scientific reports

OPEN

Investigating misclassification of type 1 diabetes in a population-based cohort of British Pakistanis and Bangladeshis using polygenic risk scores

Timing Liu[1,19], Alagu Sankareswaran[2,3,19], Gordon Paterson[4,6,19], Genes & Health Research Team*, Diane P. Fraser[5], Sam Hodgson[4], Qin Qin Huang[1], Teng Hiang Heng[1], Meera Ladwa[4,7], Nick Thomas[5], David A. van Heel[7], Michael N. Weedon[5], Chittaranjan S. Yajnik[8], Richard A. Oram[5], Giriraj R. Chandak[2,3,20], Hilary C. Martin[1,20]✉ & Sarah Finer[4,6,20]✉

Correct classification of type 1 (T1D) and type 2 diabetes (T2D) is challenging due to overlapping clinical features and the increasingly early onset of T2D, particularly in South Asians. Polygenic risk scores (PRSs) for T1D and T2D have been shown to work relatively well in South Asians, despite being derived from largely European-ancestry samples. Here we used PRSs to investigate the rate of potential misclassification of diabetes amongst British Bangladeshis and Pakistanis. Using linked health records from the Genes & Health cohort (n = 38,344) we defined two reference groups meeting stringent diagnostic criteria: 31 T1D cases, 1842 T2D cases, and after excluding these, two further groups: 839 insulin-treated diabetic individuals with ambiguous features and 5174 non-diabetic controls. Combining these with 307 confirmed T1D cases and 307 controls from India, we calculated ancestry-corrected PRSs for T1D and T2D, with which we estimated the proportion of T1D cases within the ambiguous group at ~ 6%, dropping to ~ 4.5% within the subset who had T2D codes in their health records (and are thus most likely to have been misclassified). We saw no significant association between the T1D or T2D PRS and BMI at diagnosis, time to insulin, or the presence of T1D or T2D diagnostic codes amongst the T2D or ambiguous cases, suggesting that these clinical features are not particularly helpful for aiding diagnosis in ambiguous cases. Our results emphasise that robust identification of T1D cases and appropriate clinical care may require routine measurement of diabetes autoantibodies and C-peptide.

Type 1 and type 2 diabetes (T1D and T2D respectively) are classified aetiologically, and the two conditions differ in their natural history and in the treatment required[1,2]. For clinicians, correctly classifying T1D and T2D may be challenging due to overlapping clinical features as well as the increasingly early onset of T2D due to rising obesity prevalence. Additionally, a lack of recognition of late-onset T1D cases may lead to their incorrect classification as T2D[3]. Diabetes misclassification occurs when a clinician-recorded diagnosis of T1D or T2D does not match the true (aetiological) category to which the patient belongs[4], but it can be difficult to distinguish from erroneous miscoding in health records. Previous estimates of miscoding and misclassification suggest that up to 13% of diabetes cases recorded in British general practices are affected, but these studies do not establish a ground truth and rely on a complex algorithmic definition of type 1 and type 2 diabetes using clinical features

[1]Wellcome Trust Sanger Institute, Saffron Walden, UK. [2]Genomic Research on Complex diseases Group (GRC-Group), CSIR-Centre for Cellular and Molecular Biology, Hyderabad, India. [3]Academy of Scientific and Innovative Research, Ghaziabad, India. [4]Wolfson Institute of Population Health, Queen Mary University of London, London, UK. [5]University of Exeter, Exeter, UK. [6]Barts Health NHS Trust, London, UK. [7]Blizard Institute, Queen Mary University of London, London, UK. [8]Diabetes Unit, King Edward Memorial Hospital and Research Centre, Pune, India. [19]These authors contributed equally: Timing Liu, Alagu Sankareswaran, Gordon Paterson. [20]These authors jointly supervised this work: Giriraj R. Chandak, Hilary C. Martin, Sarah Finer. *A list of authors and their affiliations appears at the end of the paper. ✉email: hcm@sanger.ac.uk; s.finer@qmul.ac.uk

and the use of oral versus insulin treatment[4–7]. The correct classification of a patient as having type 1 rather than type 2 diabetes is suggested by rapid requirement to insulin (within 3 years of diagnosis)[8]. Without specialised tests to estimate endogenous insulin secretion (serum C-peptide measurement) and/or to determine if diabetes-specific autoimmunity is present (measurements of diabetes autoantibodies), diabetes type can be challenging to determine. However, the measurement of C-peptide and diabetes autoantibody testing is discouraged in routine clinical practice in the UK[9]. The presence of multiple different diagnostic codes relating to diabetes in an individual's electronic health record may reflect either erroneous miscoding or true misclassification, making it difficult to correctly classify diabetes in real-world electronic health record datasets[10].

Misclassification rates of diabetes in people of South Asian descent are not known[11,12] but are expected to be greater than in White Europeans, due to high prevalence of T2D and its tendency to present in people who are slim, young and in some cases have features of insulin deficiency[13,14]. These epidemiological factors may result in a possible bias towards T2D diagnosis even when clinical features would be commonly considered supportive of a T1D diagnosis[8,15]. Inaccurate classification of a true T1D case as T2D could result in severe harms including diabetic ketoacidosis[16,17] and increased risk of long-term diabetic complications[16]. Misclassification of a true T2D case and subsequent prescription of insulin may lead to harms including increased risk of hypoglycaemic events[18], and to the misclassified patient missing out on oral therapies that are highly effective at reducing the risk of T2D complications.

Large-scale genome-wide association studies (GWAS) have derived major insights into the genetic aetiology of complex diseases such as T1D and T2D[19–21] and have led to the construction of polygenic risk scores (PRSs) that could have clinical benefit through use in risk prediction and characterisation of disease heterogeneity[3,22–25]. Recent work has shown that PRS for T1D and T2D are useful tools to aid the classification of people with diabetes, in combination with clinical features and metabolic measurements[14,23,26]. Although, as is the case for most diseases, the T1D GWASs have largely been focused on Europeans, a T1D PRS derived from these shows good discriminative ability in Indians[27]. Large multi-ancestry GWASs for T2D have recently been performed[19,28], and PRSs derived from these were found to work almost equally well in South Asians as in Europeans[19].

We set out to use these PRSs to investigate rates of misclassification of diabetes in a population of British South Asians. Specifically, we applied genetic ancestry-optimised T1D and T2D PRSs to estimate the proportion and misclassification of T1D amongst insulin-treated diabetic individuals with ambiguous clinical features. We emphasise that our goal was *not* to use the PRSs to predict which specific individuals had T1D versus T2D. Secondly, we tested whether T1D or T2D PRSs were associated with the clinical characteristics commonly used to determine diabetes type clinically. We used the Genes and Health (G&H) cohort based in East London, UK, which combines genomic and detailed electronic health record data for over 44,000 people of British Bangladeshi and Pakistani descent[29], who have twice the rate of T2D than the local White-European population[30]. The combination of high-quality phenotypic and genetic data gives us a unique opportunity to study misclassification of diabetes in this understudied South Asian-ancestry population.

Methods

G&H study population and clinical codes

The Genes & Health cohort has been described previously[22,29,31]. Briefly, G&H is a community-based study which recruits British Pakistani and Bangladeshi individuals aged 16 years and older from National Health Service (NHS) and community settings in the UK. All volunteers consent to lifelong electronic health record (EHR) access and donate a saliva sample for genetic studies. G&H was approved by the London South East NRES Committee of the Health Research Authority (14/LO/1240).

For this study, we used the June 2021 data release, selecting 38,344 of the 46,132 volunteers recruited in east London (median age 43 years; 54.4% female; 45.6% male) who had primary health care record data available. Additional secondary care data was available for 22,713 of the 38,244 individuals who had any interactions with Barts Health NHS Trust, the largest secondary care trust in the area providing inpatient and outpatient services, including specialist diabetes care. Individuals were excluded (N = 94) if they had registered at different GP practices with an incongruent year of birth. Definitions of clinical diagnosis can be found in Supplementary "Methods", with key codes given in Supplementary Tables 1 and 2. Quantitative clinical measures were cleaned as described in Supplementary "Methods".

Defining ambiguous cases, reference cases and controls in G&H

We filtered the available data to define four groups of mutually exclusive individuals, as follows. If electronic health records did not contain sufficient historic information to inform the T1D reference case selection (e.g. unavailability of historic prescribing information), additional case note review was undertaken by two experienced diabetologists (SF, ML).

1) T1D reference cases:

- Individuals with a clinical code for T1D or T2D (Supplementary Table 1). (In practice, as Table 1 shows, most of the thirty-one individuals ultimately included in this group did have a T1D code, but we also included seven who had only a T2D code because, based on review from two senior diabetologists (SF and ML), we believe the following criteria makes it highly likely they are truly T1D and this is a miscoding error.)
- AND had a time to insulin from diagnosis between 0.5 and 1 years, with the rationale that this would capture people who were beginning regular insulin therapy (People diagnosed with T2D who are initially not insulin treated but progress rapidly to insulin therapy are often actually misclassified T1D patients[12].)

		Median value (quartiles)				
Characteristic	**N. with data**	**Ambiguous N = 839**	**T1D N = 31**	**T2D N = 1842**	**Controls N = 5174**	***p* value (ANOVA)**
Age at diagnosis (years)	2712	40 (33, 46)$^{*\#}$	23 (13, 34)$^{\$\#}$	45 (39, 53)$^{\$*}$	NA	< 0.001
BMI (kg/m^2)	5928	27.8 (25.1, 31.8)	26.4 (20.4, 31.8)	28.4 (25.6,32.1)	26.3 (22.8, 29.0)	0.08
HDL (mmol/l)	4324	1.09 (0.90,1.23)	1.20 (1.10,1.43)	1.04 (0.90,1.21)	1.27 (1.05,1.44)	0.19
Triglycerides (mmol/l)	4134	2.00 (1.37, 2.81)	1.65 (1.05,2.54)	2.00 (1.40,2.98)	1.46 (0.85,1.75)	0.39
HbA1c (mmol/mol)	4814	67 (55,87)$^{\#}$	78 (57, 111)$^{\#}$	59 (51,76)$^{\$*}$	36 (33, 38)	< 0.001
C-peptide (pmol/L)	57	905 (564,1228)*	162 (21,168)$^{\$\#}$	1056 (880,1179)*	NA	< 0.001
Time to insulin (years)	864	10 (5,15)	5 (1,12)	NA	NA	0.025
		Number of individuals (%)				
Characteristic	**N. with data**	**Ambiguous N = 839**	**T1D N = 31**	**T2D N = 1842**	**Controls N = 5174**	***p* value (ANOVA)**
Gender	7868					< 0.001
Female		450 (54%)$^{\#}$	12 (39%)	749 (41%)$^{\$}$	2,962 (57%)$^{@}$	
Male		389 (46%)$^{\#}$	19 (61%)	1,093 (59%)$^{\$}$	2,194 (42%)$^{@}$	
Autoimmune condition present (n,%)	7886	43 (5.1%)$^{\#}$	3 (9.7%)	53 (2.9%)$^{\$}$	117 (2.2%)	0.003
Only T1D Code (n,%)	7886	18 (2%)*	15 (48%)$^{\$}$	0 (0%)	0 (0%)	< 0.001
Only T2D Code (n,%)	7886	18 (2%)*	7 (23%)$^{\$\#}$	1,822 (99%)$^{\$*}$	0 (0%)	< 0.001
T1D & T2D Code (n,%)	7886	105 (13%)$^{*\#}$	9 (29%)$^{\$\#}$	20 (1%)$^{\$*}$	0 (0%)	< 0.001
Diabetes Auto-antibodies tested (n)	69	34 (4%)	15 (48%)	20 (1%)	0 (0%)	< 0.001
Positive (n,%)		9 (26%)$^{*\#}$	11 (73%)$^{\$\#}$	0 (0%)$^{\$*}$	0 (0%)	
Negative (n,%)		25 (74%)	4 (27%)	20 (100%)	0(0%)	

Table 1. Clinical characteristics recorded within one year of diagnosis of individuals in the ambiguous, T1D and T2D groups in G&H, and in the G&H non-diabetic controls. In the top part of the table, quantitative traits are presented as the median (quartiles), and in the bottom part, count variables are presented as sample size (percentage). Significance was assessed using Kruskal–Wallis tests (non-parametric one-way ANOVA). Dunn's and Chi-squared post-hoc tests were performed where appropriate (Supplementary Table 5). *significant ($p < 0.05$) versus T1D group; $^{\#}$ significant ($p < 0.05$) versus T2D group; $^{\$}$significant ($p < 0.05$) versus ambiguous group. $^{@}$ Eighteen controls for whom both male and female genders were recorded in the health records were excluded from clinical comparisons.

- AND (a) for those aged under 30 at diagnosis either insulin deficiency OR positive diabetes autoantibodies were required; (b) for those aged 30 to 60 years at diagnosis ONLY evidence of insulin deficiency was used, based on autoantibody tests having a low positive predictive value in this age group[32]. No individuals who were aged over 60 at diagnosis passed the above criteria.

2) T2D reference cases:

- Individuals with a clinical code for T2D
- AND duration of diabetes > 3 years
- AND had received oral anti-hyperglycaemic treatment but never insulin
- AND no confirmed insulin deficiency or autoantibody positivity

3) Ambiguous group:

- Individuals with a clinical code for T1D and/or T2D
- AND did not meet the above criteria for T1D or T2D reference cases
- AND had age at diagnosis 60 years old or younger
- AND had received an insulin prescription within the most recent year of data linkage

4) Non-diabetic controls:

- Individuals with no diagnostic code ever recorded for: T1D, T2D, secondary/rare diabetes and pancreatic disease/surgery, diabetes risk states (defined in Supplementary Table 3)
- AND has no confirmed insulin deficiency or autoantibody positivity
- AND has never been prescribed diabetes medications (any)

The number of individuals remaining after these filtering steps are shown in Supplementary Table 4.

Our stringent criteria to define confirmed T1D resulted in a group of only 31 individuals (Table 1), and therefore a separate T1D reference population was used from an Indian cohort, the details of which are presented below.

Statistical analysis of clinical data

For comparisons between baseline groups, Kruskall-Wallis tests were performed with Dunn's post-hoc where appropriate. For comparisons of two groups, student's t-tests were used to compare continuous measures and chi-squared tests were used to compare binary end-points.

Preparation of G&H genotype data

We used the June 2021 data freeze which included 46,132 individuals genotyped on the Illumina Infinium Global Screening Array v3 chip (GRCh38). Quality control, inference of genetic ancestry and inference of relatedness are described in the Supplementary "Methods". After quality control, restriction to individuals fulfilling the clinical criteria described above, and removal of relatives, we retained 7886 unrelated individuals with genetically-inferred Pakistani or Bangladeshi ancestry. These comprised 31 T1D cases, 1842 T2D cases, 839 ambiguous cases and 5174 non-diabetic controls. We combined genetic data from these individuals with the Indian cohort described below.

The Indian cohort of T1D cases and controls

The Indian cohort consisted of 332 T1D cases and 317 non-diabetic controls who were recruited in Pune, India via detailed phenotypic characterisation and robust diagnostic classification, using the same methods as those described in Harrison et al.[27]. The Institutional Ethics Committee of the KEM Hospital Research Centre, Pune, India (KEMHRC ID No1737 & KEMHRC ID No PhD19) approved the study, and all methods were performed in accordance with the relevant guidelines and regulations. Individuals were genotyped on the Illumina GSA-24v3 chip at the CSIR-Centre for Cellular and Molecular Biology, Hyderabad, India (CSIR-CCMB). Quality control of the genotype data is described in the Supplementary "Methods". After removing principal component outliers and third-degree relatives or closer using PropIBD metric from KING, 307 cases and 307 controls remained in the dataset.

Principal component analysis and imputation of G&H and Indian samples

We combined the genetic data from the 7886 unrelated G&H individuals and the 614 unrelated Indian individuals to perform a principal component analysis (PCA) which we subsequently used to correct the T1D and T2D polygenic risk scores (PRSs) for genetic ancestry differences (described below). The preparation of this combined dataset, the PCA, and imputation to TOPMED r2 are described in Supplementary "Methods".

PRS calculation and PC correction

For type 1 diabetes, we used a previously-published PRS which has been shown to have good discriminative ability in Indians (area under the receiver operating characteristic curve = 0.84)[27,33]. Specifically, we used ten SNPs including two that, in combination, tag HLA-DR3/DR4-DQ8 haplotype, since Oram et al. showed that this 10-SNP score performed almost as well as the larger 30-SNP score[33].

For type 2 diabetes, we used a PRS derived from the largest recent trans-ancestry meta-analysis, which was shown to perform almost as well in South Asians as in Europeans, with pseudo-r^2 of ~ 3–6% depending on the target cohort[19]. The PRS was calculated using PRSice2.2 using SNPs that had $P < 1 \times 10^{-4}$ in that GWAS meta-analysis, as we found that the area under the receiver operating curve was highest at this P value cut-off. Clumping R^2 was set to 0.1, and European samples from 1000 Genomes Project were used as the LD reference since the majority of samples in the GWAS meta-analysis with European-ancestry.

We observed significant differences in the PRS distributions between the genetically-inferred ancestry groups even when restricting to non-diabetic controls (Supplementary Fig. 1), which are well known to exist due to differences in demographic history[34]. We thus decided to correct these by regressing out the principal components (PCs). A scree plot (Supplementary Fig. 2) suggested that five PCs were sufficient to explain most of the variation (Supplementary Fig. 3). Since ancestry was correlated with case status within our cohort (i.e. the majority of T1D cases were Indian and the T2D cases were Pakistani/Bangladeshi), we regressed the PRSs on these five PCs in the controls only for both the T1D and T2D PRSs:

$$PRS_{controlc} \sim \sum_{p=1}^{5} \beta_p PC_{p,c} + \epsilon$$

We then used the estimated β s for each PC to calculate the expected PRS for each case and control individual i as follows:

$$\widehat{PRS_i} = \sum_{p=1}^{5} \widehat{\beta}_p PC_{p,i}$$

We then calculated the residuals which we used in subsequent analyses:

$$\mathrm{PC-corrected}\ PRSi = PRS_i - \widehat{PRS_i}$$

Statistical analysis of PRSs

We employed the approach from Evans et al.[23] to estimate the prevalence of T1D in the set of ambiguous cases from G&H (f_{T1D}). We did this using three approaches. In all cases we used the PC-corrected PRSs constructed as described above. In the first approach, we used the small sample of G&H T1D cases as "true cases" and the G&H T2D cases as "non-cases" to estimate f_{T1D} using the PC-corrected T1D PRS. Since we had only a small sample of clear T1D cases in G&H, in the second approach we used the larger sample of Indian T1D cases as our "true cases" and the G&H T2D cases as "non-cases", and used the PC-corrected T1D PRSs. In the third approach, we used the T2D cases and non-diabetic controls from G&H to estimate the fraction of the ambiguous cases that had T2D (f_{T2D}) using the PC-corrected T2D PRS, and then calculated f_{T1D} as 1-f_{T2D}. With each approach, we used the three statistical methods from[23]: the means method, the Earth Mover's Distance (EMD), and the Kernel Density Estimation (KDE) method.

Association between PRS and clinical parameters

We selected three clinical parameters routinely used by clinicians and researchers to classify diabetes diagnosis: age at diagnosis, BMI at diagnosis and time to insulin start[3,35]. Age at onset[36] and BMI[37] have been associated with the T2D PRS amongst European T2D cases. We used multiple linear regression to assess the association between the T1D or T2D PRS and these clinical parameters within either the ambiguous cases or T2D cases (Fig. 3). Analysis was performed using R 4.0.1 and Python 3.8.

Results

Definition and clinical characteristics of T1D, T2D and ambiguous cases in G&H

Clinical data were available for a total of 38,344 individuals. Supplementary Table 4 indicates the results of our filtering process to define T1D, T2D, ambiguous cases and nondiabetic controls, who were subsequently filtered to remove related individuals. Using this process, we were able to identify 31 T1D reference cases, 1842 T2D reference cases, 839 further diabetes cases treated with insulin and deemed ambiguous by the criteria outlined above, and 5174 nondiabetic controls. The majority of the individuals we removed were excluded due to having diabetes risk states (e.g. 'at risk of diabetes mellitus' or 'family history of diabetes mellitus') in their health records, or due to being putative T2D cases not on oral antihyperglycemic agents.

Using ANOVA, we compared clinical features between the four groups (Table 1) and found no significant difference in BMI at diagnosis. There were significant differences across groups in age and glycosylated haemoglobin (HbA1c) at diagnosis, C-peptide (ANOVA $P<0.001$) and diabetes autoantibody positivity (ANOVA $P<0.01$), consistent with clinical expectation. However, C-peptide and diabetes autoantibodies were rarely measured (0.7% and 2.5% of diabetes cases, respectively). The ambiguous group displayed intermediate clinical features, with post-hoc pairwise comparisons (Supplementary Table 5) demonstrating differences with both T1D and T2D reference groups in age at diagnosis (40 years for the ambiguous group versus 23 years for T1D cases and 45 years for T2D cases), and C-peptide significantly higher than the T1D but not T2D reference cases (905 pmol/l for the ambiguous versus 162 pmol/l for T1D versus 1056 pmol/l for T2D groups, respectively). In contrast, HbA1c at diagnosis was significantly higher in the ambiguous group (67 mmol/mol) than T2D (59 mmol/mol) but not T1D (78 mmol/mol) cases. Importantly, clinical codes did not differentiate these groups reliably, with 39% of T1D cases having a T2D code present in their electronic health records, and only 77% actually having a T1D code. The ambiguous group was predominantly coded as having T2D: 93% had only a T2D code, but 7% had a T1D code present either with or without a T2D code.

Estimating the fraction of T1D cases using PRSs

Despite being derived from GWASs in purely or mostly European samples, PRSs for T1D and T2D have been shown to have reasonably good discriminative ability in South Asians[19,27]. We set out to use these PRSs to estimate the proportion of T1D cases within the set of 839 ambiguous cases. We applied a mathematical framework recently proposed by Evans et al. for estimating the prevalence of a given disease within a cohort using PRSs[23]. This relies on estimating the proportion of true cases and non-cases from the PRS distribution of a sample of individuals that contains a mixture of these. In this instance, we wished to estimate the proportion of T1D cases amongst the set of ambiguous cases, which we presume contains a mixture of T1D and T2D cases. However, since we only had 31 definite T1D cases from G&H, we anticipated this would be insufficient to produce an accurate estimate of the prevalence of T1D within the ambiguous group. We thus combined our data with a larger set of 307 cases and 307 controls from Pune, India, and regressed out genetic principal components from the T1D and T2D PRSs to correct for ancestry (Supplementary Fig. 3). PC-corrected PRSs did not show significant differences between the different ancestry groups amongst the controls (Supplementary Figs. 1b and 4b), but they showed significant differences between cases of the relevant diabetes type and non-diabetic controls within G&H (Fig. 2). We then applied three approaches to estimate the prevalence of T1D within the ambiguous group using the PC-corrected PRSs, and for each, used three different statistical methods to estimate the mixture proportion (see Methods).

We found that the three approaches produced very similar estimates of the fraction of T1D cases in our ambiguous group, with the point estimates ranging from 3.6% to 10.2% (median 5.9%) and confidence intervals from 0 to 15.2% (Fig. 1a). Estimates were also very similar between three different statistical approaches used to estimate the mixture proportion, with confidence intervals all overlapping. We then performed the same analysis removing individuals (n = 18) from the ambiguous group who only had a T1D code present in their clinical records, considering that they were more likely to be true T1D diagnoses. This therefore gave an estimate of the proportion of individuals who are likely to be misclassified. The point estimates of prevalence (and presumed misclassification estimates) ranged from 1.9 to 7.8%, with a median of 4.5%, and confidence intervals from 0 to 12.6% (Fig. 1b).

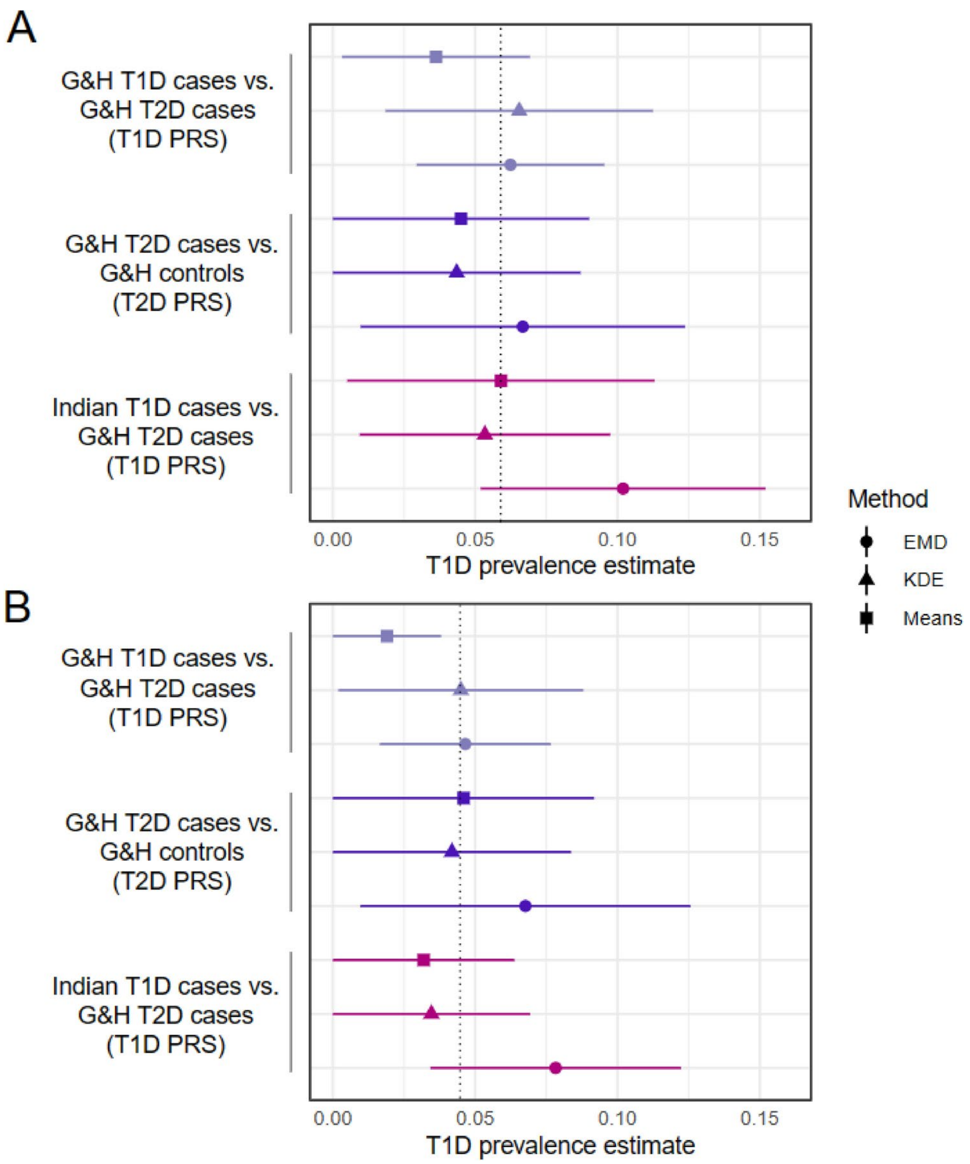

Fig. 1. Estimated prevalence of T1D in the G&H ambiguous group. Panel (**A**) includes all individuals in the ambiguous group (n = 839) and (**B**) excludes those individuals where only a T1D clinical code is present (n = 821). Points show mean estimates and horizontal lines indicate 95% confidence intervals. The dotted line indicates the median point estimate across all methods. The point type indicates the statistical method used for estimation. EMD: Earth Mover's Distance; KDE: Kernel density estimation.

Associations between PRSs and clinical characteristics in G&H

Figure 2 shows the average T1D and T2D PRSs in the four clinically-defined groups in G&H. The ambiguous group had a similar (two-tailed Wilcoxon signed-rank tests, $P = 0.39$) T1D PRS to the T2D group, which is consistent with it only containing a small proportion of T1D cases. The ambiguous group has a greater T2D PRS than the T2D group ($P = 5.6 \times 10^{-4}$), likely because the ambiguous cases were defined as having earlier age of onset, which has previously been shown to be associated with higher polygenic risk score in T2D cases[36,38]. We used multiple linear regression to assess the association between the T1D and T2D PRSs and clinical features (age at diagnosis, BMI at diagnosis, time to insulin, and the presence of T1D and T2D diagnostic code) in the ambiguous group and the reference T2D group (Fig. 3). Age of diagnosis was significantly negatively associated with the T2D PRS within the T2D cases ($P = 3.73 \times 10^{-7}$), as expected[36,38]. There was no significant association between BMI at diagnosis, time to insulin or the presence of T1D/T2D diagnostic codes with any of the PRSs.

Discussion

Our study combines routine health data from a large population-based study of British Bangladeshi and Pakistani individuals with ancestry-optimised polygenic risk scores to estimate the proportion of T1D cases, and the rate of misclassification, in a group of people with diabetes who have ambiguous clinical features. We undertook this

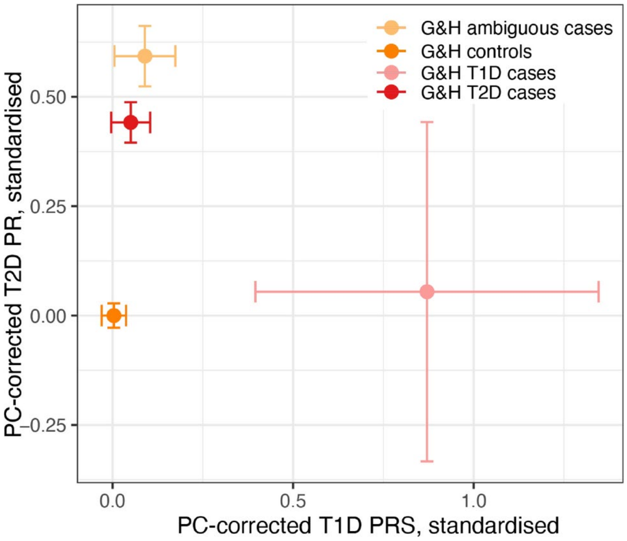

Fig. 2. Average PC-corrected T1D and T2D polygenic risk score (PRS) of subgroups in Genes and Health (G&H), with 95% confidence intervals. These have been standardised such that the controls have a mean of 0 and variance 1.

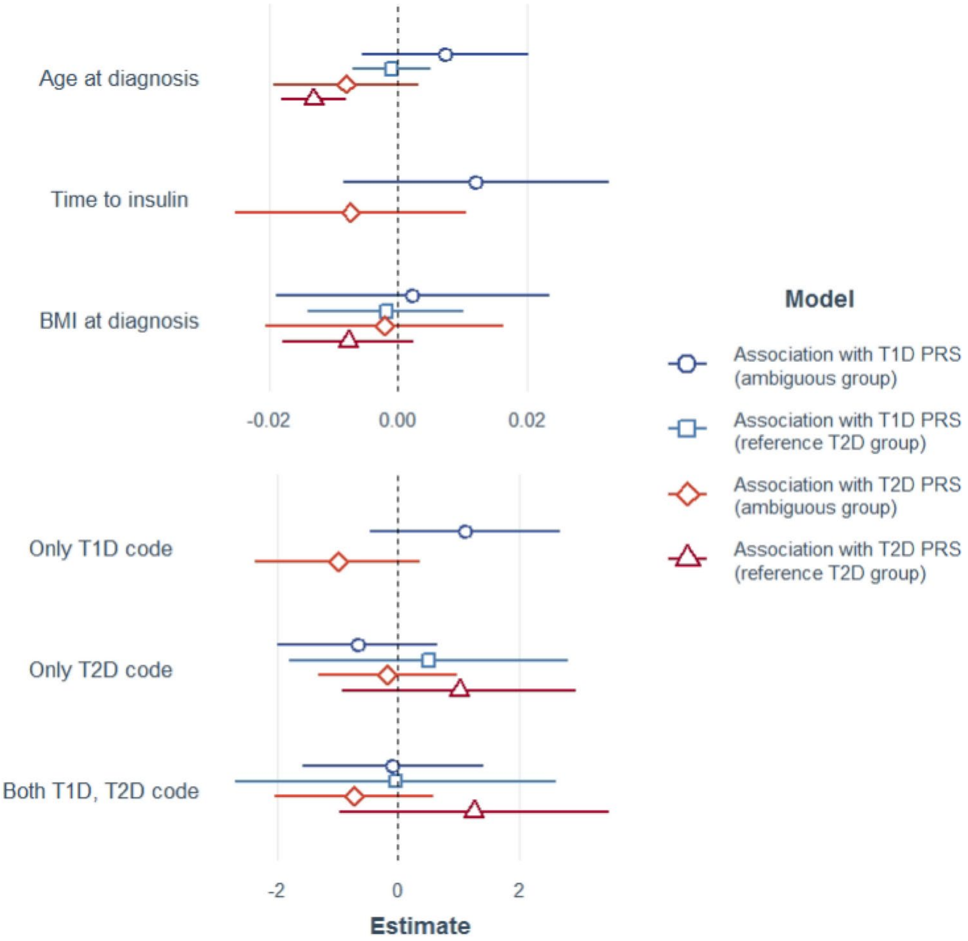

Fig. 3. Results from multiple linear regressions of the PC-corrected T1D or T2D PRSs on the indicated clinical variables within either the ambiguous or T2D cases from G&H. Points show the point estimates for the effect size and lines show the 95% confidence intervals. The estimates are split into two panels due to the difference in their scale. Note that the regression within the reference T2D group excluded 'time to insulin' and 'having only a T1D code' since these were not relevant because of how this group was defined.

study due to the well-recognised challenges of correctly classifying diabetes type in South Asian populations in which young onset T2D is increasingly common and differences in fat distribution and fat mass mean clinical features such as BMI are unhelpful. We showed that the clinical features that are commonly used to help classify T1D and T2D (age and BMI at diagnosis, and time to insulin) were not associated with the ancestry-optimised PRS within the ambiguous group, implying that they have limited utility to help distinguish T1D from T2D within this group. This study is the first to systematically assess the likely rates of misclassification in a large, real world south Asian population receiving routine diabetes care. Our goal was to use PRS as an epidemiological tool to estimate the rate of misclassification in British South Asians with ambiguous insulin-treated diabetes, as opposed to identified specific misclassified individuals. It is not clear from our findings whether combining ancestry-optimised PRSs with clinical and metabolic measurements may prove useful to help improve the diagnosis of individuals with diabetes, and this is an important area for future research.

We used standard criteria to define T1D and T2D cases from our population-based cohort with linked health record data. Diabetes autoantibodies and measures of beta cell function (serum C-peptide) were rarely recorded. We derived a large group of individuals whose diabetes diagnosis was clinically 'ambiguous' but was characterised by insulin treatment. The clinically ambiguous group included people classified (by diagnostic codes) as having T1D, T2D or both. We corrected T1D and T2D PRSs, which had previously been shown to perform reasonably well in South Asians[19,27], for genetic principal components to remove spurious differences due to population structure (Supplementary Figs. 1, 3, 4), and showed that these ancestry-corrected PRSs were significantly associated with case/control status in G&H (Fig. 2). Using data from an Indian T1D reference cohort and these ancestry-optimised PRSs, we estimated that the true proportion of T1D in this clinically ambiguous group is most likely in the range of 3.6–10.2% (median estimate across approaches = 5.9%, although with wide 95% confidence intervals). Diagnostic codes were not significantly associated with the T1D and T2D PRSs within the ambiguous group (Fig. 3), suggesting misclassification is therefore likely. When we removed those individuals who had only diagnostic codes for T1D (i.e. those where we assume clinical suspicion for T1D was highest) from the analysis, and re-estimated the proportion of T1D cases amongst the remaining ambiguous individuals, we obtained a median estimate of 4.5% across approaches, which we regard as an estimate of putative misclassification rate. This is lower than estimates obtained in previous studies in Europeans that only relied on broad clinical criteria (7–15%[6,7]), but it is difficult to draw conclusions from this due to our wide error bars, as well as the distinct populations, cohorts, and methodologies.

Our work builds on the methods developed by Evans et al.[23], and applies ancestry-adjusted PRSs to estimate disease prevalence within clinically-defined groups of individuals with diabetes. This approach allowed us to employ a set of reference cases with different recent genetic ancestry to the target population (i.e. Indian versus Pakistani/Bangladeshi). We conducted PCA on both the reference and target sample combined, to ensure that the PCs captured the genetic diversity within the full sample, particularly given the strong fine-scale population structure in South Asia[39,40]. It was critical to carry out the ancestry correction using the relationship between the PCs and PRSs defined in controls alone, since, within our sample, T1D case status was strongly correlated with ancestry and we did not wish to 'correct away' the true difference in T1D PRS between cases and controls.

Our study has certain limitations. The low numbers of T1D cases that could be robustly identified by clinical criteria in our G&H sample meant that we could not optimise the weights and choice of SNPs for the T1D PRS within this sample, or use the true T1D cases from G&H as our reference sample when estimating the fraction of T1D cases in the ambiguous group. The steps outlined above using T1D cases from a study of Indian individuals and a PCA-based method to correct for population structure has mitigated this. If we had access to a well-powered T1D GWAS within individuals of Pakistani and Bangladeshi ancestry, this would likely boost the accuracy of the PRSs and improve the accuracy of our inference. Having said that, the European-derived T1D PRS we did use still showed a very marked average difference between T1D cases and controls in G&H (Fig. 2).

The use of routine health data in our analyses has potential limitations. Most importantly, it is likely that poor conversion of clinical notes from secondary care into ICD codes has resulted in our inability to capture diabetic ketoacidosis (DKA) episodes, which are highly suggestive of true T1D. It is therefore likely that, if we had had access to the study participants' full clinical notes or improved coding of DKA data, we would have been able to assign more individuals to the T1D group rather than the ambiguous group. Thus, we may have over-estimated the rate of T1D misclassification, although we do note that our estimates fall within the range previously published in British Europeans using epidemiological methods[6,7]. Furthermore, whilst we have excluded individuals with diagnostic codes for rare types of diabetes (monogenic, secondary), but with prevalence estimates for monogenic diabetes in young populations being as high as 6.5%, it is possible that there are individuals with undiagnosed monogenic diabetes in our ambiguous group who have not been excluded and who could affect our estimate of the proportion of T1D[41]. This is particularly pertinent to our population where consanguineous marriages are common and therefore recessively inherited monogenic diabetes is more likely. Future work is needed to characterise the prevalence of these rare types of diabetes in south Asian populations and to develop south Asian-specific diagnostic aids to guide testing for monogenic diabetes[42–44]. Finally, our definition of insulin treatment used to define reference cases does not distinguish between different types of insulin regime, due to the complexity of longitudinal prescribing data, and a small number of individuals may be on a regime of rapid- or short-acting insulin alone that would not be compatible with type 1 diabetes.

Finally, our study does not attempt to estimate the *total* burden of misclassification. Rather, we restricted our analysis to a group of individuals with diabetes and ambiguous clinical features who were treated with insulin, and excluded those who were not insulin treated and were within three years of their diabetes diagnosis (i.e. those who did not meet the T2D clinical criteria). By restricting our analyses to those who are insulin-treated, we have identified an important subpopulation of people with diabetes who can be readily identified through clinical systems and targeted for further diagnostic assessment (e.g. with measurement of C-peptide and diabetes autoantibodies) to assist correct classification.

Our study confirms that correct classification of diabetes is difficult in populations of British South Asians, and that routinely recorded clinical features at diagnosis cannot be reliably used to discriminate between type 1 and type 2 diabetes. Our PRS-determined estimates of T1D prevalence and misclassification suggest that one in twenty individuals *within the ambiguous group* has not been correctly identified as having T1D. This is an important finding as it suggests that many individuals are receiving poorly-targeted clinical care, are at greater risk of hospital admission due to diabetic emergencies, and may be missing out on technology-supported care such as insulin pump therapy. Such 'ambiguous' diabetes cases could be readily identified in primary care settings by routinely collected health data. We propose that T1D could be identified robustly in the majority of these ambiguous cases using diabetes autoantibody and C-peptide measurements, and that there should be a change to clinical guidance to support their wider use.

Data availability

The Genes & Health data are available upon application as described here https://www.genesandhealth.org/research/scientists-using-genes-health-scientific-research. Access will be granted within a custom-built trusted research environment. Request for access to the Indian cohort should be sent to Giriraj R Chandak at chandakgrc@ccmb.res.in and Chittaranjan Yajnik csyajnik@gmail.com.

Received: 9 May 2024; Accepted: 18 November 2024
Published online: 13 January 2025

References

1. Classification of diabetes mellitus. https://www.who.int/publications/i/item/classification-of-diabetes-mellitus (2019).
2. American Diabetes Association. 2 Classification and diagnosis of diabetes: Standards of medical care in diabetes-2021. *Diabetes Care* **44**, S15–S33 (2021).
3. Thomas, N. J. et al. Frequency and phenotype of type 1 diabetes in the first six decades of life: A cross-sectional, genetically stratified survival analysis from UK Biobank. *Lancet Diabetes Endocrinol* **6**, 122–129 (2018).
4. de Lusignan, S. et al. Miscoding, misclassification and misdiagnosis of diabetes in primary care. *Diabet. Med.* **29**, 181–189 (2012).
5. Tate, A. R. et al. Quality of recording of diabetes in the UK: how does the GP's method of coding clinical data affect incidence estimates? Cross-sectional study using the CPRD database. *BMJ Open* **7**, e012905 (2017).
6. de Lusignan, S. et al. A method of identifying and correcting miscoding, misclassification and misdiagnosis in diabetes: A pilot and validation study of routinely collected data. *Diabet. Med.* **27**, 203–209 (2010).
7. Seidu, S., Davies, M. J., Mostafa, S., de Lusignan, S. & Khunti, K. Prevalence and characteristics in coding, classification and diagnosis of diabetes in primary care. *Postgrad. Med. J.* **90**, 13–17 (2014).
8. Lynam, A. et al. Development and validation of multivariable clinical diagnostic models to identify type 1 diabetes requiring rapid insulin therapy in adults aged 18–50 years. *BMJ Open* **9**, e031586 (2019).
9. Diagnosis - adults. https://cks.nice.org.uk/topics/diabetes-type-1/diagnosis/diagnosis-adults/#:~:text=C%2Dpeptide%20and%2For%20diabetes%2Dspecific%20autoantibody%20titres%20may,slow%20evolution%20of%20hyperglycaemia)%2C%20or.
10. Eastwood, S. V. et al. Algorithms for the capture and adjudication of prevalent and incident diabetes in UK biobank. *PLoS One* **11**, e0162388 (2016).
11. Hope, S. V. et al. Practical Classification Guidelines for Diabetes in patients treated with insulin: A cross-sectional study of the accuracy of diabetes diagnosis. *Br. J. Gen. Pract.* **66**, e315–e322 (2016).
12. Thomas, N. J. et al. Type 1 diabetes defined by severe insulin deficiency occurs after 30 years of age and is commonly treated as type 2 diabetes. *Diabetologia* **62**, 1167–1172. https://doi.org/10.1007/s00125-019-4863-8 (2019).
13. Gholap, N., Davies, M., Patel, K., Sattar, N. & Khunti, K. Type 2 diabetes and cardiovascular disease in South Asians. *Prim. Care Diabetes* **5**, 45–56 (2011).
14. Padilla-Martínez, F., Collin, F., Kwasniewski, M. & Kretowski, A. Systematic review of polygenic risk scores for type 1 and type 2 diabetes. *Int. J. Mol. Sci.* **21**, (2020).
15. Shields, B. M. et al. Can clinical features be used to differentiate type 1 from type 2 diabetes? A systematic review of the literature. *BMJ Open* **5**, e009088 (2015).
16. Tripathi, A., Rizvi, A. A., Knight, L. M. & Jerrell, J. M. Prevalence and impact of initial misclassification of pediatric type 1 diabetes mellitus. *South. Med. J.* **105**, 513–517 (2012).
17. Muñoz, C. et al. Misdiagnosis and diabetic ketoacidosis at diagnosis of type 1 diabetes: patient and caregiver perspectives. *Clin. Diabetes* **37**, 276–281 (2019).
18. Lebovitz, H. E. Insulin: potential negative consequences of early routine use in patients with type 2 diabetes. *Diabetes Care* **34**(Suppl 2), S225–S230 (2011).
19. Mahajan, A. et al. Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. *Nat. Genet.* **54**, 560–572 (2022).
20. Barrett, J. C. et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.* **41**, 703–707 (2009).
21. Bradfield, J. P. et al. A genome-wide meta-analysis of six type 1 diabetes cohorts identifies multiple associated loci. *PLoS Genet.* **7**, e1002293 (2011).
22. Hodgson, S. et al. Integrating polygenic risk scores in the prediction of type 2 diabetes risk and subtypes in British Pakistanis and Bangladeshis: A population-based cohort study. *PLoS Med.* **19**, e1003981 (2022).
23. Evans, B. D. et al. Estimating disease prevalence in large datasets using genetic risk scores. *Nat. Commun.* **12**, 6441 (2021).
24. Mansour Aly, D. et al. Genome-wide association analyses highlight etiological differences underlying newly defined subtypes of diabetes. *Nat. Genet.* **53**, 1534–1542 (2021).
25. Nair, A. T. N. et al. Heterogeneity in phenotype, disease progression and drug response in type 2 diabetes. *Nat. Med.* **28**, 982–988 (2022).
26. Oram, R. A. et al. Utility of diabetes type-specific genetic risk scores for the classification of diabetes type among multiethnic youth. *Diabetes Care* **45**, 1124–1131 (2022).
27. Harrison, J. W. et al. Type 1 diabetes genetic risk score is discriminative of diabetes in non-Europeans: Evidence from a study in India. *Sci. Rep.* **10**, 9450 (2020).
28. Suzuki, K. *et al.* Multi-ancestry genome-wide study in >2.5 million individuals reveals heterogeneity in mechanistic pathways of type 2 diabetes and complications. *medRxiv* (2023) https://doi.org/10.1101/2023.03.31.23287839.
29. Finer, S. et al. Cohort profile: East London Genes & Health (ELGH), a community-based population genomics and health study in British Bangladeshi and British Pakistani people. *Int. J. Epidemiol.* **49**, 20–21i (2020).

30. 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.* **62**, e663–e670 (2012).
31. Huang, Q. Q. et al. Transferability of genetic loci and polygenic scores for cardiometabolic traits in British Pakistanis and Bangladeshis. *Nat. Commun.* https://doi.org/10.1038/s41467-022-32095-5 (2022).
32. Jones, A. G., McDonald, T. J., Shields, B. M., Hagopian, W. & Hattersley, A. T. Latent autoimmune diabetes of adults (LADA) is likely to represent a mixed population of autoimmune (type 1) and nonautoimmune (type 2) diabetes. *Diabetes Care* **44**, 1243–1251 (2021).
33. Oram, R. A. et al. A type 1 diabetes genetic risk score can aid discrimination between type 1 and type 2 diabetes in young adults. *Diabetes Care* **39**, 337–344 (2016).
34. Martin, A. R. et al. Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).
35. Thomas, N. J. et al. Identifying type 1 and 2 diabetes in research datasets where classification biomarkers are unavailable: Assessing the accuracy of published approaches. *J. Clin. Epidemiol.* **153**, 34–44 (2023).
36. Mars, N. et al. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nat. Med.* **26**, 549–557 (2020).
37. Ghatan, S. et al. Defining type 2 diabetes polygenic risk scores through colocalization and network-based clustering of metabolic trait genetic associations. *Genome Med.* **16**, 10 (2024).
38. Ashenhurst, J. R. et al. A polygenic score for type 2 diabetes improves risk stratification beyond current clinical screening factors in an ancestrally diverse sample. *Front. Genet.* **13**, 871260 (2022).
39. Wall, J. D. et al. South Asian medical cohorts reveal strong founder effects and high rates of homozygosity. *Nat. Commun.* **14**, 3377 (2023).
40. Arciero, E. et al. Fine-scale population structure and demographic history of British Pakistanis. *Nat. Commun.* **12**, 7189 (2021).
41. Passanisi, S., Salzano, G., Bombaci, B. & Lombardo, F. Clinical and genetic features of maturity-onset diabetes of the young in pediatric patients: a 12-year monocentric experience. *Diabetol. Metab. Syndr.* **13**, 96 (2021).
42. Shields, B. M. et al. The development and validation of a clinical prediction model to determine the probability of MODY in patients with young-onset diabetes. *Diabetologia* **55**, 1265–1272 (2012).
43. Riddle, M. C. et al. Monogenic diabetes: from genetic insights to population-based precision in care. Reflections from a diabetes care editors' expert forum. *Diabetes Care* **43**, 3117–3128 (2020).
44. Misra, S. et al. South Asian individuals with diabetes who are referred for MODY testing in the UK have a lower mutation pick-up rate than white European people. *Diabetologia* **59**, 2262–2265 (2016).
45. Huang, Q. Q. et al. Transferability of genetic loci and polygenic scores for cardiometabolic traits in British Pakistani and Bangladeshi individuals. *Nat. Commun.* **13**, 4664 (2022).
46. Guo, Y. et al. Illumina human exome genotyping array clustering and quality control. *Nat. Protoc.* **9**, 2643–2662 (2014).

Acknowledgements

For the Genes & Health cohort We thank Social Action for Health, Centre of The Cell, members of our Community Advisory Group, and staff who have recruited and collected data from volunteers. We thank the NIHR National Biosample Centre (UK Biocentre), the Social Genetic & Developmental Psychiatry Centre (King's College London), Wellcome Sanger Institute, and Broad Institute for sample processing, genotyping, sequencing and variant annotation. We thank: Barts Health NHS Trust, NHS Clinical Commissioning Groups (City and Hackney, Waltham Forest, Tower Hamlets, Newham, Redbridge, Havering, Barking and Dagenham), East London NHS Foundation Trust, Bradford Teaching Hospitals NHS Foundation Trust, Public Health England (especially David Wyllie), Discovery Data Service/Endeavour Health Charitable Trust (especially David Stables), and NHS Digital for GDPR-compliant data sharing backed by individual written informed consent. Most of all we thank all of the volunteers participating in Genes & Health and UK Biobank. We thank the current Genes & Health Research Team (in alphabetical order by surname): Shaheen Akhtar, Mohammad Anwar, Omar Asgar, Samina Ashraf, Saeed Bidi, Gerome Breen, James Broster, Raymond Chung, David Collier, Charles J Curtis, Shabana Chaudhary, Megan Clinch, Grainne Colligan, Panos Deloukas, Ceri Durham, Faiza Durrani, Fabiola Eto, Sarah Finer, Joseph Gafton, Ana Angel Garcia, Chris Griffiths, Joanne Harvey, Teng Heng, Sam Hodgson, Qin Qin Huang, Matt Hurles, Karen A Hunt, Shapna Hussain, Kamrul Islam, Vivek Iyer, Ben Jacobs, Ahsan Khan, Cath Lavery, Sang Hyuck Lee, Robin Lerner, Daniel MacArthur, Daniel Malawsky, Hilary Martin, Dan Mason, Rohini Mathur, Mohammed Bodrul Mazid, John McDermott, Caroline Morton, Bill Newman, Elizabeth Owor, Asma Qureshi, Samiha Rahman, Shwetha Ramachandrappa, Mehru Reza, Jessry Russell, Nishat Safa, Miriam Samuel, Michael Simpson, John Solly, Marie Spreckley. Daniel Stow, Michael Taylor, Richard C Trembath, Karen Tricker, Nasir Uddin, David A van Heel, Klaudia Walter, Caroline Winckley, Suzanne Wood, John Wright, Julia Zollner.

Author contributions

S.F., H.C.M., and G.P. conceived the project. G.P. prepared, cleaned and analysed the clinical data and defined the clinical cohorts, with contribution from S.H. and M.L. T.L. and T.H.H. undertook Q.C. of the Genes & Health genotype data. A.S. generated the genotype data from the Indian cohort, and carried out initial Q.C. on these, in consultation with G.R.C. T.L. undertook Q.C. of Genes & Health and further Q.C. on the Indian genotype data and imputed both datasets to TopMED with input from A.S., D.F., M.W., R.O. and G.R.C. T.L. conducted the PCA and calculated the PRSs with input from A.S., Q.Q.H., R.O., M.W., and G.R.C. T.L. correlated the PRSs with clinical data and, with help from N.T., ran the estimation method. DvH and S.F. lead the G&H study and, with the G&H Research team, they supervise all data collection. C.S.Y. created the Indian clinical cohort and GRC supervised genotyping of its cases and controls with A.S. Intellectual contribution to analyses was provided by R.O., M.W., N.T., C.S.Y. and G.R.C. S.F., H.M., G.P. and T.L. wrote the first draft of the manuscript, and all authors commented on it.

Funding

Genes & Health is/has recently been core-funded by Wellcome (WT102627, WT210561), the Medical Research Council (UK) (M009017, MR/X009777/1), Higher Education Funding Council for England Catalyst, Barts Charity (845/1796), Health Data Research UK (for London substantive site), and research delivery support from the NHS National Institute for Health Research Clinical Research Network (North Thames). Genes & Health is/has recently been funded by Alnylam Pharmaceuticals, Genomics PLC; and a Life Sciences Industry Consortium of AstraZeneca PLC, Bristol-Myers Squibb Company, GlaxoSmithKline Research and Development Limited, Maze Therapeutics Inc, Merck Sharp & Dohme LLC, Novo Nordisk A/S, Pfizer Inc, Takeda Development Centre Americas Inc. GP was supported via the National Institute for Health Research funded Specialist Foundation Programme. SH is supported by a PhD Fellowship from the Wellcome Trust Health Advances for Underrepresented Populations and Diseases PhD programme at Queen Mary University of London. This research was funded in whole, or in part, by the Wellcome Trust Grant 220540/Z/20/A, 'Wellcome Sanger Institute Quinquennial Review 2021–2026'. For the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. The Indian cohort of T1D patients and non-diabetic controls has been supported from multiple funding sources. The recruitment and early phenotyping including antibodies estimation has been supported by a Diabetes UK grant (15/0005297). High throughput genotyping has been supported by funds from Council for Scientific and Industrial Research (CSIR), Ministry of Science and Technology, Govt. of India, New Delhi. GRC is thankful to the Science and Engineering Research Board, Department of Science and Technology, Ministry of Science and Technology, Govt. of India, New Delhi for the JC Bose Fellowship. AS is supported through a Fellowship from CSIR for his PhD.

Declarations

Competing interests

SF and HCM have received salary contributions via the Genes & Health Industry Consortium of AstraZeneca PLC, Bristol-Myers Squibb Company, GlaxoSmithKline Research and Development Limited, Maze Therapeutics Inc, Merck Sharp & Dohme LLC, Novo Nordisk A/S, Pfizer Inc, Takeda Development Centre Americas Inc. ML has received speaker's fees from Medtronic and Insulet. Other authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-80348-8.

Correspondence and requests for materials should be addressed to H.C.M. or S.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Genes & Health Research Team

Shaheen Akhtar[9], Ana Angel[4], Omar Asgar[10], Samina Ashraf[9], Saeed Bidi[4], Gerome Breen[11], James Broster[7], Raymond Chung[11], David Collier[13], Charles J. Curtis[11], Shabana Chaudhary[7], Grainne Colligan[12], Panos Deloukas[13], Ceri Durham[12], Faiza Durrani[4], Fabiola Eto[4], Sarah Finer[4], Sam Hodgson[4], Qin Qin Huang[1], Karen A. Hunt[7], Matt Hurles[1], Shapna Hussain[7], Kamrul Islam[7], Vivek Iyer[1], Benjamin M. Jacobs[4], Georgios Kalantzis[1], Ahsan Khan[14], Claudia Langenberg[15], Cath Lavery[7], Sang Hyuck Lee[11], Daniel MacArthur[16], Eamonn Maher[17], Daniel Malawsky[1], Sidra Malik[7], Hilary Martin[1], Dan Mason[9], Rohini Mathur[4], Mohammed Bodrul Mazid[7], John McDermott[10], Caroline Morton[4], Bill Newman[10], Vladimir Ovchinnikov[1], Elizabeth Owor[7], Iaroslav Popov[1], Asma Qureshi[7], Mehru Raza[4], Jessry Russell[7], Stuart Rison[4], Nishat Safa[7], Annum Salman[4], Miriam Samuel[4], Moneeza K. Siddiqui[4], Michael Simpson[11], John Solly[7], Marie Spreckley[7], Daniel Stow[4], Michael Taylor[13], Richard C. Trembath[11], Karen Tricker[10], David A. Heel[7], Klaudia Walter[1], Jan Whalley[7], Caroline

Winckley[18], Suzanne Wood[13], John Wright[9], Sabina Yasmin[7], Ishevanhu Zengeya[7] & Julia Zöllner[4]

[9]Bradford Teaching Hospitals NHS Foundation Trust, Bradford, UK. [10]Manchester University Hospitals, Manchester, UK. [11]King's College London, London, UK. [12]Social Action for Health (Charity), London, UK. [13]William Harvey Research Institute, Queen Mary University of London, London, UK. [14]Waltham Forest Council, London, UK. [15]Precision Healthcare University Research Institute, Queen Mary University of London, London, UK. [16]Garvan Institute, Darlinghurst, Australia. [17]Aston University, Birmingham, UK. [18]Local Clinical Research Networks Greater Manchester Core Team, National Institute Health Research, Clinical Research Network, Manchester, UK.