information has been used in haplotype association analyses of MDD[10], stratification of MDD into genetic subgroups[11] and genome-wide meta-analyses of stratified depression.[12] Data science has great potential as a catalyst for improved mental health recognition, understanding, support and outcomes.[13,14]

A second example is a genome-wide association study (GWAS) of serum urate in the Tayside regional subset of the cohort.[5] Uric acid is a medically relevant phenotype measure, with high levels leading to the formation of monosodium urate crystals that can cause gout. Hyperuricaemia has additionally been associated with a variety of diseases including type 2 diabetes, hypertension and cardiovascular disease, while hypouricaemia has been linked to neurodegenerative disorders including Parkinson’s disease and Alzheimer’s disease.[15] GWAS of uric acid was performed using EHR-derived measures from 2077 individuals[5] and shows the strong signal in the *SLC2A9* gene previously reported using data gathered specifically for research[16,17,18]. This positive control confirmed that the EHR-derived biochemistry data can be suitable for population-based analyses, despite being collected for clinical purposes. The initial analysis has now been extended into the Glasgow regional subset of the cohort (Figure 1), extending the number of individuals with both genotype data and at least one uric acid measurement to 3162, within the total of 17 877 participants where genetic and biochemistry data can be accessed (Figure 2).

Hospital admission, prescription and biochemistry EHRs can be used to infer disease status in individuals after recruitment to the study has concluded, including for diseases that were not part of the initial data collection. One such example is gout, which was not explicitly included in the pre-clinical questionnaire, but with access to EHR data it is possible to ascertain gout status for participants in GS:SFHS. Not all individuals with hyperuricaemia develop gout, which may mean that other factors predispose individuals to a greater or lesser risk of disease progression. Identifying these factors could help inform targeted prevention and personalised management of the disease. The up-to-date status of risk factors available in GS:SFHS from EHR linkage makes it an excellent resource for a case-to-high-risk-control GWAS. A static study might incorrectly class an individual as a high-risk control simply because they did not develop gout until after data collection had concluded. Here, 420 gout cases have been identified through the use of urate lowering medication, obtained from the Scottish National Prescribing Information System (PIS).[7] Additional information can be obtained from the GS:SFHS baseline phenotype dataset, including self-reported use of medications and measures such as body mass index. This information, together with the range of risk factors available in the biochemistry, prescribing and morbidity (e.g. gout ICD-10 codes in SMR01) EHR datasets, will be used to select risk-matched individuals who have not developed gout, for a case-control GWAS of GS:SFHS participants.

Strengths and Weaknesses

Data quality and costs

Extensive and detailed phenotyping, including longitudinal biochemistry data, is of considerable utility in understanding underlying biological or disease mechanisms. However, this data can be difficult and expensive to obtain directly, as measures require different assays and the quantity of donated samples (e.g. serum, plasma) is finite in a biobank such as GS. Collecting longitudinal samples requires ongoing recontact with study participants, which is both expensive and time-consuming. Access to routine EHR data is therefore of great value in genetic research across broad-based medical specialities,[19] as exemplified by the Electronic Medical Records and Genomics (eMERGE) network.[20]

Scotland has some of the most comprehensive health service data in the world. Few other countries can lay claim to national indexed data of such high quality and consistency. The Generation Scotland phenotype, genotype and imputed data have been subject to extensive quality control and are research ready. The EHR biochemistry data was generated in accredited NHS laboratories for clinical use, therefore the measures are accurate, with internal and external quality control and quality assurance processes in place for all tests and investigations. Over time assay methods, instrumentation and automation protocols will have changed, but outputs have to show consistency for clinical diagnostic purposes. On occasion an arithmetic correction has to be applied, for example some of the uric acid values in the EHR dataset are in mmol/L, while most are in the more widely used µmol/L.

The biochemistry data was generated as part of routine medical care by the NHS, and only modest cost recovery charges are made to provide access to it for research purposes. This compares favourably with the costs of commissioning new tests on blood, serum or plasma samples. The Generation Scotland data access process also incurs a cost recovery charge.

An NHS Data Quality Assurance team is responsible for ensuring Scottish Morbidity Record (SMR) datasets are accurate, consistent and comparable across time and between sources (http://www.isdscotland.org/Products-and-Services/Data-Quality/). Longitudinal studies are feasible, with mechanisms also in place for re-contact of GS participants for focussed followup including recall-by-genotype studies, enabling detailed research on chronic conditions and long-term outcomes.

Although the data available is extensive, some gaps exist. For example, data from primary care (the Scottish Primary Care Information Resource, SPIRE)[21] may become available for linkage in future. This would be particularly useful for research on conditions such as dementia where much of the initial health service contact is with general practitioners. Medical imaging data (X-rays, CT scans etc) held in an NHS Scotland Picture Archiving and Communications System (http://www.nisg.scot.nhs.uk/currently-supporting/pacs-and-ris) may also become available for research in due course. Another weakness is that any features of illness that occur outside NHS Scotland will not be documented. However, the availability of GS:SFHS study data that was collected at recruitment, together with the range of different types of data available longitudinally in the EHR, mean that (for example) accurate classification of cases and controls can be achieved.

Data Resource Access

Generation Scotland operates a managed data access process including an online application form (http://www.ed.ac.uk/generation-scotland/using-resources/access-to-resources). Five of the authors of this paper sit on the Generation Scotland Access Committee (GSAC) that assesses applications for collaborative access to the data described. Proposals are reviewed by the GSAC and, following approval, researchers and their host institutions sign a data and/or material transfer agreement with GS. A favourable opinion from a Research Ethics Committee (REC) is in place for research using study data (and samples), as GS:SFHS is a Research Tissue Bank. Research that includes access to EHR data is notified to the REC by the GS management team on behalf of the researchers, through a substantial amendment. Researchers requesting EHR data must also submit their proposal to the NHS Public Benefit and Privacy Panel for Health and Social Care (http://www.informationgovernance.scot.nhs.uk/pbpphsc/home/for-applicants/). If access to biochemistry EHR data is part of the proposed research, an additional application to the two data Safe Havens (http://www.nhsggc.org.uk/about-us/professional-support-sites/nhsggc-safe-haven/ and https://www.dundee.ac.uk/hic/hicsafehaven/) holding this data is required, with a new safe haven workspace created for each project. The GS management group can guide applicants requesting access to any EHR dataset.

Resulting research papers are expected to include GS Executive member(s) and/or their GS nominees as co-authors. A standard acknowledgement paragraph is also required in resulting publications and can be found in the GS Authorship and Acknowledgement Policy at http://www.ed.ac.uk/files/atoms/files/summary_gs_aa_policy_20161216_0.pdf. For more information on all aspects of the data access process, the Generation Scotland management group should be contacted [info@generationscotland.org].

Acknowledgements

We thank staff at the University of Dundee Health Informatics Centre and the NHS National Service Scotland electronic Data Research and Innovation Service (eDRIS) for their expert assistance with EHR data linkage.

Conflict of interest: The authors declare no conflict of interest.

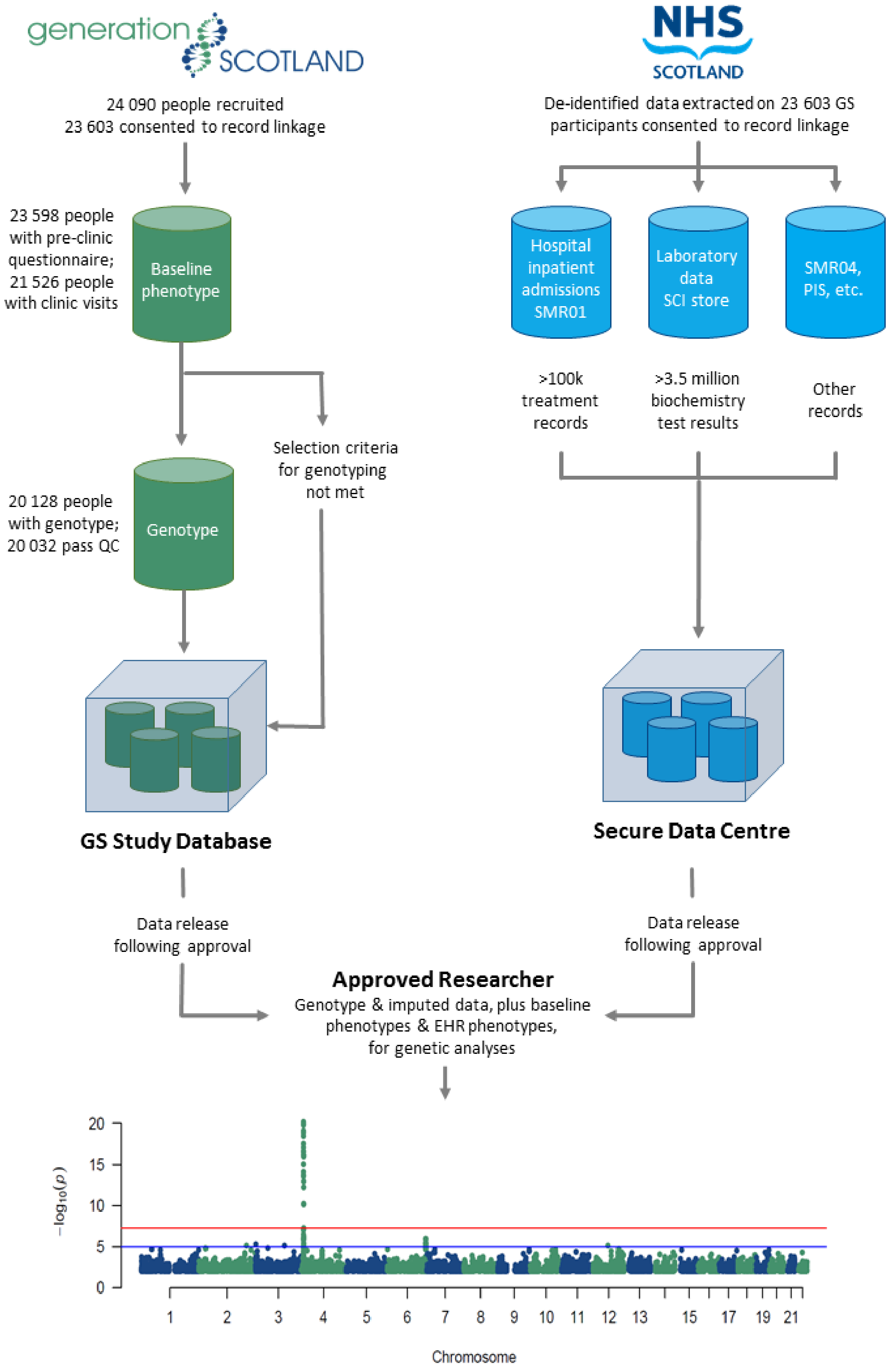
generation SCOTLAND
24 090 people recruited
23 603 consented to record linkage
23 598 people with pre-clinic questionnaire; 21 526 people with clinic visits
Baseline phenotype
Selection criteria for genotyping not met
20 128 people with genotype; 20 032 pass QC
Genotype
GS Study Database
Data release following approval
NHS SCOTLAND
De-identified data extracted on 23 603 GS participants consented to record linkage
Hospital inpatient admissions SMR01
Laboratory data SCI store
SMR04, PIS, etc.
>100k treatment records
>3.5 million biochemistry test results
Other records
Secure Data Centre
Data release following approval
Approved Researcher
Genotype & imputed data, plus baseline phenotypes & EHR phenotypes, for genetic analyses
$-\log_{10}(p)$
20
15
10
5
0
1
2
3
4
5
6
7
8
9
10
11
12
13
15
17
19
21
Chromosome

Figure 1. Schematic illustrating the scope of the resource. The datasets available in Generation Scotland and EHRs are indicated, with numbers of participants and records. The Manhattan plot displays the results of a genome-wide association analysis using genotyped SNPs and EHR-derived serum urate measurements, as an example of how the datasets can be can be used together in genetic research by approved researchers. The single highest serum urate reading was taken for each participant, with covariates and methods for accounting for relatedness as previously described.[5] The $-\log_{10}$ (P-value) is plotted on the y-axis, and chromosomal location is plotted on the x-axis. The genome-wide significance threshold accounting for multiple testing (p-value < 5 x 10^{-8}) is indicated by a red line, while suggestive significance (p-value < 10^{-5}) is indicated by a blue line.

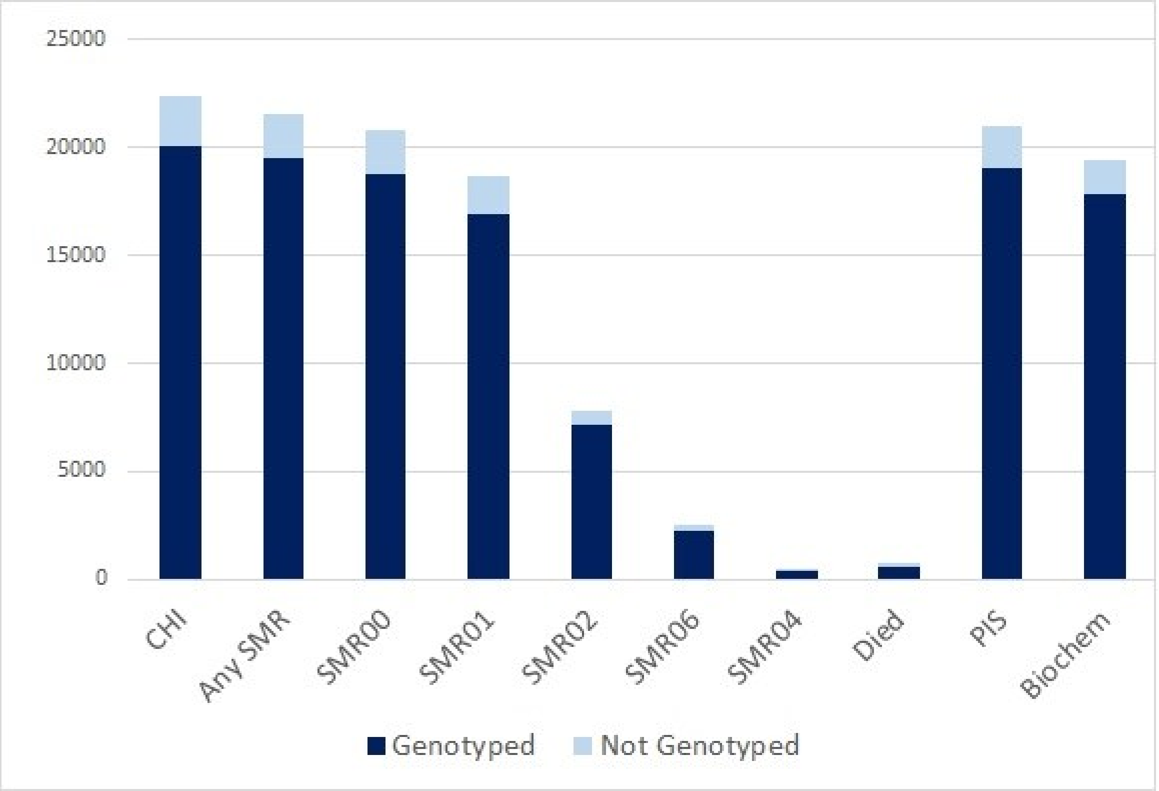

Figure 2. GS:SFHS participants with categories of EHR data available through CHI linkage. The light blue bars indicate the numbers of participants with linked data but not genetic data available. The dark blue bars indicate the numbers with genome-wide genotype data (Illumina OmniExpressExome array) available in each EHR dataset. CHI, Community Health Index;

Any SMR, Scottish Morbidity Record (n = 19 499); SMR00, Outpatient (n = 18 803); SMR01, General acute/Inpatient (n = 16 918); SMR02, Maternity Inpatient (n = 7,186); SMR06, Scottish Cancer Registry (n = 2,266); SMR04, Mental Health Inpatient (n = 498); Died (n = 634); PIS, Prescribing Information System (n = 19 035); Biochem, biochemistry laboratory data (n = 17 877).

References

1. Smith BH, Campbell A, Linksted P, et al. Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. *Int J Epidemiol* 2013; **42**: 689-700.
2. Smith BH, Campbell H, Blackwood D, et al. Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability. *BMC Med Genet* 2006; **7**: 74.
3. Pavis S, Morris AD. Unleashing the power of administrative health data: the Scottish model. *Public Health Res Pract* 2015; **25**: e2541541.
4. Kerr SM, Campbell A, Murphy L, et al. Pedigree and genotyping quality analyses of over 10,000 DNA samples from the Generation Scotland: Scottish Family Health Study. *BMC Med Genet* 2013; **14**: 38.
5. Nagy R, Boutin TS, Marten J, et al. Exploration of haplotype research consortium imputation for genome-wide association studies in 20,032 Generation Scotland participants. *Genome Med* 2017; **9**: 23.
6. Campbell A. Generation Scotland: using data linkage for longitudinal studies. *2017* 2017; **1**.
7. Alvarez-Madrazo S, McTaggart S, Nangle C, Nicholson E, Bennie M. Data Resource Profile: The Scottish National Prescribing Information System (PIS). *Int J Epidemiol* 2016; **45**: 714-5f.
8. Aldhous MC, Reynolds RM, Campbell A, et al. Sex-Differences in the Metabolic Health of Offspring of Parents with Diabetes: A Record-Linkage Study. *PLoS One* 2015; **10**: e0134883.
9. Consortium AF, ISGC MCot, Neurology Working Group of the CC. Large-scale analyses of common and rare variants identify 12 new loci associated with atrial fibrillation. *Nat Genet* 2017.
10. Howard DM, Hall LS, Hafferty JD, et al. Genome-wide haplotype-based association analysis of major depressive disorder in Generation Scotland and UK Biobank. *bioRxiv* 2017.
11. Howard DM, Clarke T-K, Adams MJ, et al. The Stratification Of Major Depressive Disorder Into Genetic Subgroups. *bioRxiv* 2017.
12. Hall LS, Adams MJ, Arnau-Soler A, et al. Genome-Wide Meta-Analyses Of Stratified Depression In Generation Scotland And UK Biobank. *bioRxiv* 2017.
13. McIntosh AM, Stewart R, John A, et al. Data science for mental health: a UK perspective on a global challenge. *Lancet Psychiatry* 2016; **3**: 993-8.
14. Smoller JW. The use of electronic health records for psychiatric phenotyping and genomics. *Am J Med Genet B Neuropsychiatr Genet* 2017.
15. Kutzing MK, Firestein BL. Altered uric acid levels and disease states. *J Pharmacol Exp Ther* 2008; **324**: 1-7.
16. Vitart V, Rudan I, Hayward C, et al. SLC2A9 is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout. *Nat Genet* 2008; **40**: 437-42.
17. Li S, Sanna S, Maschio A, et al. The GLUT9 gene is associated with serum uric acid levels in Sardinia and Chianti cohorts. *PLoS Genet* 2007; **3**: e194.
18. Doring A, Gieger C, Mehta D, et al. SLC2A9 influences uric acid concentrations with pronounced sex-specific effects. *Nat Genet* 2008; **40**: 430-6.
19. Wei WQ, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome Med* 2015; **7**: 41.
20. Gottesman O, Kuivaniemi H, Tromp G, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet Med* 2013; **15**: 761-71.
21. SPIRE. Available from http://www.spire.scot.nhs.uk/. 2017.