This is version 1 for bioRxiv submission 25 Sep 2018 with data linkage R7 (Feb 2018).
We plan to update to version 2 in bioRxiv with the next data linkage (Oct 2018) and submit to a journal.

Cohort Profile: East London Genes & Health (ELGH), a community based population genomics and health study in people of British-Bangladeshi and -Pakistani heritage.

Author List:

Sarah Finer[1], Hilary Martin[2], Karen Hunt[1], Beverley MacLaughlin[1], Richard Ashcroft[3], Ahsan Khan[4], Mark I McCarthy[5,6,7], John Robson[1], Daniel MacArthur[8,9], Chris Griffiths[1] John Wright[10], Richard C Trembath[11], David A van Heel[1]

Correspondence to s.finer@qmul.ac.uk or d.vanheel@qmul.ac.uk

Affiliations:

1. Blizard Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, 4 Newark Street, London, E1 2AT, UK.
2. Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, UK.
3. Department of Law, Queen Mary University of London, Mile End Road, London, E1 4NS, UK.
4. Waltham Forest Council, Waltham Forest Town Hall, Forest Road, Walthamstow, E17 4JF, UK.
5. Wellcome Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford, OX3 7BN, UK.
6. Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Old Road, Headington, Oxford, OX3 7LJ, UK.
7. Oxford NIHR Biomedical Research Centre, Churchill Hospital, Old Road, Headington, Oxford, OX3 7LJ UK
8. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA.
9. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA.
10. Bradford Institute for Health Research, Bradford Teaching Hospitals National Health Service (NHS) Foundation Trust, Bradford, BD9 6RJ, UK.
11. School of Basic and Medical Biosciences, Faculty of Life Sciences and Medicine, King’s College London, London, SE1 1UL, UK.

Cohort profile in a nutshell

- East London Genes & Health (ELGH) is a large scale, community genomics and health study (to date >30,000 volunteers; target 100,000 volunteers).
- ELGH was set up in 2015 to gain deeper understanding of health and disease, and underlying genetic influences, in people of British-Bangladeshi and -Pakistani heritage living in east London.
- ELGH prioritises studies in areas important to, and identified by, the community it represents. Current priorities include cardiometabolic diseases and mental illness, these being of notably high prevalence and severity. However studies in any scientific area are possible, subject to community advisory group and ethical approval.
- ELGH combines health data science (using linked UK National Health Service (NHS) electronic health record data) with exome sequencing and SNP array genotyping to elucidate the genetic influence on health and disease, including the contribution from high rates of parental relatedness on rare genetic variation and homozygosity (autozygosity), in two understudied ethnic groups. Linkage to longitudinal health record data enables both retrospective and prospective analyses.
- Through stage 2 studies, ELGH offers researchers the opportunity to undertake recall-by-genotype and/or recall-by-phenotype studies on volunteers. Sub-cohort, trial-within-cohort, and other study designs are possible.
- ELGH is a fully collaborative, open access resource, open to academic and life sciences industry scientific research partners.

Why was the cohort set up?

East London Genes & Health (ELGH) commenced recruitment in April 2015 as a community based, long term study of population health and disease in people of British-Bangladeshi and -Pakistani heritage in east London. ELGH uses a novel population-based design incorporating cutting-edge genomics with high-quality electronic health record data linkage and targeted genotype-based-recall studies in currently >30,000 volunteers with funding to expand to 100,000 volunteers by 2023. ELGH is an open access data resource, building on the expertise of studies including UK Biobank, and has been designed to generate new knowledge related to the health and disease of an population at high need, and to redress the poor representation of non-White ethnic groups in existing population genomic cohorts[1].

Almost a quarter of the world's population is of South Asian origin with over 3 million in the UK, representing 5% of the UK population[2]. The risk of coronary heart disease is 3-4 times higher, and type 2 diabetes 2-4 times higher in UK South Asians compared with Europeans[3,4]. Understanding the mechanisms underlying these ethnic differences will provide important insights into the aetiology of cardiometabolic diseases to inform new approaches to treatment and prevention, and help to reduce ethnic inequalities.

The setting for ELGH, in east London, incorporates one of the UK's largest South Asian communities (29% of a total population of 1.95 million people across its 8 local authorities), of which 70% are people of British-Bangladeshi and -Pakistani heritage. This population lives in high levels of deprivation (Tower Hamlets, Hackney, Barking and Dagenham are the 9th, 10th and 11th most deprived local authorities in England)[5] and it experiences disproportionately adverse health outcomes, especially relating to cardiometabolic health and its complications. Compared to White Europeans, South Asians living in east London have a two-fold greater risk of developing type 2 diabetes (16.4% vs. 7.5%)[6]; and faster progression of chronic kidney disease in those with diabetes[7], nearly double the risk of non-alcoholic liver disease[8], and over double the risk of multimorbidity including cardiovascular disease[9], and the onset of cardiovascular disease occurs 8 years earlier in men (60.4 years compared to 68.2 years)[9]. Determinants of poor cardiometabolic health are also noted to start early in the life course, with 35-44% of 10-11 year old children overweight/obese in east London boroughs, above the UK average of 33%[5]. Deprivation, ethnicity related ill-health, and inequalities all combine to lead to worse health outcomes.

A key feature of ELGH is the opportunity to obtain, and link, high quality individual level data from routine, real world, longitudinal (retrospective and prospective) clinical data sources to genomic data, with the ability to recall for further research studies. East London has an extensive track record of utilising routine clinical health care data (predominantly from primary care) in research studies[6–8]. Routinely collected health record data is of high quality, and electronic performance dashboards are embedded (and sometimes incentivised) in UK clinical practice, facilitating both high quality clinical care and disease monitoring[10]. However clinical data is different to traditional epidemiological researcher-collected (or participant-recorded) data, and presents different challenges in using it to measure health outcomes. These challenges are being addressed both internationally (e.g. the SNOMED collaboration) and within the UK by e.g. establishment of the new Health Data Research UK. Interestingly, population-based disease risk screening programmes, such as the National Health Service (NHS) Health Checks, are widely taken up in East London and, notably, show greatest uptake in people of South Asian ethnicity and from the most deprived quintiles[11].

ELGH volunteers report high rates of parental relatedness (e.g. are offspring of first cousin parent marriages, ***Table 1***) which leads to genomic regions of autozygosity at the DNA level, such that rare allele frequency variants normally seen as heterozygotes can be observed in the homozygous (autozygous) state. Whilst the effects of autozygosity is well studied in paediatrics and rare disease, there is much less previous research on any health effects [12,13] in adults recruited from population settings.

ELGH supports the health needs and priorities of the local community, and fosters authentic, long term engagement in its research, so that benefits to health can be delivered in the future. A community advisory group is embedded into the high level strategic management of ELGH and has helped prioritise areas for research, including type 2 diabetes, cardiovascular disease, dementia and mental health. The community focused ethos extends to a wide range of meaningful public engagement and dissemination activities, including collaboration with the award-winning Centre of the Cell science and health education facility[14].

Who is in the cohort?

ELGH (see ***Figure 1***) incorporates population-wide recruitment to Stage 1 studies, and targeted recruitment to Stage 2 recall-by-genotype studies. Stage 3 and 4 studies are planned.

During Stage 1, ELGH invites voluntary participation of all individuals aged 16 and over, of British-Bangladeshi and -Pakistani heritage living in, or within reach of, east London. There are no other exclusions to joining ELGH. The recruitment approach is wide reaching and inclusive, being undertaken in two settings: (a) community settings supported by an embedded third-sector partner organisation and volunteers from the local community, and (b) clinical settings supported by a team of bilingual health researchers, many of whom have British-Bangladeshi or -Pakistani heritage themselves. Volunteers who agree to participate in Stage 1 complete a brief participant questionnaire, give consent to electronic health record linkage (lifelong access to primary, secondary and mental health care, and national datasets and disease registries), and donate a saliva sample for DNA extraction and genetic tests. The data collected, including description of Stage 1 ELGH volunteers, are described in the next section. Between April 2015 and September 2018, ELGH has recruited 31,607 participants to Stage 1, with a recruitment target of 100,000 by 2023. At the most recent data linkage (*R7, Feb 2018*), 97.1% of 27,927 had valid NHS numbers, 65% had linked East London comprehensive primary care health record data available, 8720 volunteers had linked secondary care (Barts Health NHS Trust) data with at least one ICD-10 coded hospital encounter, including day case, inpatient, outpatient, maternity and emergency care. By 2019, near-complete (>95%) linkage to primary care health records is expected with improvements in health data connectivity and flow through initiatives supported by Health Data Research UK and the Discovery programme. Recruitment into outer London regions along with an additional study site (Bradford) is planned for late 2018 / 2019, areas with similar ethnic populations and comparable health needs.

Summary data from the stage 1 volunteer questionnaire and electronic health record data linkage are summarised in ***Table 1***, and includes both baseline characteristics and data captured from longitudinal electronic health records. Basic demographics (age group and sex) of ELGH volunteers are compared to population-wide data in ***Figure 2***. The comparison of the ELGH to the background population highlights that the 'convenience sampling' approach to volunteer recruitment in ELGH is obtaining a sample that is broadly representative of the background population, with regards age and sex, but which modestly favours recruitment of women over men in those <45 years. Data in ***Table 1*** also indicates that ELGH volunteers live in areas of high deprivation (97% live in the most deprived 2 quintiles, using the Index of Multiple Deprivation). ELGH volunteers have a high proportion of common medical conditions, including type 2 diabetes (22%), hypertension (17%), ischaemic heart disease (5%) and asthma (11%), reflective of their prevalence in the background population (5%, 9%, 2%, 5%, respectively) but also with evident over-sampling of people with chronic conditions[15].

East London Genes and Health operates under ethical approval, 14/LO/1240, from London South East NRES Committee of the Health Research Authority, dated 16 Sep 2014. All ELGH scientific activities, including applications from external researchers, are reviewed by its Executive and Community Advisory Group. The study is a National institute for Health Clinical Research Network Portfolio study supported by the NHS in England.

How often have they been followed up?

ELGH contains real-world data with data collection triggered by a broad range of clinical encounters, including routine health checks, chronic disease management, inpatient hospital admissions, surgery, maternity care and emergency care. Primary health care records in east London were digitised around 2000 and offer a rich source of data on clinical encounters since then, but also including the dates of diagnoses pre-digitisation (e.g. type 2 diabetes, diagnosed in 1992) and summarised prior clinical events.

Health data extraction and linkage takes place 3-monthly and ELGH volunteers have consented for life long access to electronic (and paper) health records (primary care, secondary care, community and mental health, national NHS datasets and registries), facilitating long term prospective follow-up.

ELGH can invite volunteers to Stage 2 studies up to 4 times per year for more detailed study visits, e.g. recall by genotype (RbG) and/or phenotype, for clinical assessment and collection of biological samples at each visit, subject to ethics approval. Other research studies (e.g. other imaging studies, other biosamples, or clinical trials) are possible subject to their own ethics approvals along with volunteer acceptability and community advisory group approval. At September 2018, around 40 ELGH volunteers have participated in Stage 2 RbG studies.

What has been measured?

Stage 1 incorporates the following procedures and data collection (summarised in ***Table 2***):

- **Participant questionnaire**

Participant stage 1 questionnaire (see *supplementary file 1*): This self-report questionnaire collects brief data on all participants including: name, date of birth, sex, ethnicity, contact details, diabetes status, parental relatedness (consanguinity, manifest as autozygosity at the genomic DNA level), and an overall assessment of general health and wellbeing. This short questionnaire has been deliberately designed and minimised to facilitate high throughput recruitment and inclusivity of groups where language and cultural differences exist, and to be used with or without researcher assistance, thereby maximising the representativeness of our population sample.

Health record data is obtained by linkage to a participant's NHS number (available for >99% of volunteers), either recorded at the time of recruitment, or at later look-up, and including an NHS number validation step (a check digit).

- **NHS primary care health record data linkage**

A data extraction template is used to extract relevant fields from electronic health record systems (both primary and secondary care data sources are accessible). Whilst raw data is potentially available, at present we are restricting analyses to directly curated research phenotypes and growing these both incrementally and on demand. The data extraction template comprises bespoke search terms, including SNOMED, ICD10 and READ (including READ2 and CTV3) diagnostic codes, prescribing data, laboratory test results and clinical measurements and processes (an example relating to type 2 diabetes is given in supplementary file 2). Electronic health record data are of high quality, and is particularly rich in settings where routine data collection is standardised and incentivised, such as by the Quality and Outcomes Framework used in NHS primary care. All electronic health record data are cleaned and checked prior to analysis. Data concordance was checked between participant questionnaire and electronic health record, with >99% concordance for gender and year of birth. Technical errors with optical character recognition (dates of birth) or user data completion (e.g. a Mr with a male first name ticking female) explained almost all cases of discordance, and were resolved with manual checking in the final analysis datasets. Data outside clinically plausible ranges (e.g. primary care-measured systolic blood pressure <60mmHg or >250mmHg, diastolic blood pressure <30mmHg or >200mmHg), or with clear data entry errors (e.g. height recorded as 167 metres instead of 1.67 metres) are removed. For the purposes of this data summary, anthropometric measurements recorded historically in the electronic health records were only used if the volunteer was >16 years old at the time of measurement. Missing data exist, but at relatively low frequency in routinely collected and incentivised clinical measures, e.g. smoking status has been recorded in the primary care record of 90% of ELGH volunteers in the 5 years prior to the most recent data linkage. Repeated measures of routinely collected data, and cross-validation across information sources can mitigate the impact of missing data where it exists, and statistical techniques, such as sensitivity analysis (for missing-not-at-random data) and multiple imputation (for missing-at-random) will be required in data analysis[16].

Electronic health record data represents a ‘living’ dataset allowing both retrospective, cross-sectional and prospective data collection.

- **NHS secondary care health record data linkage**

Barts Health NHS Trust is the major network of secondary care hospitals across in East London, and is the UK's largest NHS Trust. Currently, NHS number of ELGH volunteers are linked to the Barts Health data warehouse, containing clinician-coded SNOMED acute and chronic problem lists, laboratory results, pathology results, imaging results, and ICD-10 clinical coding which is used at every finished episode of care. Data is available for all ELGH volunteers who have attended the Barts Health hospital system - at the last linkage, this included 8720 ELGH volunteers. As an exemplar, maternity data linkage within Barts Health identified 2402 female ELGH volunteers with maternity records available for at least one pregnancy (2972 live single live births, 27 twin/multiple live births and 17 stillbirths). We intend to expand secondary care data linkage to other local East London NHS Trusts in 2019.

- **Planned linkage to health record datasets**

In 2019, East London Genes & Health will link to further datasets, including:

NHS Hospital Episode Statistics (anticipated linkage in 2019). These are national datasets curated by NHS England and NHS Digital, and are compulsorily provided by all NHS Trusts, clinical commissioning groups and local area teams in England[17]. Data comprises: admitted inpatient data - from 1997, including Admissions and Discharge, diagnosis and operation codes, maternity, psychiatric, critical care; outpatient data - from 2003; accident and Emergency data - from

2008. ICD-10 (World Health Organisation International Classification of Diseases and Related Health Problems) is used for diagnoses, and OPCS-4 (Office for Population, Censuses and Surveys) for operative procedures.

NHS Mortality Data (anticipated linkage in 2019)[18]. These are national datasets curated by the Office for National Statistics and NHS Digital. Data is linked to the Hospital Episode Statistics dataset, and provides additional valuable information, such as the cause of deaths and deaths outside of hospital.

Other potential data linkages in the future include national cancer datasets (National Cancer Registration and Analysis Service, NCRAS) and cardiovascular disease audits managed by the National Cardiovascular Outcomes Research (NICOR).

- **Genomics**

DNA is extracted from Oragene (DNA Genotek) saliva system and stored from all Stage 1 participants.

To date, low/mid depth exome sequencing has been performed (n=3781, data available) or is in progress (n=1492) on those participants reporting parental relatedness in the participant questionnaire.

In late 2018/2019 (funding secured) 50,000 samples from all stage 1 volunteers will be genotyped on the Illumina Infinium Global Screening Array v2.0 (with additional 46,662 Consortia defined multi-disease variants)[19]. Array content includes variants selected for rare disease mutations, from large exome sequencing projects, pharmacogenomics and for genome wide coverage, enabling association studies, polygenic risk score, Mendelian randomisation studies.

In 2019/2020, if support is secured from an evolving Life Sciences Industry Consortium, high-depth exome sequencing will be performed on up to 50,000 volunteer samples.

By 2023, the intention (subject to funding) is for both genotyping and high depth exome sequencing will be performed on up to 100,000 volunteer samples.

- **Samples for other -omics**

ELGH takes “core” study samples from all volunteers recalled for stage 2 or later stage studies, including a blood cell pellet (for DNA, protein, and other assays), plasma aliquots, and a blood cell RNA preservation tube (Paxgene) to enable further studies including methylation assays, transcriptomics, proteomics, lipidomics and metabolomics.

What has it found: key findings and publications?

East London Genes and Health is a new resource that continues to grow, and to date has been used for three main areas of work:

- **Characterisation of common phenotypes**

Using Type 2 diabetes as an exemplar condition, we show the feasibility of the ELGH study design to generate high quality electronic health record data for phenotypic characterisation of volunteers (***Table 3***). Of 19165 participants in ELGH with available linked electronic health record data, 4312 (22%) participants have a diagnosis of Type 2 diabetes (T2D) in their primary care record. Basic sociodemographic data (age, gender, ethnicity) of participants was recorded in

100%, and smoking status in 96% of these volunteers had been obtained within 2 years of the most recent data linkage (in 2018). Country of birth was incompletely recorded, but at least half of ELGH participants with T2D were born in Bangladesh or Pakistan. Real world clinical data recorded is of high quality, with body mass index, markers of glucose control (HbA1c) and serum cholesterol measured within 2 years prior to participating in ELGH in at least 96% of participants with T2D. The high uptake of routine care processes and high quality data capture from these shows the potential for ELGH to study participants in cross-section at study entry from electronic health record data. Hypertension, ischaemic heart disease and chronic kidney disease were observed in 47%, 15% and 10% of the 4312, and erectile dysfunction was present in 26% of men. Retinal complications of T2D are recorded and graded in the electronic health record, with 35% of ELGH participants having retinopathy and/or maculopathy in screening undergone within the last 2 years. Prescribing data is available on all ELGH volunteers with T2D, showing recent insulin prescriptions in 16%, and the use of single or multiple non-insulin agents as well as drugs for the prevention of cardiovascular disease (e.g. lipid lowering therapy).

Data summarised from ELGH participants with T2D also shows the potential for ELGH to study phenotypic traits longitudinally, both retrospectively and prospectively. The median duration of T2D in ELGH participants was 9 years (range 0-50 years) with electronic health record data available during this time. All volunteers with T2D had a year of onset of the condition recorded, and clinical measurements (including body mass index, HbA1c and serum cholesterol) at the time of diagnosis (+/- 6 months) was available for nearly two-thirds of participants. Historic prescribing data was available for similar proportions of participants (data not shown). A diagnosis of pre-diabetes has been made prior to diagnosis of T2D in 23% (993) of these individuals, and 16% (350) of women had a prior diagnosis of gestational diabetes, with clinical data available during at these times, highlighting the potential to obtain longitudinal data to inform prevalence and progression of disease states within ELGH volunteers.

Multimorbidity is an increasing problem in populations with high rates of chronic long-term conditions and ageing, in the ELGH population we identified a high rate of cardiovascular multimorbidities (including hypertension, stroke, ischaemic heart disease, heart failure, atrial fibrillation, chronic kidney disease stage 3+, advanced diabetic retinopathy) associated with type 2 diabetes. Only 14% of ELGH volunteers with type 2 diabetes (n=4312) had this as a single condition; 30% had 2 cardiovascular multimorbidities, 27% had 3, and 29% had 4 or more (***Table 3***).

The high quality of the electronic health record data available reflects the robust clinical care systems and incentivised data collection methods[10] used in east London. These data, and the consent procedures facilitating lifelong access, will provide an invaluable longitudinal data resource to facilitate genomic studies (e.g. linking phenotype to rare gene variants in Stage 2 recall studies), future at-scale population studies of common phenotypes (Stage 3) and intervention studies (Stage 4).

- **Rare allele frequency gene variants occurring as homozygotes, including predicted loss of function knockouts.**

The British-Bangladeshi and -Pakistani populations of east London have high rates of parental relatedness (ELGH volunteers self-report ~20%). All those volunteers self-reporting parental relatedness have been selected for exome sequencing. Genomic autozygosity (homozygous regions of the genome identical by descent from a recent common ancestor) means that rare allele frequency variants normally only seen as heterozygotes are enriched for homozygote genotypes. We and others previously investigated the health and population effects of such variants, with a focus on predicted protein loss of function variants[12,20,21], in smaller samples of Pakistani ethnicity, and we now expand the datasets with the ELGH study.

To inform analyses using self-reported parental relatedness, we tested the accuracy of this self reported trait to actual autozygosity measured at the DNA level by exome sequencing (***Figure 3***). We find that whilst self-reported parental relationship is a modest predictor of actual autozygosity, for example 8.2% of individuals who declare that their parents are not related in fact have >2.5% genomic autozygosity. We find that for British-Bangladeshi subjects mean autozygosity is slightly lower than expected given the reported parental relationship (possibly due to confusion over the meaning of e.g. “first cousin” versus “second cousin”), whereas for British-Pakistani subjects mean autozygosity is slightly higher than expected (possibly due to historical parental relatedness).

- **Recall by genotype (and/or phenotype) studies**

Recall-by-genotype (RbG) studies, applied to population cohorts with genomic data, are of increasing interest to researchers[22]. Studies that recall by extremes of phenotype form classical epidemiological investigations but suffer from the limitations of observational studies. RbG studies use the random allocation of alleles at conception (Mendelian Randomisation) which aid causal inference in population studies and reduce problems seen with observational studies. RbG studies can be based on genotype groups at a single variant (or an allelic series for a gene), but also permit polygenic variant designs (e.g. extremes of polygenic risk scores). RbG studies may focus on functional variants believed to alter biological pathways (e.g. amino acid changes in critical protein regions, or complete gene knockouts). RbG studies are efficient in that individuals with rare genotypes can be directly recalled from a large population sample.

ELGH is undertaking RbG studies, using its ‘Stage 2’ procedures, whereby existing volunteers may be invited to participate in recall studies up to 4 times per year. Recall studies may incorporate bespoke clinical phenotyping tailored to the genotype or phenotype under study. To date, ELGH has worked with two consortia of researchers on RbG studies, one interested in immune phenotypes who are recalling volunteers with loss-of-function gene variants in relevant genes, and another phenotyping individuals with rare variants in genes relevant to Type 2 diabetes and obesity. Successful recall completion rates to these ELGH RbG studies are between 30-40%. External academic and life sciences industry researchers may apply to ELGH to undertake RbG studies. All applications for RbG studies are subject to ELGH executive and community advisory group approval.

What are the main strengths and weaknesses?

ELGH has multiple strengths as a large, population-based study. First, its large scale and novel design offers many opportunities to investigate health and disease longitudinally (with both retrospective and prospective data) and in cross-section. ELGH reaches a British-Bangladeshi and -Pakistani population with a high burden of disease and disease risk, and affected by high levels of socioeconomic deprivation, generalisable to a wider global population. The use of a community-based approach to recruitment offers a broad reach into the target population, therefore ensuring generalisability. The study remit is congruent with the needs and wishes of this population, and its embedded community advisory group ensures acceptability of all studies being undertaken within ELGH.

To date, ELGH has modestly over-recruited British-Bangladeshi versus British-Pakistani volunteers, partly because the study first commenced in Tower Hamlets, London with mainly a British-Bangladeshi population. So as to recruit even numbers, we are expanding recruitment into outer London boroughs and will open a Bradford Genes & Health site in 2019. Bradford has a large British-Pakistani population, with many similar health issues to the community in London.

High rates of autozygosity lead to homozygous genotypes at variants with rare allele frequencies that are usually only found as heterozygotes in European ethnicity populations, including variants leading to gene loss of function. We previously reported an unexpected phenotype of an individual with a *PRDM9* gene knockout[12], and others reported the phenotype of individuals with *APOC3* knockouts[21], using this study design. The study also permits investigation of the impact of autozygosity and individual genotypes on health through population-level and individual studies. As an open-access data resource, ELGH offers an important platform for RbG studies to drive high-impact, world-class translational studies designed by academic and industrial researchers.

The use of real-world electronic health record data in ELGH is both a strength and weakness of the study. Strengths include the high quality of data available on multiple diseases and disease risks via primary care, the ability to recruitment of large numbers of participants in a feasible and cost-effective manner, and the ability to study participants longitudinally - with data available for decades prior to study participation as well as lifelong consent into the future. However, it has not yet been possible to link data to 34% of participants. At present, secondary care data is only linked to through the most local NHS Trust, Barts Health, and this may miss data from other hospital providers. Data linkage will improve in 2019, with greater coverage of primary care data access and the new Health Data Research UK, as well as linkage to national registries and databases. The real-world nature of electronic health record data may be inferior to a traditional observational studies in ascertaining recent diseases, particularly those of minor severity (which do not require necessity healthcare access) or subclinical disease or disease risk states. Additionally, whilst outcomes can be studied relatively well, real world data gives researchers limited ability to study exposures of interest, e.g. health behaviours, diet or other environmental influences, that are not part of routine clinical data collection.

ELGH is not a traditionally designed epidemiological cohort with deep data collection at recruitment (such as would be expected in a birth cohort) but does reflect an increasing trend towards pragmatic, 21st century health data-driven population study design[23]. The ability to invite all Stage 1 participants to recall studies opens the possibility to develop sub-cohorts (including collection of research grade data, as well as routine clinical care data), trials-within-cohorts and other innovative study designs in the future.

Can I get hold of the data? Where can I find out more?

External researchers and invited to participate in ELGH through the use of data generated in Stage 1 and Stage 2, as well as through the design of bespoke Stage 2 studies targeting gene variants and/or phenotypes of interest. ELGH offers an open-access resource to researchers, whether national or international, academic or industrial.

Data access is managed at several levels depending on the sensitivity and identifiable nature of the data:

- Level 1 - fully open data. We distribute summary level data via our website e.g. current summary genotype counts and annotation of knockout variants from exome sequencing of 3,782 volunteers (www.genesandhealth.org/research/scientific-data-downloads) to date downloaded by >100 users.

- Level 2 - Genotype data (from SNP chip genotyping, or high throughput sequencing) is (or will be made) available under Data Access Agreement. Individual sequencing (e.g. cram) and genotype files (e.g. vcf) are available within

6 months on the European Genome- phenome Archive[24] (EGA). Access approval is granted by the Wellcome Sanger Institute Data Access Committee, who are independent of ELGH investigators.

- Level 3 - Individual-level phenotype data is held in an ISO27001 and NHS Information Governance compliant Data Safe Haven environment under Data Access Agreement, which also contains the latest genetic data linked to the questionnaire and health record phenotypes. This "bring researchers to the data" model allows us to present the most recent data to researchers easily, update data easily, maintain complex linkages between multiple datasets easily, and avoids multiple large file data transfers for genomic datasets. This model also permits us to reassure volunteers that their sensitive health data will be carefully looked after - in particular providing maximum security against large data breaches (e.g. as experienced by Facebook and British Airways in 2018). External data export is controlled, and individual level data export will not be allowed without very good reason. The current data safe haven is the UK Secure e-Research Platform[25], hosted by Swansea University, based on the SAIL databank[26], and supporting Dementias Platform UK amongst other UK cohort studies.

ELGH also supports research studies recalling volunteers by genotype or phenotype (local, external and industry). Two RbG studies, both led by non-ELGH academic researchers, are underway. The first stage 2 RbG study led by a life sciences industry partner is about to commence. External researchers and consortia are able to apply to undertake research with East London Genes and Health via a formal application process, the details of which are available on its website. Applications are assessed by both the Executive Board and Community Advisory Group, according to community prioritisation and acceptability and scientific merit. Most external researchers will be required to have their own research ethics approval in order to work with ELGH.

Funding and competing interests

We acknowledge funding from the Wellcome Trust (102627, 210561), the Medical Research Council (M009017), Higher Education Funding Council for England Catalyst, Barts Charity (845/1796), and research delivery support from the NHS National Institute for Health Research Clinical Research Network (North Thames). We thank all volunteers for participating in the study. We are most grateful for the generous participation of volunteers who have provided samples and health record access, and have consented for recall. We thank all the general practitioners in the local health system who, as producers/controllers of this routinely collected data, have made it available for research for patient benefit and the Queen Mary University of London Clinical Effectiveness Group Research Team for making these data accessible and assisting in interpretation of coding and construction of clinical concepts. Conflict of interest: none declared.

Figures and tables

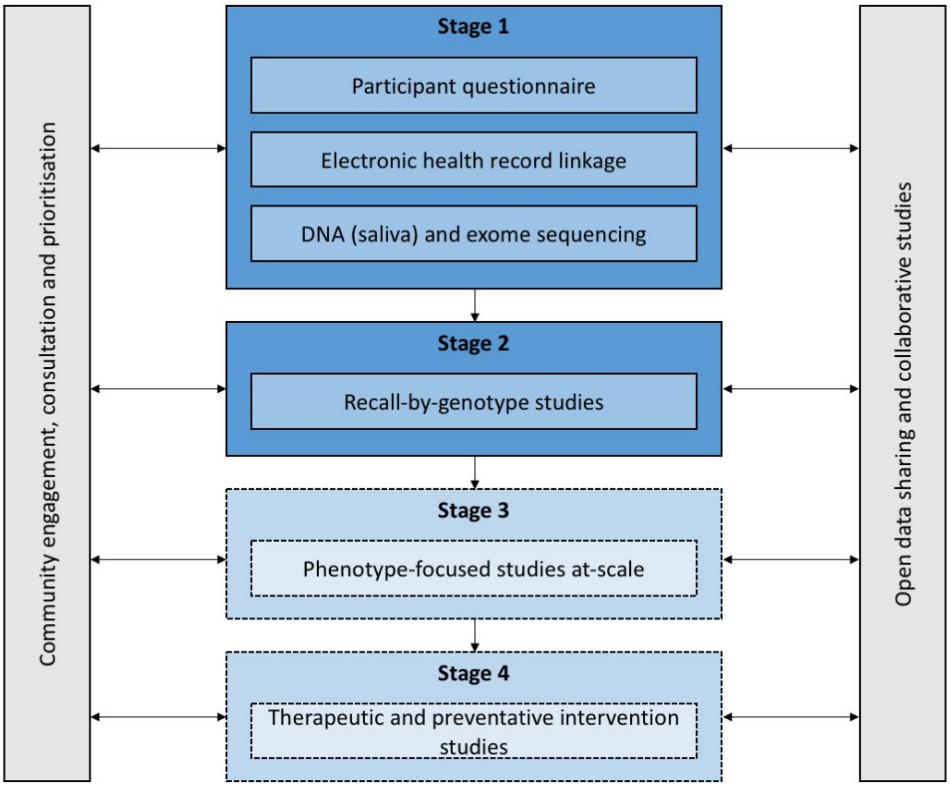

Community engagement, consultation and prioritisation
Stage 1
Participant questionnaire
Electronic health record linkage
DNA (saliva) and exome sequencing
Stage 2
Recall-by-genotype studies
Stage 3
Phenotype-focused studies at-scale
Stage 4
Therapeutic and preventative intervention studies
Open data sharing and collaborative studies

Figure 1: ELGH study design. Stage 1 and 2 studies have commenced. Stage 3 and Stage 4 studies will commence in 2019.

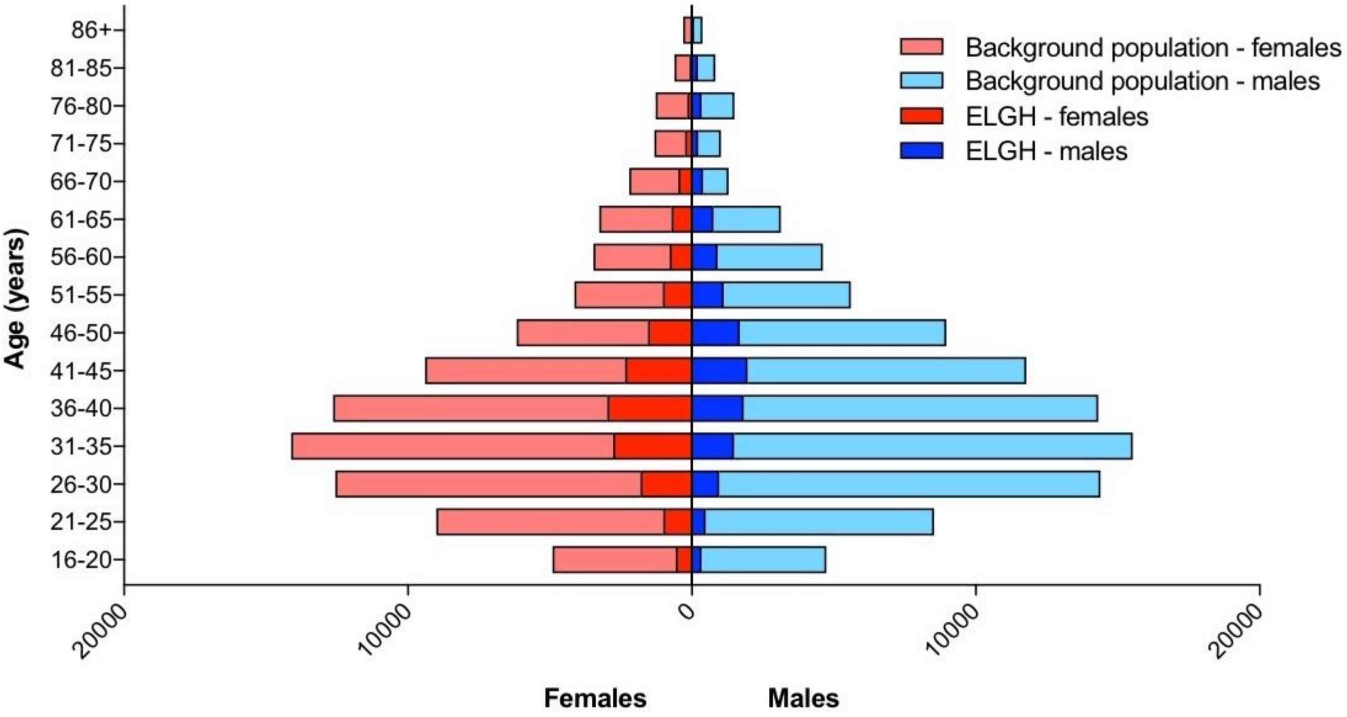
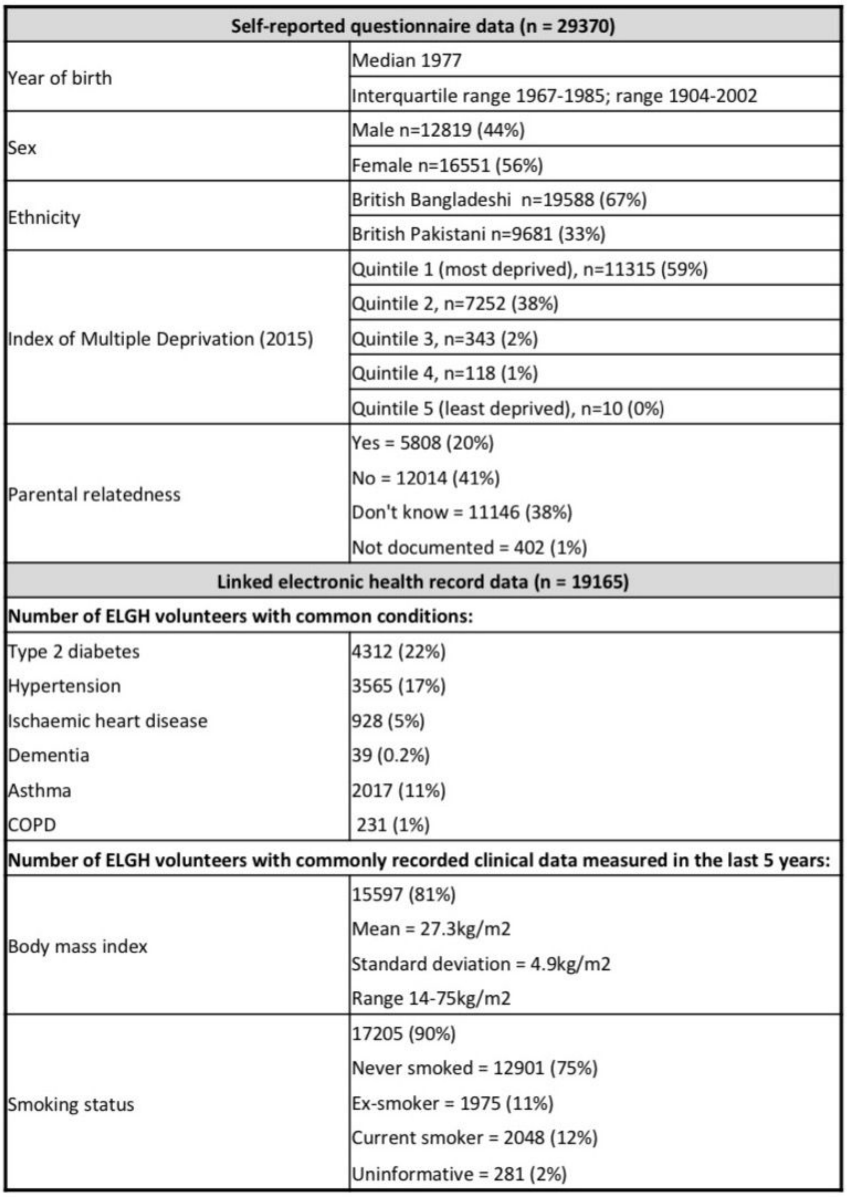


Figure 2: Population pyramid showing age and sex of ELGH volunteers (n=29,370) versus the total population of people of British-Bangladeshi and -Pakistani heritage (n=152,564) in East London (all NHS GP-registered adults residing in the London Boroughs of City and Hackney, Newham, Tower Hamlets, Waltham Forest), aged ≥ 16 years.

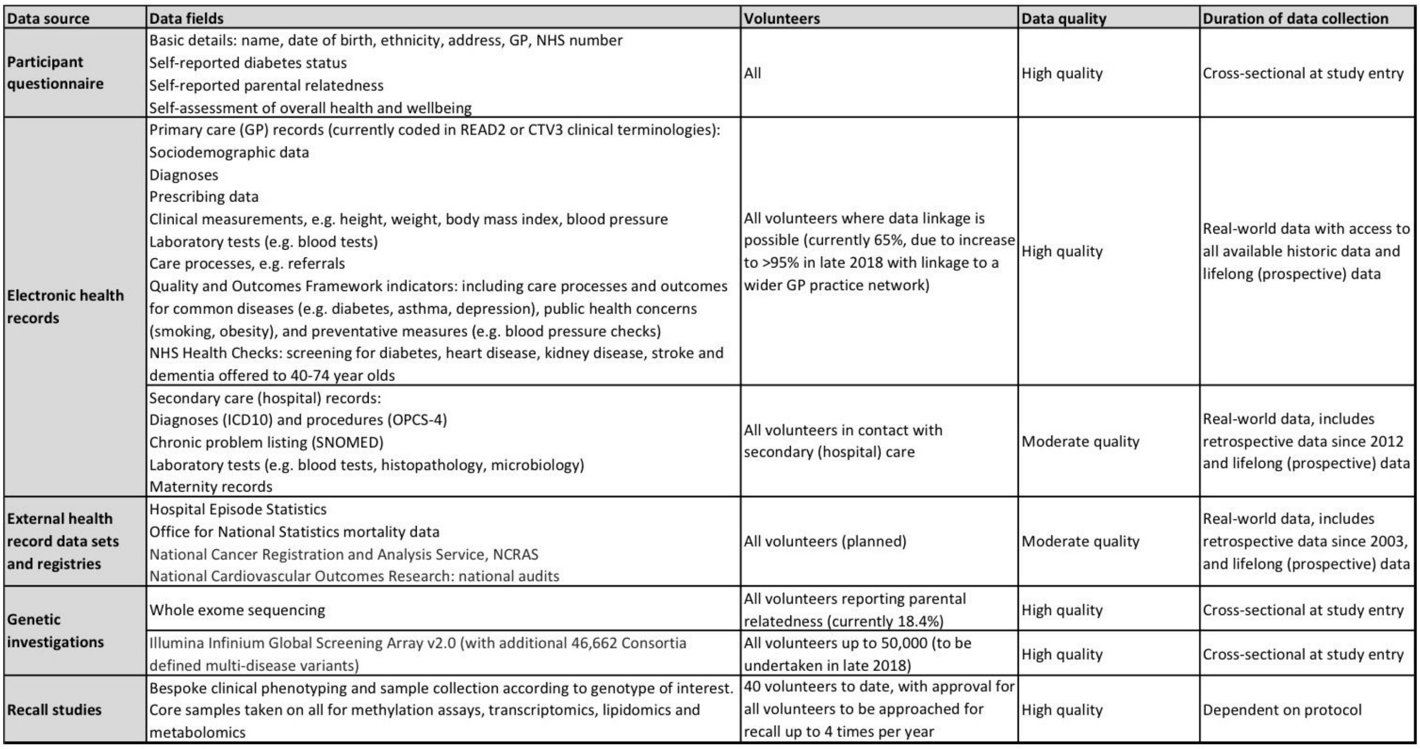


Self-reported questionnaire data (n = 29370)	
Year of birth	Median 1977
	Interquartile range 1967-1985; range 1904-2002
Sex	Male n=12819 (44%)
	Female n=16551 (56%)
Ethnicity	British Bangladeshi n=19588 (67%)
	British Pakistani n=9681 (33%)
Index of Multiple Deprivation (2015)	Quintile 1 (most deprived), n=11315 (59%)
	Quintile 2, n=7252 (38%)
	Quintile 3, n=343 (2%)
	Quintile 4, n=118 (1%)
	Quintile 5 (least deprived), n=10 (0%)

Parental relatedness	Yes = 5808 (20%)
	No = 12014 (41%)
	Don't know = 11146 (38%)
	Not documented = 402 (1%)

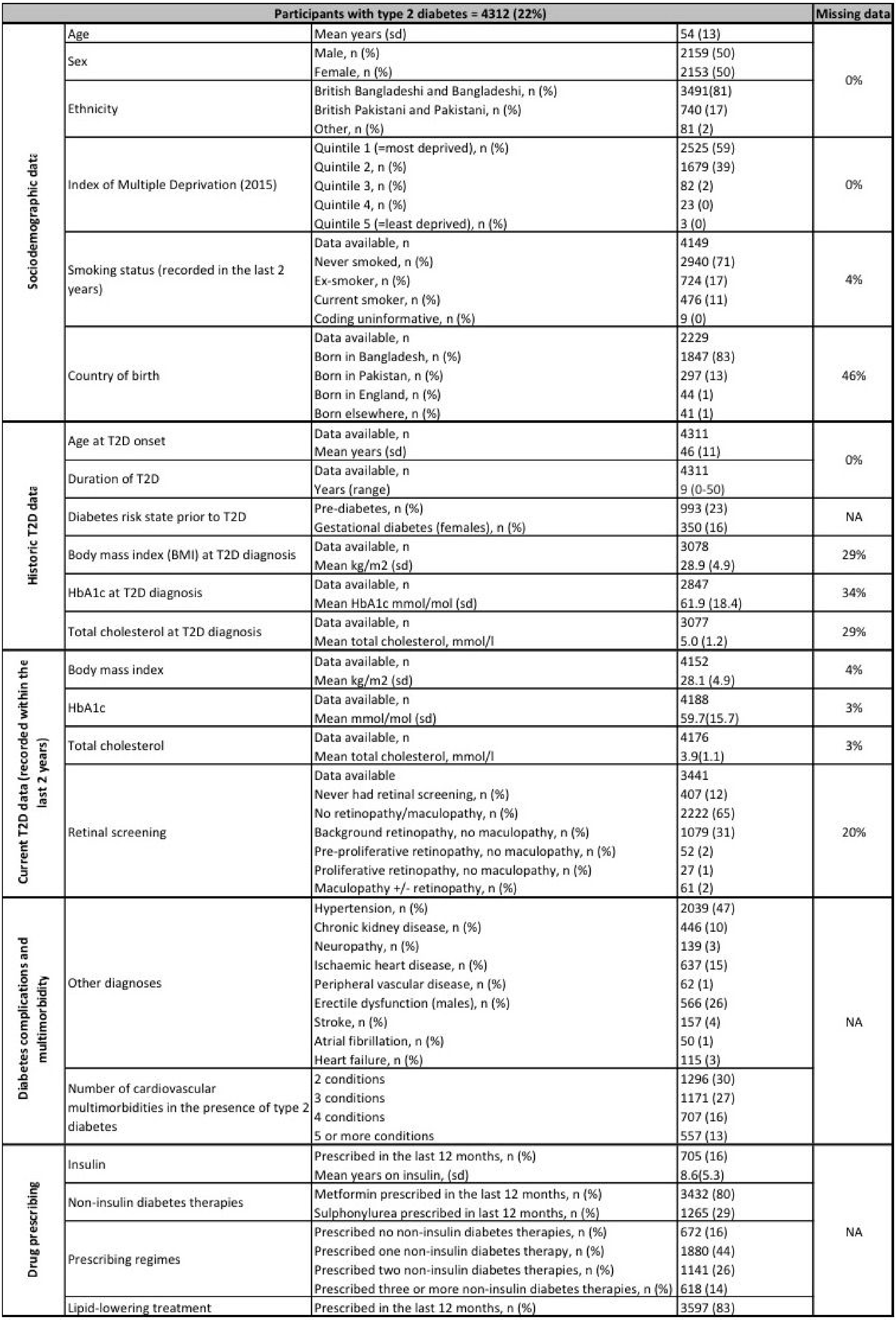
Linked electronic health record data (n = 19165)	
Number of ELGH volunteers with common conditions:	
Type 2 diabetes	4312 (22%)
Hypertension	3565 (17%)
Ischaemic heart disease	928 (5%)
Dementia	39 (0.2%)
Asthma	2017 (11%)
COPD	231 (1%)
Number of ELGH volunteers with commonly recorded clinical data measured in the last 5 years:	
Body mass index	15597 (81%) Mean = 27.3kg/m2 Standard deviation = 4.9kg/m2 Range 14-75kg/m2
Smoking status	17205 (90%) Never smoked = 12901 (75%) Ex-smoker = 1975 (11%) Current smoker = 2048 (12%) Uninformative = 281 (2%)

Table 1. Baseline characteristics of ELGH volunteers from self-reported questionnaire and electronic health record data.



Data source	Data fields	Volunteers	Data quality	Duration of data collection
Participant questionnaire	Basic details: name, date of birth, ethnicity, address, GP, NHS number Self-reported diabetes status Self-reported parental relatedness Self-assessment of overall health and wellbeing	All	High quality	Cross-sectional at study entry
Electronic health records	Primary care (GP) records (currently coded in READ2 or CTV3 clinical terminologies): Sociodemographic data Diagnoses Prescribing data Clinical measurements, e.g. height, weight, body mass index, blood pressure Laboratory tests (e.g. blood tests) Care processes, e.g. referrals Quality and Outcomes Framework indicators: including care processes and outcomes for common diseases (e.g. diabetes, asthma, depression), public health concerns (smoking, obesity), and preventative measures (e.g. blood pressure checks) NHS Health Checks: screening for diabetes, heart disease, kidney disease, stroke and dementia offered to 40-74 year olds	All volunteers where data linkage is possible (currently 65%, due to increase to >95% in late 2018 with linkage to a wider GP practice network)	High quality	Real-world data with access to all available historic data and lifelong (prospective) data
	Secondary care (hospital) records: Diagnoses (ICD10) and procedures (OPCS-4) Chronic problem listing (SNOMED) Laboratory tests (e.g. blood tests, histopathology, microbiology) Maternity records	All volunteers in contact with secondary (hospital) care	Moderate quality	Real-world data, includes retrospective data since 2012 and lifelong (prospective) data
External health record data sets and registries	Hospital Episode Statistics Office for National Statistics mortality data National Cancer Registration and Analysis Service, NCRAS National Cardiovascular Outcomes Research: national audits	All volunteers (planned)	Moderate quality	Real-world data, includes retrospective data since 2003, and lifelong (prospective) data
Genetic investigations	Whole exome sequencing	All volunteers reporting parental relatedness (currently 18.4%)	High quality	Cross-sectional at study entry
	Illumina Infinium Global Screening Array v2.0 (with additional 46,662 Consortia defined multi-disease variants)	All volunteers up to 50,000 (to be undertaken in late 2018)	High quality	Cross-sectional at study entry
Recall studies	Bespoke clinical phenotyping and sample collection according to genotype of interest. Core samples taken on all for methylation assays, transcriptomics, lipidomics and metabolomics	40 volunteers to date, with approval for all volunteers to be approached for recall up to 4 times per year	High quality	Dependent on protocol

Table 2: Summary of all data types currently available in ELGH for Stage 1 volunteers, and planned for late 2018 onwards.



Participants with type 2 diabetes = 4312 (22%)				Missing data
Sociodemographic data	Age	Mean years (sd)	54 (13)	0%
	Sex	Male, n (%)	2159 (50)	
		Female, n (%)	2153 (50)	
	Ethnicity	British Bangladeshi and Bangladeshi, n (%)	3491(81)	
		British Pakistani and Pakistani, n (%)	740 (17)	
		Other, n (%)	81 (2)	
	Index of Multiple Deprivation (2015)	Quintile 1 (=most deprived), n (%)	2525 (59)	0%
		Quintile 2, n (%)	1679 (39)	
		Quintile 3, n (%)	82 (2)	
		Quintile 4, n (%)	23 (0)	
		Quintile 5 (=least deprived), n (%)	3 (0)	
	Smoking status (recorded in the last 2 years)	Data available, n	4149	4%
		Never smoked, n (%)	2940 (71)	
		Ex-smoker, n (%)	724 (17)	
		Current smoker, n (%)	476 (11)	
		Coding uninformative, n (%)	9 (0)	
	Country of birth	Data available, n	2229	46%
		Born in Bangladesh, n (%)	1847 (83)	
		Born in Pakistan, n (%)	297 (13)	
		Born in England, n (%)	44 (1)	
		Born elsewhere, n (%)	41 (1)	
Historic T2D data	Age at T2D onset	Data available, n	4311	0%
		Mean years (sd)	46 (11)	
	Duration of T2D	Data available, n	4311	
		Years (range)	9 (0-50)	
	Diabetes risk state prior to T2D	Pre-diabetes, n (%)	993 (23)	NA
		Gestational diabetes (females), n (%)	350 (16)	
	Body mass index (BMI) at T2D diagnosis	Data available, n	3078	29%
		Mean kg/m2 (sd)	28.9 (4.9)	
	HbA1c at T2D diagnosis	Data available, n	2847	34%
		Mean HbA1c mmol/mol (sd)	61.9 (18.4)	
	Total cholesterol at T2D diagnosis	Data available, n	3077	29%
		Mean total cholesterol, mmol/l	5.0 (1.2)	
Current T2D data (recorded within the last 2 years)	Body mass index	Data available, n	4152	4%
		Mean kg/m2 (sd)	28.1 (4.9)	
	HbA1c	Data available, n	4188	3%
		Mean mmol/mol (sd)	59.7(15.7)	
	Total cholesterol	Data available, n	4176	3%
		Mean total cholesterol, mmol/l	3.9(1.1)	
	Retinal screening	Data available	3441	20%
		Never had retinal screening, n (%)	407 (12)	
		No retinopathy/maculopathy, n (%)	2222 (65)	
		Background retinopathy, no maculopathy, n (%)	1079 (31)	
		Pre-proliferative retinopathy, no maculopathy, n (%)	52 (2)	
		Proliferative retinopathy, no maculopathy, n (%)	27 (1)	
		Maculopathy +/- retinopathy, n (%)	61 (2)	
Diabetes complications and multimorbidity	Other diagnoses	Hypertension, n (%)	2039 (47)	NA
		Chronic kidney disease, n (%)	446 (10)	
		Neuropathy, n (%)	139 (3)	
		Ischaemic heart disease, n (%)	637 (15)	
		Peripheral vascular disease, n (%)	62 (1)	
		Erectile dysfunction (males), n (%)	566 (26)	
		Stroke, n (%)	157 (4)	
		Atrial fibrillation, n (%)	50 (1)	
		Heart failure, n (%)	115 (3)	
	Number of cardiovascular multimorbidities in the presence of type 2 diabetes	2 conditions	1296 (30)	
		3 conditions	1171 (27)	
		4 conditions	707 (16)	
		5 or more conditions	557 (13)	
Drug prescribing	Insulin	Prescribed in the last 12 months, n (%)	705 (16)	NA
		Mean years on insulin, (sd)	8.6(5.3)	
	Non-insulin diabetes therapies	Metformin prescribed in the last 12 months, n (%)	3432 (80)	
		Sulphonylurea prescribed in last 12 months, n (%)	1265 (29)	
	Prescribing regimes	Prescribed no non-insulin diabetes therapies, n (%)	672 (16)	
		Prescribed one non-insulin diabetes therapy, n (%)	1880 (44)	
		Prescribed two non-insulin diabetes therapies, n (%)	1141 (26)	
		Prescribed three or more non-insulin diabetes therapies, n (%)	618 (14)	
	Lipid-lowering treatment	Prescribed in the last 12 months, n (%)	3597 (83)	

Table 3. Example of a specific disease phenotype: characteristics of ELGH volunteers with type 2 diabetes. Data are presented in summary and descriptive formats as indicated. Missing data is estimated where available, e.g. for clinical care processes and measurements, but not diagnostic coding where the absence of a code is taken to indicate the absence of a diagnosis.

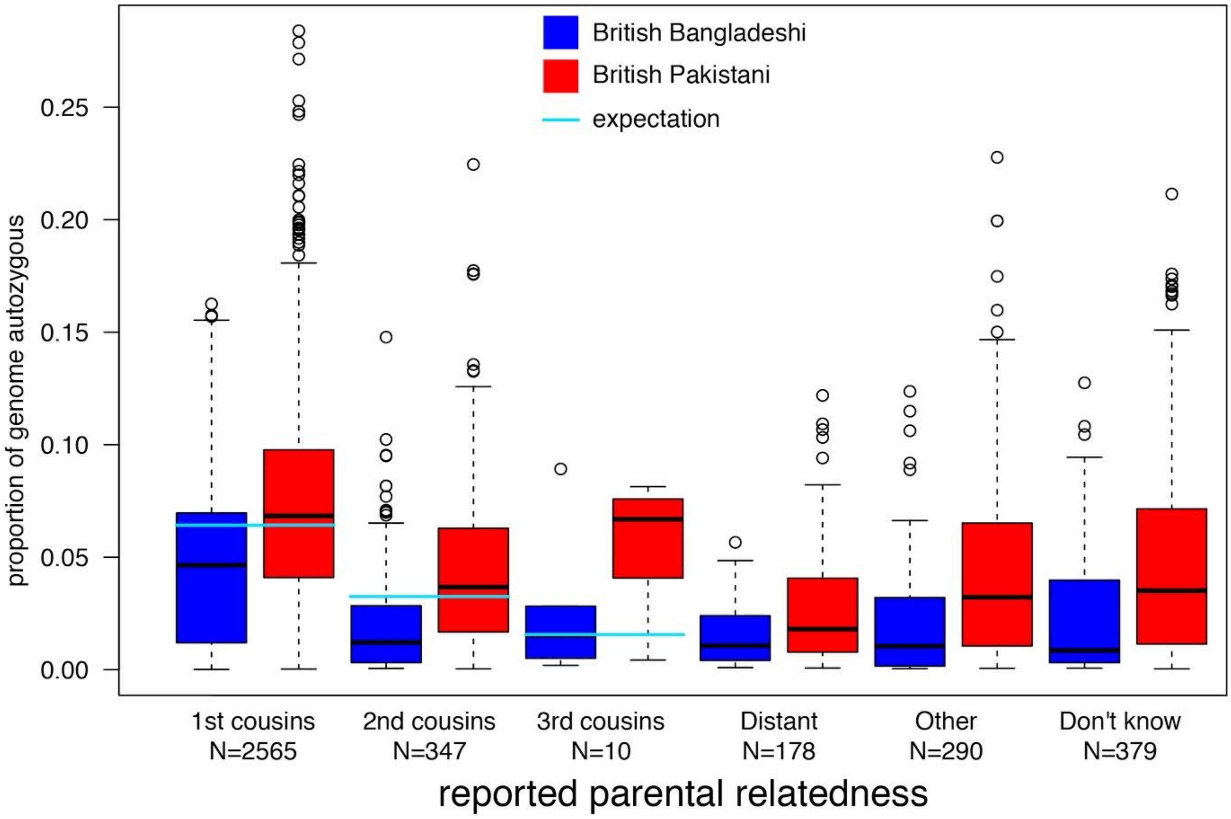


Figure 3. Distribution of levels of autozygosity as a fraction of the genome in British Bangladeshi and British Pakistani volunteers, split according to self-reported parental relatedness (Tukey box plot showing median, lower and upper quartiles, quartiles +/- 1.5x interquartile range, and outliers).

Supplementary files

Supplementary file 1: Stage 1 participant questionnaire (.pdf file)

https://docs.google.com/document/d/1PIPoXMvCeJqcZ-po-h360THBcHeOtxShE-X93EGTVoA/edit?usp=sharing

Supplementary file 2: Example primary care electronic health record data extraction template for type 2 diabetes (.xls file)

https://docs.google.com/document/d/15BjI8EazcVqRM4onDZKr2rScuGZ2LWoynzBkeXns7qA/edit?usp=sharing

References

1. Popejoy AB, Fullerton SM. Genomics is failing on diversity. *Nature*. 2016 Oct 13;**538**(7624):161–164.

2. 2011 Census: Ethnic group, local authorities in the United Kingdom [Internet]. *Office for National Statistics* 2013 [cited 2015 Nov]. Available from: www.ons.gov.uk/ons/rel/census/key-statistics-for-local-authorities-inengland-

3. Barnett AH, Dixon AN, Bellary S, et al. Type 2 diabetes and cardiovascular risk in the UK south Asian community. *Diabetologia*. 2006 Oct;**49**(10):2234–2246.

4. Sattar N, Gill JMR. Type 2 diabetes in migrant south Asians: mechanisms, mitigation, and management. *Lancet Diabetes Endocrinol*. 2015 Dec;**3**(12):1004–1016.

5. Goodyear M. Public Health Profile of north east London for NE London Sustainability and Transformation Plan [Internet]. 2016 [cited 2018 Sep 6]. Available from: http://archive.eastlondonhcp.nhs.uk/wp-content/uploads/2017/06/NEL-STP-JSNA-2016.pdf

6. Mathur R, Noble D, Smith D, Greenhalgh T, Robson J. Quantifying the risk of type 2 diabetes in East London using the QDScore: a cross-sectional analysis. *Br J Gen Pract*. 2012 Oct;**62**(603):e663–70.

7. Dreyer G, Hull S, Mathur R, Chesser A, Yaqoob MM. Progression of chronic kidney disease in a multi-ethnic community cohort of patients with diabetes mellitus. *Diabet Med*. 2013 Aug;**30**(8):956–963.

8. Alazawi W, Mathur R, Abeysekera K, et al. Ethnicity and the diagnosis gap in liver disease: a population-based study. *Br J Gen Pract*. 2014 Nov;**64**(628):e694–702.

9. George J, Mathur R, Shah AD, et al. Ethnicity and the first diagnosis of a wide range of cardiovascular diseases: Associations in a linked electronic health record cohort of 1 million patients. *PLoS One*. 2017 Jun 9;**12**(6):e0178945.

10. Hull S, Chowdhury TA, Mathur R, Robson J. Improving outcomes for patients with type 2 diabetes using general practice networks: a quality improvement project in east London. *BMJ Qual Saf*. 2014 Feb;**23**(2):171–176.

11. Robson J, Dostal I, Madurasinghe V, et al. NHS Health Check comorbidity and management: an observational matched study in primary care. *Br J Gen Pract*. 2017 Feb;**67**(655):e86–e93.

12. Narasimhan VM, Hunt KA, Mason D, et al. Health and population effects of rare gene knockouts in adult humans with related parents. *Science*. 2016 Apr 22;**352**(6284):474–477.

13. Joshi PK, Esko T, Mattsson H, et al. Directional dominance on stature and cognition in diverse human populations. *Nature*. 2015 Jul 23;**523**(7561):459–462.

14. Centre of The Cell [Internet]. *Centre of The Cell*. Available from: https://www.centreofthecell.org/

15. Flora Ogilvie GM. Health Equity in Primary Care in East London and the City: Data analysis to inform Joint Strategic Needs Assessment [Internet]. 2015 [cited 2018 Sep 6]. Available from: https://www.towerhamlets.gov.uk/Documents/Public-Health/TH-JSNA-Health-Equity-in-Primary-Care-2014.pdf

16. Farmer R, Mathur R, Bhaskaran K, Eastwood SV, Chaturvedi N, Smeeth L. Promises and pitfalls of electronic health record analysis. *Diabetologia*. 2018 Jun;**61**(6):1241–1248.

17. Hospital Episode Statistics [Internet]. *NHS Digital*. Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics

18. Linked hes ons mortality data [Internet]. *NHS Digital*. Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/linked-hes-ons-mortality-data

19. GSA Array Datasheet [Internet]. *Illumina*. Available from: http://emea.support.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/infinium-commercial-gsa-data-sheet-370-2016-016.pdf

20. Narasimhan VM, Rahbari R, Scally A, et al. Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes. *Nat Commun*. 2017 Aug 21;**8**(1):303.

21. Saleheen D, Natarajan P, Armean IM, et al. Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature*. 2017 Apr 12;**544**(7649):235–239.

22. Corbin LJ, Tan VY, Hughes DA, et al. Formalising recall by genotype as an efficient approach to detailed phenotyping and causal inference. *Nat Commun*. 2018 Feb 19;**9**(1):711.

23. Prados-Torres A, Poblador-Plou B, Gimeno-Miguel A, et al. Cohort Profile: The Epidemiology of Chronic Diseases and Multimorbidity. The EpiChron Cohort Study. *Int J Epidemiol*. 2018 Apr 1;**47**(2):382–384f.

24. European Genome-phenome Archive [Internet]. [cited 2018 Sep 18]. Available from: https://www.ebi.ac.uk/ega

25. Jones KH, Ford DV, Ellwood-Thompson S, Lyons RA. The UK Secure eResearch Platform for public health research: a case study. *Lancet*. 2016 Nov 1;**388**:S62.

26. Ford DV, Jones KH, Verplancke J-P, et al. The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC Health Serv Res*. 2009 Sep 4;**9**:157.