Assessing the prediction of type 2 diabetes risk using polygenic and clinical risk scores in South Asian study populations

Madhusmita Rout, Gurpreet S. Wander, Sarju Ralhan, Jai Rup Singh, Christopher E. Aston, Piers R. Blackett, Steven Chernausek and Dharambir K. Sanghera

Abstract

Background: Genome-wide polygenic risk scores (PRS) have shown high specificity and sensitivity in predicting type 2 diabetes (T2D) risk in Europeans. However, the PRS-driven information and its clinical significance in non-Europeans are underrepresented. We examined the predictive efficacy and transferability of PRS models using variant information derived from genome-wide studies of Asian Indians (AIs) (PRS_{AI}) and Europeans (PRS_{EU}) using 13,974 AI individuals.

Methods: Weighted PRS models were constructed and analyzed on 4602 individuals from the Asian Indian Diabetes Heart Study/Sikh Diabetes Study (AIDHS/SDS) as discovery/training and test/validation datasets. The results were further replicated in 9372 South Asian individuals from UK Biobank (UKBB). We also assessed the performance of each PRS model by combining data of the clinical risk score (CRS).

Results: Both genetic models (PRS_{AI} and PRS_{EU}) successfully predicted the T2D risk. However, the PRS_{AI} revealed 13.2% odds ratio (OR) 1.80 [95% confidence interval (CI) 1.63–1.97; $p = 1.6 \times 10^{-152}$] and 12.2% OR 1.38 (95% CI 1.30–1.46; $p = 7.1 \times 10^{-237}$) superior performance in AIDHS/SDS and UKBB validation sets, respectively. Comparing individuals of extreme PRS (ninth decile) with the average PRS (fifth decile), PRS_{AI} showed about two-fold OR 20.73 (95% CI 10.27–41.83; $p = 2.7 \times 10^{-17}$) and 1.4-fold OR 3.19 (95% CI 2.51–4.06; $p = 4.8 \times 10^{-21}$) higher predictability to identify subgroups with higher genetic risk than the PRS_{EU}. Combining PRS and CRS improved the area under the curve from 0.74 to 0.79 in PRS_{AI} and 0.72 to 0.75 in PRS_{EU}.

Conclusion: Our data suggest the need for extending genetic and clinical studies in varied ethnic groups to exploit the full clinical potential of PRS as a risk prediction tool in diverse study populations.

Keywords: Asian Indians, clinical risk score, diabetes, polygenic risk score, UK Biobank

Correspondence to:
Dharambir K. Sanghera
Department of Pediatrics, College of Medicine, University of Oklahoma Health Sciences Center, 940 Stanton L. Young Blvd., Rm 317 BMSB, Oklahoma City, OK 73104, USA

Harold Hamm Diabetes Center, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA

Department of Pharmaceutical Sciences, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA

Department of Physiology, College of Medicine, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA

Oklahoma Center for Neuroscience, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA
Dharambir-sanghera@ ouhsc.edu

Madhusmita Rout
Department of Pediatrics, Section of Genetics, College of Medicine, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA

Gurpreet S. Wander
Sarju Ralhan
Hero DMC Heart Institute, Ludhiana, Punjab, India

Jai Rup Singh
Central University of Punjab, Bathinda, Punjab, India

Christopher E. Aston
Section of Developmental and Behavioral Pediatrics, College of Medicine, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA

Introduction

Type 2 diabetes (T2D) is a complex chronic disease that challenges public health globally due to its continued increased prevalence. The International Diabetes Federation estimated that the global prevalence of 463 million people with diabetes in 2019 will increase to 700 million by 2045; ~90% of the total will comprise T2D cases.[1] The region of South Asia (consisting of people from India, Pakistan, Bangladesh, Nepal, Bhutan, and Sri Lanka) is the epicenter of the growing epidemic of T2D due to rapid urbanization, immense population growth, and aging.[2,3] Studies performed on immigrant Indians outside India have shown that T2D and coronary artery disease incidence is three to six times higher in immigrant Asian Indians (AIs) than Euro-Caucasians. The onset of T2D is roughly a decade earlier [even at

Piers R. Blackett
Steven Chernausek
Department of Pediatrics,
Section of Endocrinology,
College of Medicine,
University of Oklahoma
Health Sciences Center,
Oklahoma City, OK, USA

Harold Hamm Diabetes
Center, University
of Oklahoma Health
Sciences Center,
Oklahoma City, OK, USA

a lower body mass index (BMI)] in AIs than in Europeans.[4–7] While environmental factors play an essential role in T2D susceptibility, T2D has a strong genetic component, as has been established by many studies of different designs.[8] The heritability estimates of T2D range from 40% to 70%, out of which at least 10–20% is explained by common variants identified in extensive genome-wide association studies (GWAS).[9–11] However, the genetics of T2D is poorly characterized in people of AI descent. Thus far, only a handful of T2D GWAS has been published on South AIs, who comprise more than a quarter of the world population.[12–14] Most of the genetic studies on South Asians have been performed on immigrants, such as Indians living in the UK or on Pakistani populations. These studies do not putatively reflect the underlying genetic architecture of phenotypic traits and their interactions with other clinical and lifestyle factors of native AIs.

There has been a growing interest in using a cumulative genetic score of T2D-associated variants combined into a polygenic risk score (PRS) to identify individuals with a high genetic risk for the clinical prediction of future occurrence, early prognosis, intervention, and prevention of T2D.[15,16] The PRS is known to predict the risk for cancer,[17] T2D,[18–20] cardiovascular diseases,[21,22] and traits like height[23] and obesity.[24–26] A recent study demonstrated that European-derived PRS effectively predicted T2D incidence in an indigenous population from the Southwestern USA. Their adult cohort had an area under the curve (AUC) of 0.728, and the hazard ratio (HR) was 1.27 per SD.[27] A similar study was performed on FinnGen biobank data to determine the role of PRS in predicting susceptibility to five common diseases including T2D.[28] However, these studies have predominantly focused on populations of European origin, with little information on the transferability of genetic risk loci identified in Europeans onto populations of AI ancestries.[29,30] Both genetic and environmental factors contribute to causing T2D. The predictive value of multiple nongenetic factors in relation to PRS has not been thoroughly examined. Recent studies by He *et al.* have suggested a polyexposure score (PXS) that combines multiple correlated nongenetic exposures and lifestyle factors and compared against PRS to elucidate which score had better efficacy in predicting T2D. Their results highlighted the addition of PGS and PXS to clinical risk score (CRS) improved T2D classification accuracy.[31,32] In this study, we constructed an ancestry-specific PRS (PRS$_{AI}$) for T2D using candidate variants derived from our Asian Indian Diabetic Heart Study/Sikh Diabetes Study (AIDHS/SDS).[13,33–36] We also built a PRS (PRS$_{EU}$) using summary statistics from GWAS meta-analyses from seven European cohorts.[37] We compared the predictive efficacy and transferability between PRS derived from AIDHS/SDS and Europeans using validation datasets-1 and 2, including data from South Asians from the UK Biobank (UKBB). We also evaluated the performance of genetic scores by integrating CRS in the risk assessment models.

Materials and methods

Study subjects
All study participants of AIDHS/SDS were from the Northern part of India and were recruited from 2003 to 2009.[13,36,38,39] Study protocol and consent documents were reviewed and approved by the University of Oklahoma Health Sciences Center's Institutional Review Board (IRB #: 2911). Clinical characteristics and demographic details of the studied subjects are presented in Table 1. The Sikh population is a relatively homogenous endogamous community from India. Sikhs are mostly non-smokers with the current study population having only 1.4% of smokers, and ~50% of them are vegetarian. However, they have a high prevalence of T2D and cardiovascular diseases with familial aggregation.[40] The diagnosis of T2D was confirmed by scrutinizing medical records for symptoms, use of medications, and measuring fasting glucose levels following the guidelines of the American Diabetes Association[41] as described previously.[39] The selection of controls was based on a fasting glucose <100.8 mg/dl or a 2-h glucose <141 mg/dl as described previously.[13] Details of other demographic characteristics, including anthropometric measurements, physical activity, smoking, alcohol consumption, and diet, are described elsewhere.[38,42] Briefly, BMI was calculated as weight (kg)/height (meter)^2, and waist-to-hip ratio (WHR) was calculated as the ratio of waist circumference to hip circumference. Blood pressure (BP) was measured twice after a 5-min seated rest period with the participant's feet flat on the floor. All blood samples were obtained at the baseline visit.[38] Subjects with type 1 diabetes, or those with a family member with type 1 diabetes, or rare forms of T2D

Table 1. Clinical characteristics of the AIDHS/SDS and UKBB South Asians.

Trait	Discovery (AIDHS/SDS) ($N = 1616$)			Validation dataset-1 (AIDHS/SDS) ($N = 2986$)			Validation dataset-2 (UKBB South Asians) ($N = 9372$)		
	Controls ($N = 773$)	Cases ($N = 843$)	*p* Value	Controls ($N = 1255$)	Cases ($N = 1731$)	*p* Value	Controls ($N = 7429$)	Cases ($N = 1943$)	*p* Value
Males (%)	53	54		56	58		52	62	
Age (years)	52.21 ± 13.76	53.98 ± 10.59	4×10^{-3}	46.36 ± 14.44	55.33 ± 11.72	3×10^{-73}	52.43 ± 8.35	56.64 ± 7.97	3×10^{-4}
BMI (kg/m^2)	26.26 ± 4.89	27.32 ± 5.07	2×10^{-5}	26.08 ± 4.63	27.25 ± 4.78	3.2×10^{-11}	26.77 ± 4.20	28.68 ± 4.81	1.4×10^{-9}
Waist (cm)	92.36 ± 12.00	94.90 ± 11.64	1×10^{-5}	89.69 ± 12.40	94.10 ± 12.27	5.5×10^{-21}	89.90 ± 11.48	97.41 ± 11.42	3.5×10^{-139}
Waist-to-hip ratio	0.94 ± 0.08	0.96 ± 0.07	8×10^{-6}	0.92 ± 0.09	0.96 ± 0.09	5.1×10^{-47}	0.89 ± 0.08	0.95 ± 0.08	1.5×10^{-8}
FBG (mg/dL)	95.99 ± 12.36	176.35 ± 71.05	3×10^{-159}	96.90 ± 14.62	175.77 ± 72.40	3×10^{-235}	$76.93 \pm 32.33*$	$113.72 \pm 71.34*$	1.3×10^{-259}

Values are in mean $\pm$ SD.
*Random glucose levels in the UK Biobank population.
AIDHS/SDS, Asian Indian Diabetic Heart Study/Sikh Diabetes Study; BMI, body mass index; FBG, fasting blood glucose; UKBB, UK Biobank.

subtypes (maturity-onset diabetes of the young[43]) or secondary diabetes (from, e.g. hemochromatosis or pancreatitis) were excluded from the study based on clinical reports, as previously described.[13]

Genotyping, imputation, and quality controls

Genomic DNA was extracted from buffy coats using QIAamp blood kits (Qiagen, Chatsworth, CA, USA) or by the salting-out procedure.[44] Samples were genotyped using the Illumina 660W Quad BeadChip (Illumina Inc., San Diego, CA, USA), Illumina Global Screening Arrays (GSA), and GSA with multi-disease content (GSA+) arrays as described previously.[13,35,43] Also, samples with genotyping call rate <95%, cryptic relatedness, and population outliers were removed, and single nucleotide polymorphisms (SNPs) with genotyping call rate <90%, departures from Hardy–Weinberg equilibrium (HWE) ($p < 10^{-7}$) or minor allele frequency (MAF) <5% were excluded before association testing. To increase genome coverage, data was imputed using Minimac4[45] with 1000G Phase3 v5 multiethnic reference panel in NCBI Build 37 (hg19) and coordinates as described.[35,36] Quality control for the imputed SNPs included removing variants with an imputation certainty 'info score' $R^2 < 0.8$, and SNPs significantly deviated from HWE ($p < 1 \times 10^{-6}$) before further analysis.

UKBB study participants (validation dataset-2)

To validate the results of PRS$_{AI}$, we used data from UKBB for South Asian ancestry ($n = 9372$), including Indians, Pakistani, Bangladeshi, and any other Asian background from the UKBB following the approval of the current research project (application # 78635).[46] T2D was characterized based on doctor-diagnosed disease phenotype and glycated hemoglobin levels (HbA1c). For genetic analysis, we used imputed data released by the UKBB specific for South Asian subjects. We excluded outliers for heterozygosity or genotype missing rates (0.2 > missing rate) as well as ambiguous SNPs (MAF > 0.44). Participants with inconsistent reports and genotypic inferred sex inconsistencies or withdrawn consent were removed, as explained previously.[46]

Statistical analysis

The genome-wide PRS analysis was performed by using discovery and validation datasets. The discovery/training set included 1616 individuals (843 cases/773 controls) genotyped using Illumina 660 Quad chip arrays. The validation dataset-1 was comprised of 2986 (1731 cases/1255 controls) individuals who were genotyped using Illumina's GSA+ and GSA arrays. Both discovery and validation dataset-1 comprised individuals from AIDHS/SDS.[13] The additional validation dataset-2 included 9372 South Asians (1943 T2D cases and 7429 controls) from UKBB. To adjust for residual population stratification, age, sex, BMI, and five principal components (PCs) were included as covariates. As the existing HapMap2 or HapMap3 and 1000 Genomes data do not include Sikhs, the PCs used for this correction were estimated using our Sikh population

sample and not the HapMap populations.[13] The PC information for the UKBB dataset was obtained from the data provided by the UKBB. Associations of directly genotyped and imputed SNPs with T2D were tested using logistic regression and an additive genetic model.

The selection criteria of SNPs from 46,985,978 (common and rare) for constructing Sikh-specific PRS_{AI} were based on: (1) SNPs from South Asian T2D GWASs[12,13,47,48]; (2) SNPs with independent association signals with $p < 10^{-2}$; (3) including SNPs with MAF > 0.01 and MAF < 0.45, and excluding SNPs with info score $R^2 < 0.8$, insertion/deletions, and multiallelic SNPs; (4) including SNPs with independent association signals with $p < 10^{-4}$; and (5) linkage disequilibrium (LD) pruning using R^2 (LD) = 0.80. After the screening, 2921 significant SNPs were selected for the construction of the PRS_{AI} (Supplemental Table 1). The individual-level regression coefficients were multiplied by the number of risk alleles to compute the PRS in training and test sets as described previously.[49] A fixed and random-effect, inverse-variance meta-analysis (implemented in METAL)[50] was used to combine the results of the AIDHS/SDS and UKBB.

To construct European PRS (PRS_{EU}), we used the summary statistics data from O'Connor *et al.*,[51] which comprised datasets from seven European cohorts ($n = 312,646$) containing 33,122,978 variants, and available for both additive and recessive models. We used the additive model for our analysis. The selection criteria of SNPs for constructing European PRS_{EU} were based on: (1) SNPs with independent association signals with $p < 10^{-2}$; (2) including SNPs with MAF > 0.01 and MAF < 0.45 and selecting only biallelic SNPs; (3) SNPs with independent association signals with $p < 10^{-4}$ after LD pruning using $R^2 = 0.80$. A total of 1847 significant SNPs were selected for the construction of the PRS_{EU} (Supplemental Table 1). Age, sex, BMI, and five PCs were used as covariates. To test the discrimination capability at the extreme tail of the PRS, we divided the PRS into deciles and calculated the odds ratio (OR) of these high-risk individuals in the ninth decile *versus* the fifth decile of the samples. Since the extreme deciles (e.g. decile 1st and 10th) had no individuals who were T2D and non-T2D, respectively, risk prediction of extreme polygenic score *versus* average score was computed comparing ninth and tenth deciles (Supplemental Table 2).

The CRS was calculated following the modified version of the Joint British Society (JBS) risk score in both AIDHS/SDS and UKBB.[52] The modified JBS risk factors include age (<30 = 0, ⩾30–50 = 1, ⩾51–70 = 2, >70 = 3); gender (female = 1, male = 2); BMI (⩽23 = 0, 23–27.5 = 1, >27.5 = 2); smoking habits (yes = 1, no = 0); hypertension {hypertensive = 1 [systolic BP (SYSBP) ⩾110 mmHg and diastolic BP (DBP) ⩾90 mmHg], non-hypertensive = 0}; an independent assessment of SYSBP (<89 mmHg = 0, 90–130 mmHg = 1, >130 mmHg = 2); family history of diabetes (yes = 1, no = 0); and any other metabolic disorders such as arthritis or kidney disease (yes = 1, no = 0). Note that we use a lower BMI cut-off for defining obesity in AIs based on the ethnicity-specific guidelines proposed by the World Health Organization (WHO).[53]

The prediction efficiency of PRS models was assessed by generating the receiver operative characteristic curve (ROC), which is a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity). It calculates the corresponding AUC, which ranges from 0.5 as a total lack of discrimination to a maximum of 1.0 (perfect discrimination).[54] All analyses were performed using PLINK 2.0,[55] SVS version 8.9.1 (Golden Helix, Bozeman, MT, USA), and SPSS software version 27 (IBM, Ney York City, USA).

Results

The clinical characteristics of the study participants are presented in Table 1. As expected, individuals with T2D had significantly higher BMI, waist circumference, WHR, and fasting blood glucose than controls in both the discovery and the validation datasets (Table 1).

In comparison to the European-derived PRS, the ancestry-derived PRS_{AI} revealed a stronger association of T2D reflected by the higher ORs in validation dataset-1 OR 1.80 (95% CI 1.63–1.97; $p = 1.6 \times 10^{-152}$) compared to the PRS_{EU} 1.59 (95% CI 1.42–1.77; $p = 2.4 \times 10^{-27}$). A similar trend was observed in validation dataset-2, showing slightly higher ORs of PRS_{AI} 1.38 (95% CI 1.30–1.46; $p = 7.1 \times 10^{-237}$) compared to the

PRS_{EU} 1.23 (95% CI 1.15–1.31; $p = 7.5 \times 10^{-36}$). Interestingly, the risk for T2D associated with CRS alone showed the ORs of 1.77 (95% CI 1.64–1.90; $p = 1.9 \times 10^{-254}$) in validation dataset-1 and 1.55 (95% CI 1.47–1.63; $p = 7.2 \times 10^{-320}$) in validation dataset-2. Combining validation datasets 1 and 2, the CRS was the strongest predictor of T2D risk 1.56 (95% CI 1.48–1.64; $p = 6.7 \times 10^{-29}$), followed by PRS_{AI} 1.44 (95% CI 1.37–1.52; $p = 3.2 \times 10^{-23}$). The lowest predictive outcome was observed in the PRS_{EU} 1.28 (1.21–1.35; $p = 1.4 \times 10^{-11}$; Table 2, Figure 1).

We further assessed the joint effect of PRS models integrating with the CRS. There was a significant improvement in the performance of PRS models showing an increase in ORs of 1.80 (95% CI 1.63–1.97; $p = 1.6 \times 10^{-152}$) to 1.90 (95% CI 1.84–1.95; $p = 7.0 \times 10^{-285}$) in PRS_{AI} and 1.59 (95% CI 1.42–1.77; $p = 2.4 \times 10^{-27}$) to 1.89 (95% CI 1.82–1.96; $p = 2.4 \times 10^{-280}$) in PRS_{EU} in validation dataset-1. A similar strong trend was observed in the increase in ORs from 1.38 (95% CI 1.30–1.46; $p = 7.1 \times 10^{-237}$) to 1.47 (95% CI 1.45–1.50; $p = 4.8 \times 10^{-315}$) in PRS_{AI} and 1.23 (95% CI 1.15–1.31; $p = 7.5 \times 10^{-36}$) to 1.45 (95% CI 1.43–1.48; $p = 3.0 \times 10^{-303}$) in PRS_{EU} for the joint effects of PRS + CRS models (Table 3, Figure 1).

Next, we compared individuals from the extreme PRS with the average PRS, and divided the PRS scores in deciles. Comparing participants at the ninth decile *versus* the middle fifth decile revealed an OR of 20.73 (95% CI 10.27–41.83; $p = 2.7 \times 10^{-17}$) for the PRS_{AI} and 11.29 (95% CI 6.00–21.24; $p = 6.0 \times 10^{-14}$) for the PRS_{EU}. Similar results were observed for the UKBB, which showed a higher OR of 3.19 (95% CI 2.51–4.06; $p = 4.8 \times 10^{-21}$) for the PRS_{AI} while the OR was 2.31 (95% CI 2.07–2.56; $p = 1.2 \times 10^{-19}$) for PRS_{EU} (Table 4). Comparing participants at the first decile *versus* the middle fifth decile revealed an OR of 162.76 (95% CI 62.51–423.79; $p = 1.8 \times 10^{-35}$) for the PRS_{AI} and 17.38 (95% CI 10.75–28.09; $p = 2.4 \times 10^{-31}$) for the PRS_{EU}. Similar results were observed for the UKBB, which showed a higher OR of 4.08 (95% CI 2.82–5.92; $p = 1.2 \times 10^{-13}$) for the PRS_{AI} while the OR was 2.78 (95% CI 2.08–3.72; $p = 6.7 \times 10^{-12}$) for the PRS_{EU} (Table 5). Next, by analyzing the individuals in the lower three deciles *versus* the upper three deciles, we identified 446

Table 2. Association of PRS with type 2 diabetes.

Model	Discovery (AIDHS/SDS) (N=1616)		Validation dataset-1 (AIDHS/SDS) (N=2986)		Validation dataset-2 (UKBB South Asians) (N=9372)		Meta-analysis ^ (validation dataset-1 + validation dataset-2) (N=12,358)	
	OR (95% CI)	p Value	OR (95% CI)	p Value	OR (95% CI)	p Value	OR (95% CI)	p Value
PRS_{AI}*	2.10 (1.98–2.21)	3.1×10^{-218}	1.80 (1.63–1.97)	1.6×10^{-152}	1.38 (1.30–1.46)	7.1×10^{-237}	**1.44 (1.37–1.52)**	3.2×10^{-23}
PRS_{EU}*	1.95 (1.80–2.09)	9.0×10^{-194}	1.59 (1.42–1.77)	2.4×10^{-27}	1.23 (1.15–1.31)	7.5×10^{-36}	**1.28 (1.21–1.35)**	1.4×10^{-11}
CRS	1.56 (1.39–1.74)	2.8×10^{-79}	1.77 (1.64–1.90)	1.9×10^{-254}	1.55 (1.47–1.63)	7.2×10^{-320}	**1.56 (1.48–1.64)**	6.7×10^{-29}

*Model adjusted for age, gender, BMI.
^Meta-analysis ORs and p-values are marked in bold.
AIDHS/SDS, Asian Indian Diabetic Heart Study/Sikh Diabetes Study; BMI, body mass index; CI, confidence interval; CRS, clinical risk score; OR, odds ratio; PRS_{AI}, Asian Indian ancestry-derived PRS; PRS_{EU}, European-derived PRS; UKBB, UK Biobank.
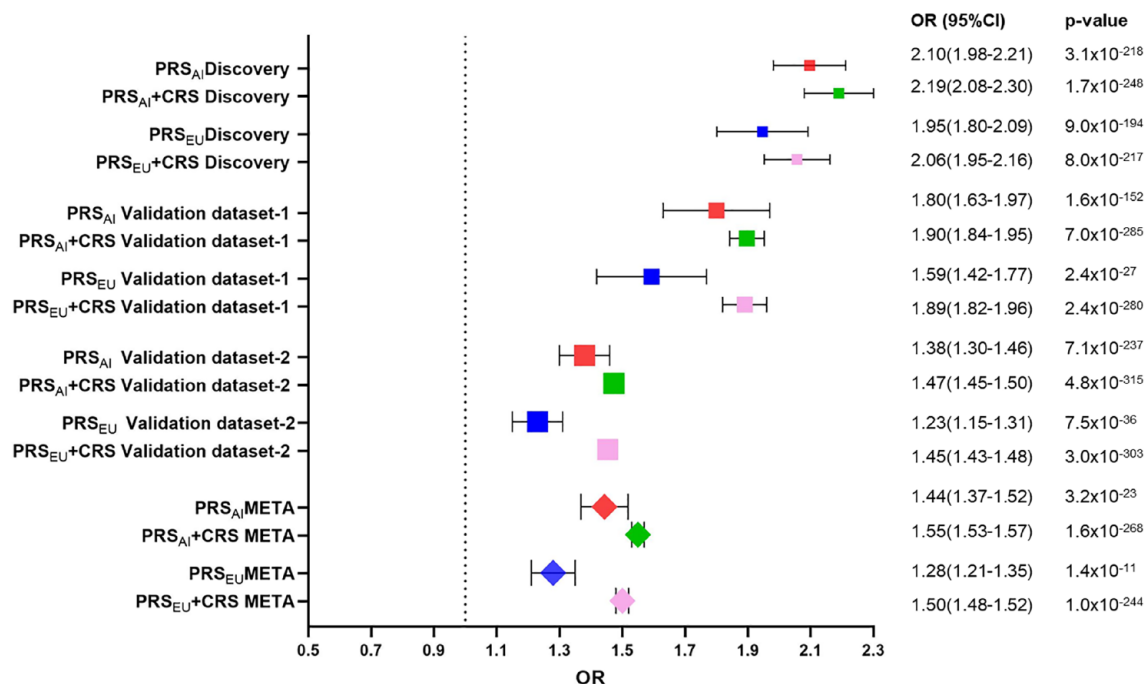
Figure 1. Forest plot showing effect sizes and confidence interval for type 2 diabetes risk using Asian Indian (PRS$_{AI}$), European (PRS$_{EU}$), and combined with CRS trained on Discovery set and tested on validation dataset-1 (AIDHS/SDS) and validation dataset-2 (South Asians from UKBB).

AIDHS/SDS, Asian Indian Diabetic Heart Study/Sikh Diabetes Study; CRS, clinical risk score; PRS$_{AI}$, Asian Indian ancestry-derived PRS; PRS$_{EU}$, European-derived PRS; UKBB, UK Biobank.

and 195 genes and several SNPs in the unnamed gene regions uniquely present in PRS$_{AI}$ and PRS$_{EU}$, respectively, and their MAF differed significantly between the decile extremes (Supplemental Table 5).

Lastly, we performed a sensitivity analysis to test the discriminative accuracy of the PRS models using ROC curve analysis. For the validation dataset-1, the AUC was 0.74 (95% CI 0.69–0.80; $p = 2.1 \times 10^{-11}$) for the PRS$_{AI}$ compared to an AUC of 0.72 (95% CI 0.67–0.78; $p = 9.2 \times 10^{-10}$) for the PRS$_{EU}$. For the validation dataset-2, the AUC was 0.71 (95% CI 0.69–0.73; $p = 2.6 \times 10^{-70}$) for the PRS$_{AI}$ compared to an AUC of 0.69 (95% CI 0.68–0.72; $p = 1.7 \times 10^{-62}$) for the PRS$_{EU}$ (Figure 2). Combining PRS and CRS improved the AUC from 0.74 (95% CI 0.69–0.80; $p = 2.1 \times 10^{-11}$) to 0.79 (95% CI 0.75–0.83; $p = 5.4 \times 10^{-15}$) in PRS$_{AI}$, and 0.72 (95% CI 0.67–0.78; $p = 9.2 \times 10^{-10}$) to 0.75 (95% CI 0.70–0.80; $p = 1.3 \times 10^{-11}$) in PRS$_{EU}$ in validation dataset-1. In validation dataset-2 combining PRS and CRS improved the AUC from 0.71 (95% CI 0.69–0.73; $p = 2.6 \times 10^{-70}$) to 0.73 (95% CI

0.71–0.75; $p = 6.5 \times 10^{-85}$) in PRS$_{AI}$ and 0.69 (95% CI 0.68–0.72; $p = 1.7 \times 10^{-62}$) to 0.71 (95% CI 0.69–0.74; $p = 2.3 \times 10^{-74}$) in PRS$_{EU}$. The AUC was 0.83 (95% CI 0.82–0.85; $p = 6.4 \times 10^{-206}$) and 0.80 (95% CI 0.79–0.81; $p = 2.8 \times 10^{-235}$) for CRS in validation dataset-1 and -2, respectively (Figure 2).

Discussion

Recent genome-wide studies of complex traits have an overwhelming abundance of European-focused information. The underrepresentation of data on other ethnic groups challenges the generalizability of genetic findings across population groups. Even the PRS derived from well-powered European GWAS have shown poor-risk prediction in non-Europeans, suggesting the need for expanding the genetic evaluations globally to improve the clinical utility of PRS.[56] In this study, we compared the predictive efficacy and transferability of the PRS models derived from South Asian GWAS meta-analysis studies and using GWAS results of European T2D consortia studies.

Table 3. Association of combined PRS and CRS with type 2 diabetes.

Model	Discovery (AIDHS/SDS) (N = 1616)				Validation dataset-1 (AIDHS/SDS) (N = 2986)				Validation dataset-2 (UKBB South Asians) (N = 9372)			
	PRS*		PRS + CRS		PRS*		PRS + CRS		PRS*		PRS + CRS	
	OR (95% CI)	p Value	OR (95% CI)	p Value	OR (95% CI)	p Value	OR (95% CI)	p Value	OR (95% CI)	p Value	OR (95% CI)	p Value
PRS_{AI}	2.10 (1.98–2.21)	3.1×10^{-218}	2.19 (2.08–2.30)	1.7×10^{-248}	1.80 (1.63–1.97)	1.6×10^{-152}	1.90 (1.84–1.95)	7.0×10^{-285}	1.38 (1.30–1.46)	7.1×10^{-237}	1.47 (1.45–1.50)	4.8×10^{-315}
PRS_{EU}	1.95 (1.80–2.09)	9.0×10^{-194}	2.06 (1.95–2.16)	8.0×10^{-217}	1.59 (1.42–1.77)	2.4×10^{-27}	1.89 (1.82–1.96)	2.4×10^{-280}	1.23 (1.15–1.31)	7.5×10^{-36}	1.45 (1.43–1.48)	3.0×10^{-303}

*Model adjusted for age, gender, BMI.
AIDHS/SDS, Asian Indian Diabetic Heart Study/Sikh Diabetes Study; BMI, body mass index; CI, confidence interval; CRS, clinical risk score; OR, odds ratio; PRS, polygenic risk scores; PRS_{AI}, Asian Indian ancestry-derived PRS; PRS_{EU}, European-derived PRS; UKBB, UK Biobank.

Both ancestry-specific and European derived PRS predicted T2D risk in AIDHS/SDS and South AIs from UKBB. However, the PRS_{AI} was a better predictor of T2D risk than the European-derived PRS (Table 2 and Figure 1). The overall performance of our PRS_{AI} model was 13.2% and 12.2% superior in validation dataset-1 and 2, respectively. Interestingly, our CRS model showed a strong and independent association with T2D, revealing ORs of 1.77 (95% CI 1.64–1.90) and 1.55 (95% CI 1.47–1.63) in both validation dataset-1 (AIDHS/SDS) and dataset-2 (UKBB). The integration of CRS in the PRS improved the performance of both genetic models. The predictive power of PRS_{AI} after including CRS was improved 5.6% in validation dataset-1 and 6.5% in validation dataset-2. At the same time, the performance of the PRS_{EU} was increased by 18.9% and 17.9% in validation dataset-1 and -2, respectively. Notably, both AI and European models performed equally well, showing the similar ORs of 1.90 (95% CI 1.84–1.95) in PRS_{AI} and 1.89 (95% CI 1.82–1.96) in PRS_{EU} for the validation dataset-1 and ORs of 1.47 (95% CI 1.45–1.50) in PRS_{AI} and 1.45 (95% CI 1.43–1.48) in PRS_{EU} for the validation dataset-2 after integrating CRS in the models. In the combined meta-analysis of validation sets 1 and 2, the PRS_{AI} was 3.3% more efficient than the PRS_{EU}, showing respective ORs of 1.55 *versus* 1.50 (Figure 1). Similarly, combining PRS and CRS improved the AUC from 0.74 to 0.79 in PRS_{AI} and 0.72 to 0.75 in PRS_{EU} in validation dataset-1 and 0.71 to 0.73 in PRS_{AI} and 0.69 to 0.71 in PRS_{EU} in validation dataset-2. Combining clinical risk factors in the PRS has been shown to enhance the prediction of incident T2D in British South Asians.[57] It is also possible that using the lower BMI cut-offs (based on the WHO guidelines) might have improved the sensitivity of our CRS model, and consequently, its integration into genetic risk assessment enhanced the performance of both PRS models.

Upon comparing individuals in the top PRS distribution, the individuals with the ninth decile had more than 20-fold higher T2D risk predictability than those in the middle (fifth decile). At the same time, the difference was 11-fold higher in PRS_{EU} in the validation dataset-1, showing nearly two times higher predictability to identify individuals who would be genetically predisposed to having a higher risk for T2D over the European-derived PRS. Similarly, in the UKBB, the PRS_{AI}

Table 4. Effect sizes and CIs for type 2 diabetes risk comparing PRS (fifth *versus* ninth) deciles in validation datasets 1 and 2.

Cohort	PRS$_{AI}$		PRS$_{EU}$	
	OR (95% CI)	*p* Value	OR (95% CI)	*p* Value
Validation dataset-1 (AIDHS/SDS)	20.73 (10.27–41.83)	2.7×10^{-17}	11.29 (6.00–21.24)	6.0×10^{-14}
Validation dataset-2 (UKBB South Asians)	3.19 (2.51–4.06)	4.8×10^{-21}	2.31 (2.07–2.56)	1.2×10^{-19}

The individuals with extreme PRS in the ninth decile were compared with those in the fifth (middle) decile to determine the risk for PRS$_{AI}$ and PRS$_{EU}$. AIDHS/SDS, Asian Indian Diabetic Heart Study/Sikh Diabetes Study; CI, confidence interval; PRS$_{AI}$, Asian Indian ancestry-derived PRS; PRS$_{EU}$, European-derived PRS; UKBB, UK Biobank.

Table 5. Effect sizes and CIs for type 2 diabetes risk comparing PRS (first *versus* fifth) deciles in validation datasets 1 and 2.

Cohort	PRS$_{AI}$		PRS$_{EU}$	
	OR (95% CI)	*p* Value	OR (95% CI)	*p* Value
Validation dataset-1 (AIDHS/SDS)	162.76 (62.51–423.79)	1.8×10^{-35}	17.38 (10.75–28.09)	2.4×10^{-31}
Validation dataset-2 (UKBB South Asians)	4.08 (2.82–5.92)	1.2×10^{-13}	2.78 (2.08–3.72)	6.7×10^{-12}

The individuals with extreme PRS in the first decile were compared with those in the fifth (middle) decile to determine the risk for PRS$_{AI}$ and PRS$_{EU}$. AIDHS/SDS, Asian Indian Diabetic Heart Study/Sikh Diabetes Study; CI, confidence interval; OR, odds ratio; PRS, polygenic risk scores; PRS$_{AI}$, Asian Indian ancestry-derived PRS; PRS$_{EU}$, European-derived PRS; UKBB, UK Biobank.

showed a 1.4 times higher likelihood of detecting high-risk individuals in the ninth decile over the PRS$_{EU}$. Comparing extreme scores between the first decile *versus* fifth deciles, our models captured high-risk individuals with 7.85-fold in the validation dataset-1 and 1.28-fold in UKBB compared to fifth *versus* ninth decile, using PRS$_{AI}$ model. While these differences were 1.54- and 1.22-fold higher using validation dataset 1 and UKBB, respectively, using the PRS$_{EU}$ model (Tables 4 and 5). These results further suggest the sensitivity and effectiveness of the PRS models even at the lower extremes. These analyses further helped us identify 446 genes uniquely found in PRS$_{AI}$, which could be involved in the T2D development in South Asians (Supplemental Table 5).

Our study has several advantages and limitations. First, our robust genetic analyses include high-quality data from a single sub-population originating from North India with well-characterized clinical phenotypes is a significant strength.[13,49] Second, the PRS$_{AI}$ provided a better prediction over the PRS$_{EU}$ derived from European GWAS. This was further confirmed by comparing the individuals with the extremes of genetic scores *versus* the median scores. Our PRS$_{AI}$ model will

have 2 times and 1.4 times higher likelihood of capturing high-risk individuals in validation cohorts 1 and 2, respectively, than the European-derived model. However, both models performed equally well after integrating CRS in the genetic scores. Third, this is the first study using the PRS approach in a Punjabi population from North India. Limitations include a relatively smaller size of the discovery/training dataset of AIDHS/SDS. The size of validation set-1 is nearly two times, and validation set-2 is six times the size of the discovery set.

The discovery set comprises a homogeneous Punjabi Sikh population, the majority being the Khatri ethnic group recruited in only a small geographical region of Punjab. The validation set-1 ($n = 2986$) was relatively heterogeneous and contained mixed Punjabi communities from north India. The second validation set ($n = 9372$) comprised highly heterogeneous samples of South Asian communities, including Indian, Pakistani, Bangladeshi, and any other Asian background. The number of T2D cases in the discovery and validation dataset-1 were 52% and 58%, respectively, whereas validation set-2 had only 21% T2D cases. These differences must have resulted in a wide gap in the PRS range in the discovery
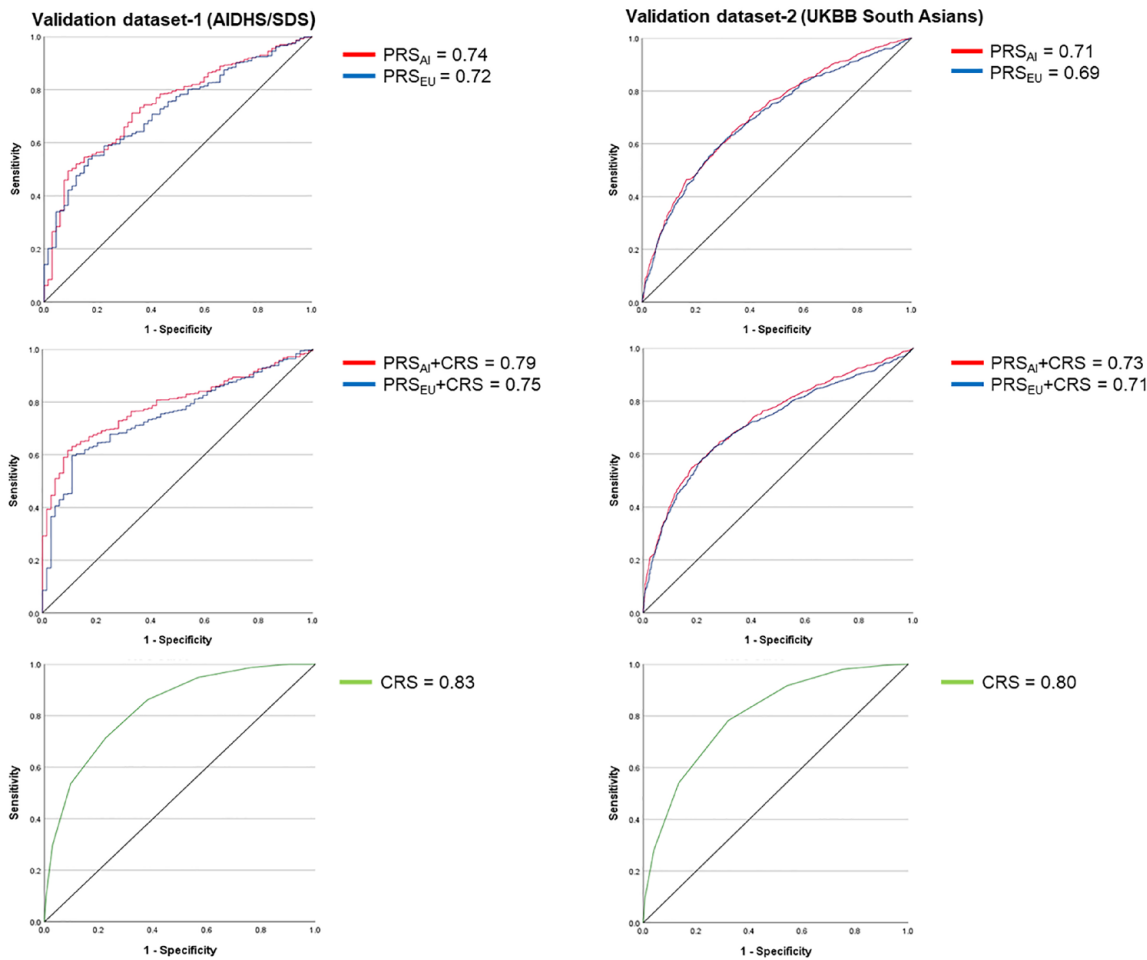
Figure 2. ROC curve showing AUC for type 2 diabetes in PRS_{AI} and PRS_{EU} with and without CRS.
AUC, area under the curve; CRS, clinical risk score; PRS_{AI}, Asian Indian ancestry-derived PRS; PRS_{EU}, European-derived PRS; ROC, receiver operating characteristic.

versus the validation datasets. Because of these differences, the ORs were wider when we analyzed the T2D risk by extreme deciles (Tables 4 and 5). Despite these differences, our PRS models captured individuals at increased risk for T2D in both ancestry-specific and European-derived PRS and suggest the strength of our models. Additionally, the ancestry-derived PRS has relatively higher transportability than the European-derived PRS even in the heterogenous population from UKBB.

clinical application of genetic and clinical risk prediction models is still underdeveloped because there is a lack of diversity in the clinical trials and genome-wide studies that remain predominantly an over-representation of European-ancestry study populations. Our results support the implications of diversity in genomic studies to improve the knowledge and utility of clinical and genomics tools for identifying and treating individuals at higher risk for developing T2D.

Conclusion
The PRS_{AI} has provided a relatively better T2D risk prediction outcome and higher transportability than the PRS_{EU}. However, both models performed equally well after integrating ethnicity-defined CRS in the genetic scores. The

Declarations

Ethics approval and consent to participate
Study protocol and consent documents were reviewed and approved by the University of Oklahoma Health Sciences Center's Institutional

Review Board (IRB#: 2911). Verbal and written consent was obtained from participants.

ORCID iD
Dharambir K. Sanghera https://orcid.org/0000-0001-7203-8655

References

1. Saeedi P, Petersohn I, Salpea P, *et al*. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: results from the International Diabetes Federation Diabetes Atlas, 9th edition. *Diabetes Res Clin Pract* 2019; 157: 107843.

2. Ranagalage M, Morimoto T, Simwanda M, *et al*. Spatial analysis of urbanization patterns in four rapidly growing South Asian cities using sentinel-2 data. *Remote Sens (Basel)* 2021; 13: 1531.

3. Trivedi JK, Sareen H and Dhyani M. Rapid urbanization – its impact on mental health: a South Asian perspective. *Indian J Psychiatry* 2008; 50: 161–165.

4. McKeigue PM, Pierpoint T, Ferrie JE, *et al*. Relationship of glucose intolerance and hyperinsulinaemia to body fat pattern in south Asians and Europeans. *Diabetologia* 1992; 35: 785–791.

5. Nakagami T, Qiao Q, Carstensen B, *et al*. Age, body mass index and type 2 diabetes-associations modified by ethnicity. *Diabetologia* 2003; 46: 1063–1070.

6. Ogurtsova K, da Rocha Fernandes JD, Huang Y, *et al*. IDF diabetes atlas: global estimates for the prevalence of diabetes for 2015 and 2040. *Diabetes Res Clin Pract* 2017; 128: 40–50.

7. Oldroyd J, Banerjee M, Heald A, *et al*. Diabetes and ethnic minorities. *Postgrad Med J* 2005; 81: 486–490.

8. Zimmet P. Type 2 (non-insulin-dependent) diabetes – an epidemiological overview. *Diabetologia* 1982; 22: 399–411.

9. Willemsen G, Ward KJ, Bell CG, *et al*. The concordance and heritability of type 2 diabetes in 34,166 twin pairs from international twin registers: the discordant twin (DISCOTWIN) consortium. *Twin Res Hum Genet* 2015; 18: 762–771.

10. Avery AR and Duncan GE. Heritability of type 2 diabetes in the Washington state twin registry. *Twin Res Hum Genet* 2019; 22: 95–98.

11. Vujkovic M, Ramdas S, Lorenz KM, *et al.* A multiancestry genome-wide association study of unexplained chronic ALT elevation as a proxy for nonalcoholic fatty liver disease with histological and radiological validation. *Nat Genet* 2022; 54: 761–771.

12. Kooner JS, Saleheen D, Sim X, *et al.* Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat Genet* 2011; 43: 984–989.

13. Saxena R, Saleheen D, Been LF, *et al.* Genome-wide association study identifies a novel locus contributing to type 2 diabetes susceptibility in Sikhs of Punjabi origin from India. *Diabetes* 2013; 62: 1746–1755.

14. Tabassum R, Chauhan G, Dwivedi OP, *et al.* Genome-wide association study for type 2 diabetes in Indians identifies a new susceptibility locus at 2q21. *Diabetes* 2013; 62: 977–986.

15. Reisberg S, Iljasenko T, Lall K, *et al.* Comparing distributions of polygenic risk scores of type 2 diabetes and coronary heart disease within different populations. *PLoS One* 2017; 12: e0179238.

16. Khera AV, Chaffin M, Aragam KG, *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* 2018; 50: 1219–1224.

17. Fritsche LG, Ma Y, Zhang D, *et al.* On cross-ancestry cancer polygenic risk scores. *PLoS Genet* 2021; 17: e1009670.

18. Weedon MN, McCarthy MI, Hitman G, *et al.* Combining information from common type 2 diabetes risk polymorphisms improves disease prediction. *PLoS Med* 2006; 3: e374.

19. Chikowore T, van Zyl T, Feskens EJ, *et al.* Predictive utility of a genetic risk score of common variants associated with type 2 diabetes in a black South African population. *Diabetes Res Clin Pract* 2016; 122: 1–8.

20. Vujkovic M, Keaton JM, Lynch JA, *et al.* Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat Genet* 2020; 52: 680–691.

21. King A, Wu L, Deng HW, *et al.* Polygenic risk score improves the accuracy of a clinical risk score for coronary artery disease. *BMC Med* 2022; 20: 385.

22. Lu X, Liu Z, Cui Q, *et al.* A polygenic risk score improves risk stratification of coronary artery disease: a large-scale prospective Chinese cohort study. *Eur Heart J* 2022; 43: 1702–1711.

23. Aulchenko YS, Struchalin MV, Belonogova NM, *et al.* Predicting human height by Victorian and genomic methods. *Eur J Hum Genet* 2009; 17: 1070–1075.

24. Li S, Zhao JH, Luan J, *et al.* Physical activity attenuates the genetic predisposition to obesity in 20,000 men and women from EPIC-Norfolk prospective population study. *PLoS Med* 2010; 7: e1000332.

25. Cheung CY, Tso AW, Cheung BM, *et al.* Obesity susceptibility genetic variants identified from recent genome-wide association studies: implications in a Chinese population. *J Clin Endocrinol Metab* 2010; 95: 1395–1403.

26. Peterson RE, Maes HH, Holmans P, *et al.* Genetic risk sum score comprised of common polygenic variation is associated with body mass index. *Hum Genet* 2011; 129: 221–230.

27. Wedekind LE, Mahajan A, Hsueh WC, *et al.* The utility of a type 2 diabetes polygenic score in addition to clinical variables for prediction of type 2 diabetes incidence in birth, youth and adult cohorts in an indigenous study population. *Diabetologia* 2023; 66: 847–860.

28. Mars N, Koskela JT, Ripatti P, *et al.* Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nat Med* 2020; 26: 549–557.

29. Anjana RM, Baskar V, Nair ATN, *et al.* Novel subgroups of type 2 diabetes and their association with microvascular outcomes in an Asian Indian population: a data-driven cluster analysis: the INSPIRED study. *BMJ Open Diabetes Res Care* 2020; 8: e001506.

30. Unnikrishnan R, Anjana RM and Mohan V. Diabetes in South Asians: is the phenotype different? *Diabetes* 2014; 63: 53–55.

31. He Y, Lakhani CM, Rasooly D, *et al.* Comparisons of polyexposure, polygenic, and clinical risk scores in risk prediction of type 2 diabetes. *Diabetes Care* 2021; 44: 935–943.

32. He Y and Patel CJ. Shared exposure liability of type 2 diabetes and other chronic conditions in the UK Biobank. *Acta Diabetol* 2022; 59: 851–860.

33. Abdullah N, Attia J, Oldmeadow C, *et al.* The architecture of risk for type 2 diabetes: understanding Asia in the context of global findings. *Int J Endocrinol* 2014; 2014: 593982.

34. Sanghera DK, Sapkota BR, Aston CE, *et al.* Vitamin D status, gender differences, and cardiometabolic health disparities. *Ann Nutr Metab* 2017; 70: 79–87.

35. Sapkota BR, Hopkins R, Bjonnes A, *et al.* Genome-wide association study of 25(OH) vitamin D concentrations in Punjabi Sikhs: results of the Asian Indian diabetic heart study. *J Steroid Biochem Mol Biol* 2016; 158: 149–156.

36. Saxena R, Bjonnes A, Prescott J, *et al.* Genome-wide association study identifies variants in casein kinase II (CSNK2A2) to be associated with leukocyte telomere length in a Punjabi Sikh diabetic cohort. *Circ Cardiovasc Genet* 2014; 7: 287–295.

37. Grarup N, Sandholt CH, Hansen T, *et al.* Genetic susceptibility to type 2 diabetes and obesity: from genome-wide association studies to rare variants and beyond. *Diabetologia* 2014; 57: 1528–1541.

38. Sanghera DK, Bhatti JS, Bhatti GK, *et al.* The Khatri Sikh Diabetes Study (SDS): study design, methodology, sample collection, and initial results. *Hum Biol* 2006; 78: 43–63.

39. Sanghera DK, Ortega L, Han S, *et al.* Impact of nine common type 2 diabetes risk polymorphisms in Asian Indian Sikhs: PPARG2 (Pro12Ala), IGF2BP2, TCF7L2 and FTO variants confer a significant risk. *BMC Med Genet* 2008; 9: 59.

40. Sanghera DK and Dodani S. Cardiovascular disease in South Asians; risk factors, genetics and environment. In GS Wander MD and KK Pareek MD (Eds), *Medicine Update 2016-1*. New Delhi, London, Philadelphia, Panama: The Health Sciences Publishers, 2016, p. 2.

41. American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes Care* 2004; 27(Suppl. 1): S5–S10.

42. Sanghera DK, Been LF, Ralhan S, *et al.* Genome-wide linkage scan to identify loci associated with type 2 diabetes and blood lipid phenotypes in the Sikh Diabetes Study. *PLoS One* 2011; 6: e21188.

43. Goyal S, Tanigawa Y, Zhang WH, *et al.* APOC3 genetic variation, serum triglycerides, and risk of coronary artery disease in Asian Indians, Europeans, and other ethnic groups. *Lipids Health Dis* 2021; 20: 113.

44. Miller SA, Dykes DD and Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* 1988; 16: 1215.

45. Das S, Forer L, Schonherr S, *et al.* Next-generation genotype imputation service and methods. *Nat Genet* 2016; 48: 1284–1287.

46. Sudlow C, Gallacher J, Allen N, *et al.* UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015; 12: e1001779.

47. Zhao W, Rasheed A, Tikkanen E, *et al.* Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease. *Nat Genet* 2017; 49: 1450–1457.

48. Loh M, Zhang W, Ng HK, *et al.* Identification of genetic effects underlying type 2 diabetes in South Asian and European populations. *Commun Biol* 2022; 5: 329.

49. Bejar CA, Goyal S, Afzal S, *et al.* A bidirectional Mendelian randomization study to evaluate the causal role of reduced blood vitamin D levels with type 2 diabetes risk in South Asians and Europeans. *Nutr J* 2021; 20: 71.

50. Willer CJ, Li Y and Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010; 26: 2190–2191.

51. O'Connor MJ, Schroeder P, Huerta-Chagoya A, *et al.* Recessive genome-wide meta-analysis illuminates genetic architecture of type 2 diabetes. *Diabetes* 2022; 71: 554–565.

52. Board JBS. Joint British Societies' consensus recommendations for the prevention of cardiovascular disease (JBS3). *Heart* 2014; 100(Suppl. 2): ii1–ii67.

53. WHO Expert Consultation. Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies. *Lancet* 2004; 363: 157–163.

54. Hanley JA and Mcneil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; 143: 29–36.

55. Weeks JP. plink: an R package for linking mixed-format tests using IRT-based methods. *J Stat Softw* 2010; 35: 1–33.

56. Martin AR, Kanai M, Kamatani Y, *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* 2019; 51: 584–591.

57. Hodgson S, Huang QQ, Sallah N, *et al.*; Genes & Health Research Team. 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* 2022; 19: e1003981.