RESEARCH ARTICLE

CLINICAL AND TRANSLATIONAL MEDICINE

WILEY

Trans-ethnic polygenic risk scores for body mass index: An international hundred K+ cohorts consortium study

Hui-Qi Qu[1] | **John J Connolly**[1] | **Peter Kraft**[2] | **Jirong Long**[3] |
Alexandre Pereira[2,4] | **Christopher Flatley**[5] | **Constance Turman**[2] | **Bram Prins**[6] |
Frank Mentch[1] | **Paulo A Lotufo**[7,8] | **Per Magnus**[9,10] | **Meir J Stampfer**[2,11,12] |
Rulla Tamimi[2] | **A Heather Eliassen**[2] | **Wei Zheng**[3] |
Gun Peggy Stromstad Knudsen[5] | **Oyvind Helgeland**[5] | **Adam S. Butterworth**[6,13,14,15] |
Hakon Hakonarson[1,16,17,18,19] | **Patrick M. Sleiman**[1,16,17] | on behalf of the IHCC
consortium

[1]The Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA

[2]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA

[3]Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, USA

[4]Department of Population Health Sciences, Weill Cornell Medicine, New York, New York, USA

[5]Division of Health Data and Digitalization, Department of Genetics and Bioinformatics, Norwegian Institute of Public Health, Oslo, Norway

[6]British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

[7]Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil

[8]Centro de Pesquisas Clínicas e Epidemiológicas, Hospital Universitário, Universidade de São Paulo, São Paulo, Brazil

[9]University of Oslo, Oslo, Norway

[10]Center for Fertility and Health, Norwegian Institute of Public Health, Oslo, Norway

[11]Department of Nutrition, Harvard T.H., Chan School of Public Health, Boston, Massachusetts, USA

[12]Channing Division of Network Medicine, Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA

[13]The National Institute for Health Research Blood and Transplant Research Unit (NIHR BTRU) in Donor Health and Genomics, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

[14]British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge, UK

[15]Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK

[16]Department of Pediatrics, The Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[17]Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA

[18]Division of Pulmonary Medicine, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA

[19]Faculty of Medicine, University of Iceland, Reykjavik, Iceland

Correspondence
Dr. Hakon Hakonarson, Center for Applied
Genomics, 3615 Civic Center Blvd, Abramson Building, Philadelphia, PA 19104,
USA.
Email: hakonarson@chop.edu

Graphical Abstract
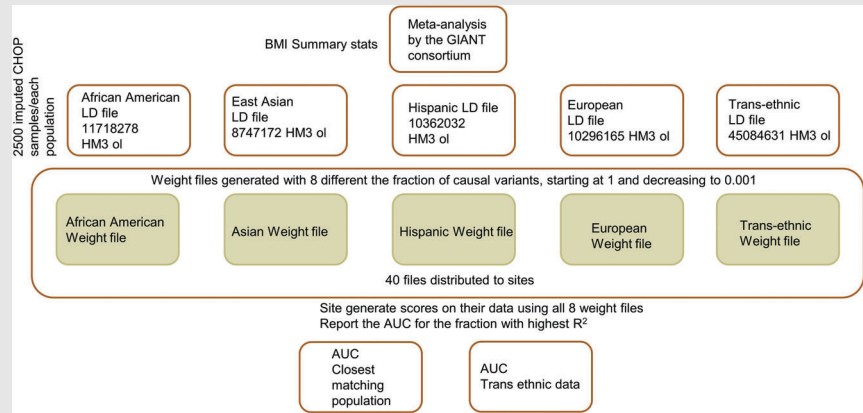
- This study is a project for the newly established global consortium, the International Hundred K+ Cohorts Consortium.
- Polygenic risk scores have significant potential to inform clinical risk, however, research efforts in minor populations are warranted to avoid health disparities.
- We present an international collaborative effort on the development of a transethnic PRS for BMI.

CLINICAL AND TRANSLATIONAL MEDICINE *Open Access* WILEY

RESEARCH ARTICLE

Trans-ethnic polygenic risk scores for body mass index: An international hundred K+ cohorts consortium study

Hui-Qi Qu[1] | John J Connolly[1] | Peter Kraft[2] | Jirong Long[3] |
Alexandre Pereira[2,4] | Christopher Flatley[5] | Constance Turman[2] | Bram Prins[6] |
Frank Mentch[1] | Paulo A Lotufo[7,8] | Per Magnus[9,10] | Meir J Stampfer[2,11,12] |
Rulla Tamimi[2] | A Heather Eliassen[2] | Wei Zheng[3] |
Gun Peggy Stromstad Knudsen[5] | Oyvind Helgeland[5] | Adam S. Butterworth[6,13,14,15] |
Hakon Hakonarson[1,16,17,18,19] | Patrick M. Sleiman[1,16,17] | on behalf of the IHCC
consortium

[1]The Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA

[2]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA

[3]Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, USA

[4]Department of Population Health Sciences, Weill Cornell Medicine, New York, New York, USA

[5]Division of Health Data and Digitalization, Department of Genetics and Bioinformatics, Norwegian Institute of Public Health, Oslo, Norway

[6]British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

[7]Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil

[8]Centro de Pesquisas Clínicas e Epidemiológicas, Hospital Universitário, Universidade de São Paulo, São Paulo, Brazil

[9]University of Oslo, Oslo, Norway

[10]Center for Fertility and Health, Norwegian Institute of Public Health, Oslo, Norway

[11]Department of Nutrition, Harvard T.H., Chan School of Public Health, Boston, Massachusetts, USA

[12]Channing Division of Network Medicine, Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA

[13]The National Institute for Health Research Blood and Transplant Research Unit (NIHR BTRU) in Donor Health and Genomics, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

[14]British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge, UK

[15]Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK

[16]Department of Pediatrics, The Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[17]Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA

[18]Division of Pulmonary Medicine, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA

[19]Faculty of Medicine, University of Iceland, Reykjavik, Iceland

Correspondence
Dr. Hakon Hakonarson, Center for Applied Genomics, 3615 Civic Center Blvd, Abramson Building, Philadelphia, PA 19104, USA.

Abstract

Background: While polygenic risk scores hold significant promise in estimating an individual's risk of developing a complex trait such as obesity, their

application in the clinic has, to date, been limited by a lack of data from non-European populations. As a collaboration model of the International Hundred K+ Cohorts Consortium (IHCC), we endeavored to develop a globally applicable trans-ethnic PRS for body mass index (BMI) through this relatively new international effort.

Methods: The polygenic risk score (PRS) model was developed, trained and tested at the Center for Applied Genomics (CAG) of The Children's Hospital of Philadelphia (CHOP) based on a BMI meta-analysis from the GIANT consortium. The validated PRS models were subsequently disseminated to the participating sites. Scores were generated by each site locally on their cohorts and summary statistics returned to CAG for final analysis.

Results: We show that in the absence of a well powered trans-ethnic GWAS from which to derive marker SNPs and effect estimates for PRS, trans-ethnic scores can be generated from European ancestry GWAS using Bayesian approaches such as LDpred, by adjusting the summary statistics using trans-ethnic linkage disequilibrium reference panels. The ported trans-ethnic scores outperform population specific-PRS across all non-European ancestry populations investigated including East Asians and three-way admixed Brazilian cohort.

Conclusions: Here we show that for a truly polygenic trait such as BMI adjusting the summary statistics of a well powered European ancestry study using trans-ethnic LD reference results in a score that is predictive across a range of ancestries including East Asians and three-way admixed Brazilians.

KEYWORDS
body mass index, obesity, polygenic risk score, population admixture, trans-ethnic

1 | INTRODUCTION

Obesity is a global health issue,[1] with an adult prevalence of about 13% across the world (https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight). As a difficult condition to treat, obesity prevention is important

which has led us to develop a trans-ethnic polygenic risk score (PRS) for body mass index (BMI) through the International HundredK+ Cohorts Consortium (IHCC).[2] PRS aggregates the effects of many genetic variants across the human genome into a single score, which may effectively improve the prediction of a complex disease/trait and assist

the differential diagnosis.[3] BMI-associated variants have been under natural selection, and explain 30−40% variance for BMI.[4] Previous study has shown that PRS is an important determinant of BMI across life.[5] In addition to obesity prevention, PRS for BMI may have clinical applications, for example, the prediction of cardiometabolic health.[6] Currently, like other health outcomes, there is a serious lacking of genomic information for BMI in minor populations. The development of PRS for BMI in minorities warrants for research efforts to avoid health disparities.

Our obesity PRS was based on the published GWAS meta-analysis of BMI that included 339 224 individuals of European ancestry. The study identified 97 genome-wide significant BMI-associated loci that account for approximately 2.7% of BMI variation alone. Genome-wide estimates suggest that common variation accounts for over 20% of variation in BMI. Various approaches for PRS calculation have been developed to date.[7–9] The standard approaches for calculating risk scores involve linkage disequilibrium (LD)-based marker pruning followed by p-value thresholding of GWAS-based association statistics. While effective, these approaches lose information and can reduce predictive accuracy particularly where the test population differs in genetic ancestry from the GWAS sample. Bayesian approaches such LDpred,[7] a method that infers the posterior mean effect size of each marker by using a prior on effect sizes and LD information from an external reference panel may therefore improve prediction accuracy in multi-ethnic studies of diverse populations.[7] As most large scale GWAS have been conducted using only individuals of European ancestry there is a need to develop approaches that can port PRS using European ancestry derived effect estimates. More importantly, the different prevalences of BMI and obesity across human populations are closely related to environmental factors and cultural diversity.[10,11] Therefore, it is essential to validate a multi-ethnic PRS in different regional populations, especially admixed populations. Leveraging an international effort supported by the IHCC that has brought together large scale cohorts with genotyping data from around the world, we explored the development of an LDpred-based, trans-ethnic (TE) obesity PRS through a collaboration of 6 international research centers (Figure 1).

2 | METHODS

2.1 | Model training at the Children's Hospital of Philadelphia (CHOP)

The PRS model development and training was carried out at the Center for Applied Genomics (CAG) according to

the pipeline shown in Figure 1. SNP weights (i.e. posterior mean effect sizes) were calculated using Markov chain Monte Carlo (MCMC) Gibbs sampling as implemented in LDpred.[7] The summary statistics of genetic association with BMI were based on the meta-analysis of genome-wide association studies by the GIANT consortium.[12] We restricted the variants to SNPs included in the HapMap3 data. Five sets of SNP weights were generated based on the respective LD patterns from the following populations, African American (AA), Hispanic American (HAMR), East Asian (ASN), Northern European (EUR) and all of the above populations (i.e. trans-ethnic). For each group we selected 2500 CAG participants who clustered with the 1000 genome project reference data[13] to generate the LD files. The trans-ethnic group included all populations tested. Each group of SNP weights includes 8 different sets corresponding, respectively, to the mixture probability values (i.e. fractions of causal markers with non-zero effects used in the Gibbs sampler) of [infinitesimal prior (LDpred-inf), 1, 0.3, 0.1, 0.03, 0.01, 0.003, and 0.001].

2.2 | Initial validation of PRS models in-house

The initial assessment of the PRS models was based on the genomic data in-house at CAG, CHOP. The pediatric biobank built at CAG has archived samples from 500 000 children from USA, Europe, South America, Canada, Saudi Arabia and Australia.[14] The definition of normal BMI for children is age and sex specific.[15] Instead of developing a TE PRS for adult BMI, this study aimed for a binary TE PRS model for extreme BMI, that is, the top 1% and 5% BMI. These percentiles are in line with the clinical definition of childhood obesity, that is, childhood overweight as BMI ≥95th percentile,[16] and severe obesity as BMI ≥99th percentile.[17] For the initial validation of the trans-ethnic PRS model, 57 613 randomly selected individuals (51% males and 49% females) including individuals of European ancestry, African Americans, Hispanics/Latinos, and East / South Asians in order of frequency, with genotypes and BMI data from the CAG biobank were used for the validation. The principal component analysis (PCA) plot of the population structure is shown in Figure 1. All the individuals have been genotyped with an Illumina Genotyping BeadChip with at least 550 000 SNPs genotyped. Genome-wide imputation was done using the TOPMed Imputation Server (https://imputation.biodatacatalyst.nhlbi.nih.gov) with the TOPMed (Version R2 on GRCh38) Reference Panel. Population ancestries of the research subjects were confirmed by principal component analysis (PCA) with genomic DNA markers, compared with the reference populations in the 1000 Genomes project.[13]
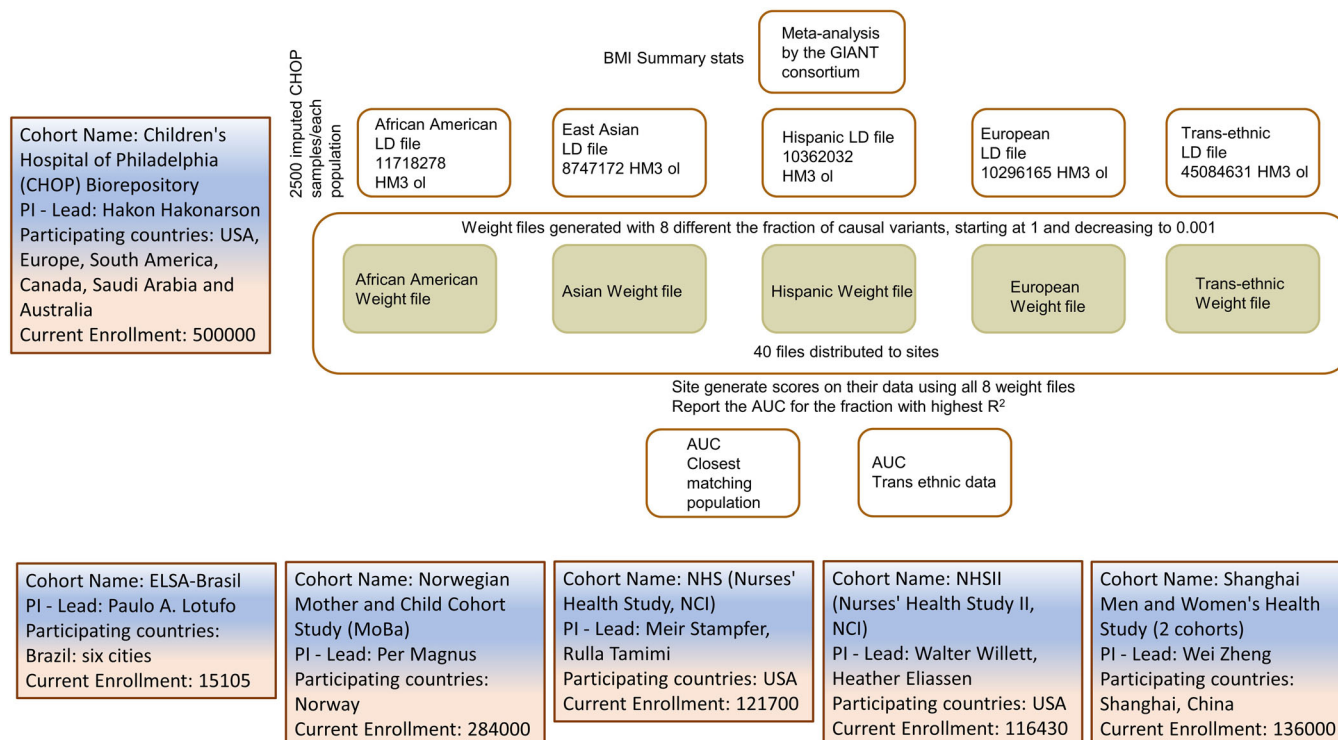
FIGURE 1 Development of a trans-ethnic polygenic risk score for body mass index.

Harmonization of SNP alleles in the PRS model was confirmed by comparing with the reference alleles of the TOPMed imputation.

2.3 | Validation of PRS models in regional populations

Having validated the trans-ethnic PRS models, we shared the protocol of the PRS models, as well as the SNPs and weights with the IHCC collaborators for assessment in 7 different cohorts of regional populations. The population sites included a three-way admixed Brazilian cohort from ELSA-Brasil,[18,19] the Norwegian Mother, Father and Child Cohort Study (MoBa, conducted by the Norwegian Institute of Public Health), two US based studies The Nurses' Health Study (NHS) and Nurses' Health Study II (NHSII), the UK based INTERVAL BioResource and two Chinese population samples from the Shanghai Men's Health Study (SMHS) and Shanghai Women's Health Study (SWHS)[20,21](Figure 1).

Each site generated the PRS on their cohort following the same protocol and using the same tools and the software packages. To compare the performance of trans-ethnic PRS vs. that of population-specific PRS in different regional populations, we requested each collaboration group run two calculations if possible, that is, one PRS calculation for the specific population that is closest to

their dataset, and one PRS for the trans-ethnic score. To do the dichotomous receiver operating characteristic (ROC) curve analysis, the top 5% and 1% of the BMI distribution within each study were defined as cases.

2.4 | Participating IHCC cohorts

The **Nurses' Health Study I** recruited 121 700 married registered nurses in 1976. Blood samples were collected from 33 000 participants in 1989−90, and cheek cells from another 33 000 in 2001−4. Genome-wide association data are available on over 17 000 participants as part of studies of multiple complex diseases and traits, including breast cancer, type 2 diabetes, venous thromboembolism, and depression. All women included in these analyses are of European ancestry (cluster with European reference samples and do not self-report as other than European ancestry).

The **Nurses' Health Study II** recruited 116 430 married registered nurses between the ages of 25 and 42 in 1989. Blood samples were collected from 29 000 participants in 1996−99, and cheek cells from another 30 000 in 2006. Genome-wide association data are available on over 12 000 participants as part of studies of multiple complex diseases and traits, including breast cancer, type 2 diabetes, venous thromboembolism, and depression. All women included in these analyses are of European ancestry (cluster with

European reference samples and do not self-report as other than European ancestry).

Body mass index was self-reported at time of blood draw or cheek cell collection. Diabetes cases were defined as self-reported diabetes confirmed by a validated supplementary questionnaire.

The **Shanghai Women's Health Study** (**SWHS**) is a large population-based prospective cohort study initiated in 1996.[20] Approximately 75 000 Chinese women who lived in Shanghai were recruited into the study. In addition to survey data, blood and urine samples were collected from most study participants at the baseline recruitment.

The **Shanghai Men's Health Study** (**SMHS**) is a population-based cohort study of 61 480 Chinese men between ages 40 and 74 who lived in eight urban communities in Shanghai at enrollment (2002–2006).[22] Detailed information on dietary and other lifestyle factors was collected at baseline and updated in follow-up surveys. Biological samples (blood, and or urine) were collected from 89% of cohort members.

The **Norwegian Mother, Father and Child Cohort Study (MoBA)** was established and is conducted by the Norwegian Institute of Public Health (NIPH). MoBa is an ambitious family-oriented cohort study that aims to find causes of diseases and explain trajectories and variability of health-related traits over a life-course span. Between 1999 and 2008, pregnant women were invited to take part in the study around the time of the ultrasound examination in week 17−20 of gestation. The fathers of the children were also invited to participate. Biological material has been collected from mothers, fathers and children and has been stored in a biobank. Self-reported data are collected from regular questionnaires about general health, diet and environmental exposure. The cohort includes approximately 109 000 children, 91 000 women and 71 700 men. 50 290 Northern European adult males and females were analysed in this study, with mean BMI = 24.96(SD = 3.9).

The **Brazilian Longitudinal Study of Adult Health (ELSA-Brasil)** enrolled 15 105 civil servants aged 35 to 74 years-old living in six cities,[18] addressing the incidence of non-communicable diseases. From the 15 105 participants, 9333 DNA samples were analyzed for genetic ancestry using a software tool for maximum likelihood estimation of individual ancestries from multilocus SNP genotype datasets.[19]

The **INTERVAL BioResource** recruited 45 263 whole blood donors (22 466 men and 22 797 women) between 11 June 2012 and 15 June 2014.[23] Donors were aged 18 years or older from 25 NHS Blood and Transplant (NSHBT) blood donation centres distributed across England, UK. Donors provided blood samples at baseline to enable DNA extraction and self-reported their height and weight for estimation of BMI. Genotyping was conducted using the

TABLE 1 BMI PRS test by the NHS and NHSII cohorts.*

Top 1% BMI	EUR-NHS1	TE-NHS1
NHS	0.730	0.727
NHSII	0.736	0.735
Top 5% BMI		
NHS	0.703	0.701
NHSII	0.721	0.718

*NHS: the Nurses' Health Study. Fraction of causal variants = 0.03.

Affymetrix Axiom UK Biobank array with imputation using a combined 1000 Genomes Phase 3/UK10K reference panel. A total of 38 319 European ancestry adult participants were included in the final analyses

3 | RESULTS

3.1 | Initial validation of PRS models in-house

The performance of the trans-ethnic PRS model tested in-house is shown in Figure 2. As shown by our analysis, AUC > 0.720 to predict top 1% BMI was achieved in all the PRS models with mixture probability≥0.03. The highest AUC is seen at the mixture probabilities of 0.03–0.1.

3.2 | Evaluation of the PRS in the four European ancestry cohorts

The variance explained by the PRS in European was 0.095 by the EUR PRS, and 0.092 by the TE PRS. Across all four European ancestry cohorts both European-specific PRS models and the trans-ethnic PRS models had the largest AUCs with a mixture probability of 0.03. Across the three cohorts where both the European-specific PRS model and the trans-ethnic PRS model were run (INTERVAL, NHS and NHSII), the performance of the scores was comparable (Table 1).

Only the European-specific PRS models were tested in the Norwegian MoBa study. The PRS model showed AUC≥0.686 to predict top 5% BMI, and AUC≥0.735 to predict top 1% BMI, in the models with mixture probability ≥0.03 (Table 2).

3.3 | Validation of the trans-ethnic PRS models in regional populations

By the Chinese population in the SMHS and SWHS cohorts, the PRS models were tested in 866 men in the SMHS cohort and 3120 women in the SWHS cohort and
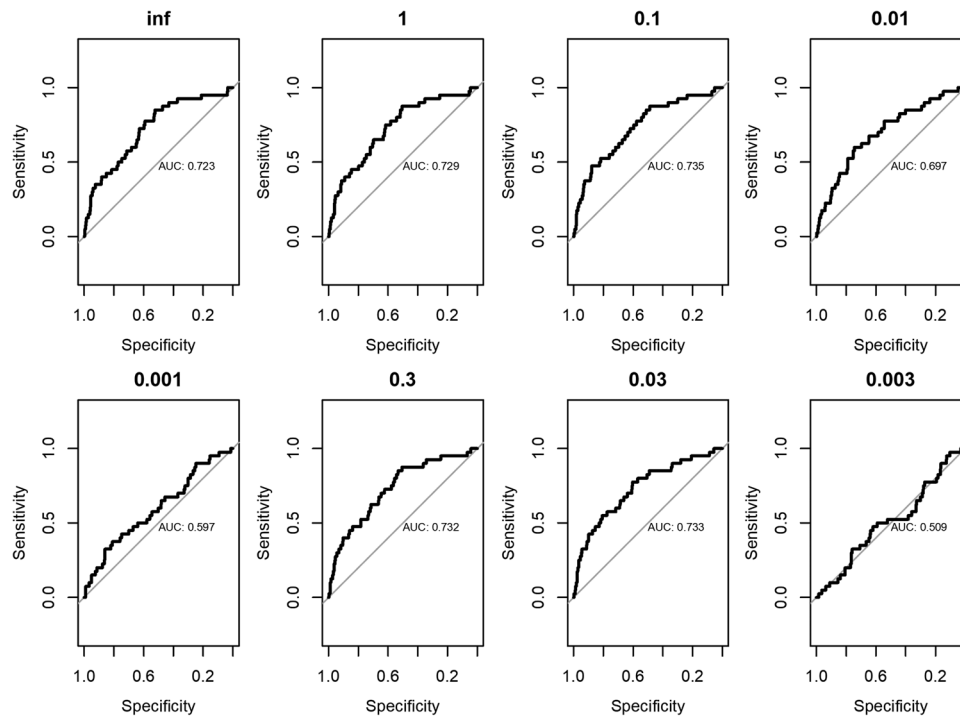
FIGURE 2 AUC values following the initial validation of PRS models in-house. 57 613 randomly selected individuals with genotypes and BMI data from the the Center for Applied Genomics biobank at the Children's Hospital of Philadelphia were used for the validation.

TABLE 2 BMI EUR PRS test by the Norwegian Mother and Child Cohort Study (MoBA).

MoBa	Top 5% BMI	Top 1% BMI
Inf*	0.686	0.735
1	0.689	0.738
0.3	0.694	0.745
0.1	0.699	0.751
0.03	0.704	0.756
0.01	0.658	0.7
0.003	0.594	0.626
0.001	0.552	0.565

*Infinitesimal prior.

the combination of both. Across all three analyses the trans-ethnic (TE) score outperformed the ancestry specific East-Asian reference with a TE AUC of 0.76 compared with 0.479 for the ASN reference in men, TE AUC of 0.737 vs ASN AUC of 0.719 in women, and a TE AUC of 0.717 vs ASN AUC of 0.706 in the combined dataset at the most informative fraction (Table 3).

The Brazilian population in the ELSA-Brasil cohort is three-way admixed. We therefore compared the trans-ethnic score in this population to each of the three founders: American, African and European. The trans-ethnic score outperformed all three founder population

ancestry-specific scores at the most informative fraction of 0.03 at both 95th and 99th percentiles (Table 4).

4 | DISCUSSION

Clinically, the trans-ethnic PRS may assist the prediction of dynamic BMI for primary prevention of overweight. A prediction model with AUC = 0.845 developed by Welten et al.[24] took into account of the predictors including maternal BMI, paternal BMI, as well as birthweight, sex, and a number of socioeconomic and environmental factors. The PRS can effectively address the impreciseness of inheritance information represented by maternal and paternal BMI as there is 50% chance of inheritance for each parental allele. In order for polygenic risk scores to achieve their clinical potential, and avoid exacerbating health disparities due to the lack of genomic information in minorities, they have to be universally applicable regardless of a patient's genetic ancestry. Ideally, trans-ethnic PRS would be generated from true trans-ethnic GWAS; however, well powered trans-ethnic studies remain the exception for the majority of traits and phenotypes. In this study we evaluated the performance of a BMI PRS that is based on European ancestry GWAS effect sizes combined with trans-ethnic LD patterns using a Bayesian approach as implemented in LDpred.[7] In a federated model, we developed the PRS score at CAG and disseminated the standard operating procedure along with

TABLE 3 BMI PRS test by the Shanghai Men and Women's Health Study.

Cohorts Fractions of causal variants	Racial groups of weights	SMHS (N = 866)			SWHS (N = 3120)			SWMHS (N = 3986)		
		P*	AUC	P #	P*	AUC	P #	P*	AUC	P #
INF	ASN	0.07	0.667	1.12E-10	6.61E-06	0.725	5.21E-33	1.91E-06	0.706	4.66E-42
	TE	0.044861	0.684	3.03E-11	4.88E-06	0.732	1.80E-32	1.15E-06	0.717	5.41E-42
1	ASN	0.05	0.682	1.99E-11	1.71E-05	0.714	3.92E-33	7.66E-06	0.694	7.74E-43
	TE	0.040165	0.684	1.91E-11	5.09E-06	0.728	1.32E-33	1.78E-06	0.709	2.43E-43
0.3	ASN	0.03	0.707	3.40E-12	1.10E-05	0.719	2.85E-34	4.90E-06	0.699	1.13E-44
	TE	0.030129	0.692	4.93E-12	2.54E-06	0.737	1.65E-35	1.16E-06	0.713	8.18E-46
0.1	ASN	0.02	0.721	2.10E-13	8.01E-06	0.722	1.83E-35	3.55E-06	0.702	6.26E-47
	TE	0.017299	0.716	2.96E-13	2.15E-06	0.735	8.27E-37	1.21E-06	0.711	3.32E-48
0.03	ASN	0.03	0.719	8.04E-15	3.62E-05	0.703	2.24E-35	1.94E-05	0.684	5.40E-48
	TE	0.015892	0.733	1.10E-14	1.34E-05	0.714	3.65E-37	8.83E-06	0.694	9.08E-50
0.01	ASN	0.22	0.65	0.04	0.98	0.515	0.08	0.82	0.497	0.01
	TE	0.11957	0.647	0.96	0.57	0.533	0.15	0.88	0.516	0.19
0.003	ASN	0.21	0.622	0.41	0.91	0.517	0.36	0.35	0.471	0.64
	TE	0.73	0.527	0.28	0.66	0.541	0.15	0.35	0.561	0.08
0.001	ASN	0.87	0.479	0.91	0.27	0.568	0.2	0.29	0.569	0.22
	TE	0.017789	0.76	0.13	0.66	0.534	0.23	0.42	0.551	0.08

*Logistic regression analyses were performed after coding top 1% distribution of BMI as cases and the remainings as controls.

#Linear regression analyses were performed using BMI as continuous variable.

the SNPs and weights files and the population specific LD matrices to participant sites within the IHCC. This model allowed us to quickly test the hypothesis in various world populations without the need for data transfer and hence time-consuming data sharing agreements. By providing detailed protocols and all required files to run the scores, we made efforts to minimize the work load on each collaboration group while enabling direct comparisons between groups. LDpred has been demonstrated of comparable performance in BMI PRS to pruning and thresholding (P + T) approaches, for example, PRSice-2.[25] In particular, LDpred applies the LD information from a population-specific reference panel. For the purpose of testing a TE PRS score, LDpred allows the comparison of TE score with the score by a population-specific reference panel.

In the TE PRS model, the gene enrichment analysis using the WEB-based Gene Set Analysis Toolkit (WebGestalt) [7] based on the Molecular Signatures Database (MSigDB) hallmark gene set collection[26] showed that the SNPs with absolute(β) ≥2.0E-04 are enriched in the gene sets HALLMARK_UV_RESPONSE_DN and HALLMARK_ESTROGEN_RESPONSE_EARLY (False discovery rate < 0.05, Data, Supporting Information). It is worth to mention that, in contrast to the LDpred[7] used in this study, a new version of the method LDpred2 has been released.[27] LDpred2 made a significant effort to address the potential bias of Gibbs sampling in the human leukocyte antigen (HLA) region at chromosome 6 with extended

LD.[27] In contrast, the HLA region has been removed in the LDpred modeling. The HLA region is highly polymorphic, with highly diverse frequencies across different populations, as well as extended and strong LD due to significant evolutionary selection pressure in human populations.[28,29] Including the HLA region will cause significant difference of PRS across different ethnicities, as we have observed in the trans-ethnic scoring of an autoimmune disease, type 1 diabetes (T1D), with HLA as a major risk factor.[30] In addition, there has been no GWAS study to date suggesting the role of the HLA region in BMI or obesity. Nevertheless, it is interesting to leverage the IHCC resources to examine the performance of LDpred2 in the trans-ethnic scoring of BMI.

The performance of the trans-ethnic PRS in different cohorts is comparable to the published PRS models for the prediction of obesity in populations with European ancestries, for example, AUC = 0.708 in the European-ancestry participants of the UK Biobank,[31] AUC = 0.619 to 0.704 in the Quebec Family Study.[32] In our study, LDpred outperformed LDpred-inf with the fractions of causal markers (1, 0.3, 0.1, 0.03), but not with the other fractions (0.01, 0.003, and 0.001) lower than the above. As the results show, our trans-ethnic PRS models outperformed the ancestry specific models in the non-European populations tested including Chinese and Brazilian. Importantly the UK and US data from the INTERVAL and NHS cohorts demonstrate that there is no appreciable loss in

TABLE 4 BMI PRS test by the Brazilian population in the ELSA-Brasil cohort.

Fractions of causal variants	Top BMI	AFR	AMR	EUR	TE
Inf*	1%	0.569	0.575	0.548	0.589
	5%	0.575	0.553	0.538	0.564
1	1%	0.604	0.582	0.558	0.595
	5%	0.575	0.554	0.542	0.569
0.1	1%	0.635	0.63	0.579	0.616
	5%	0.581	0.559	0.544	0.57
0.3	1%	0.612	0.587	0.563	0.602
	5%	0.575	0.554	0.543	0.566
0.03	1%	0.558	0.636	0.606	0.645
	5%	0.517	0.571	0.553	0.579
0.01	1%	0.472	0.56	0.634	0.538
	5%	0.546	0.526	0.573	0.511
0.003	1%	0.524	0.615	0.525	0.535
	5%	0.522	0.547	0.497	0.521
0.001	1%	0.475	0.585	0.538	0.448
	5%	0.519	0.521	0.508	0.492

*Infinitesimal prior.

Abbreviations: AFR: African; AMR: American Indian; ELSA-Brasil: The Brazilian Longitudinal Study of Adult Health; EUR: European; TE: trans-ethnic.

predictive power in European ancestry individuals when using a trans-ethnic score. As such, and in the absence of trans-ethnic effect sizes from diverse GWAS, we propose that using trans-ethnic reference data to adjust the summary statistics for the effects of LD patterns improves the performance of PRS in populations that have not been included in the generation of the summary stats. On the other hand, in the absence of reference panels for different populations, the genome-wide LD scores and matrices from the Pan-UK Biobank resource, or calculated by genome sequencing data of the Genome Aggregation Database (gnomAD), may provide sufficient resolution.

As the first effort by the IHCC to leverage the existing datasets that reside within this large scale consortium for a trans-ethnic PRS on BMI, we envision an opportunity to scale this to other cohorts within the consortium and expand the number of traits that can be analyzed. As such, the IHCC presents a rich resource of data for collaborative research with trans-ethnic focus, where there is much unmet need at the present time and an area of research that has been largely ignored. In this study, the performance of the trans-ethnic PRS model is relatively poorer in the Brazilian cohort. The prevalence of obesity is high in the Brazilian population.[33] In addition to genetic heterogeneity of the admixed population, the underperformed PRS may be also due to uncounted environmental and socioeconomic factors in this population.[34] The current study is a proof-of-principle cross-network pilot for

the IHCC consortium in demonstrating feasibility across a condensed timeline, and needs benchmarking. The strategy presented in the current study was to develop and share a technical protocol easily applicable to different sites. In the meantime, a number of large-scale GWAS studies have been published in other populations, for example, East Asian populations.[35–37] An updated meta-analysis with the GWASs in other populations may improve the current PRS model. However, to redo the meta-analysis will need access to individual data to redo the genotype imputation. To address data sharing barriers across different international research centers warrants for more extensive research efforts.

CONFLICT OF INTEREST STATEMENT
A.S.B declares grants outside of this work from AstraZeneca, Bayer, Biogen, BioMarin, Merck and Sanofi. All other authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CONSENT FOR PUBLICATION
All authors have provided consent for publication of the manuscript.

ORCID

Hui-Qi Qu https://orcid.org/0000-0001-9317-4488

Hakon Hakonarson https://orcid.org/0000-0003-2814-7461

REFERENCES

1. Caballero B. The global epidemic of obesity: an overview. *Epidemiol Rev.* 2007;29(1):1-5.

2. Manolio TA, Goodhand P, Ginsburg G. The International Hundred Thousand Plus Cohort Consortium: integrating large-scale cohorts to address global scientific challenges. *The Lancet Digital Health.* 2020;2(11):e567-e568. doi:10.1016/s2589-7500(20)30242-9

3. Lambert SA, Abraham G, Inouye M. Towards clinical utility of polygenic risk scores. *Hum Mol Genet.* 2019;28(R2):R133-R142. doi:10.1093/hmg/ddz187

4. Yang J, Bakshi A, Zhu Z, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet.* 2015;47(10):1114-1120. doi:10.1038/ng.3390

5. Bann D, Wright L, Hardy R, Williams DM, Davies NM. Polygenic and socioeconomic risk for high body mass index: 69 years of follow-up across life. *PLoS Genet.* 2022;18(7):e1010233. doi:10.1371/journal.pgen.1010233

6. Shi M, Chen W, Sun X, et al. Association of genome-wide polygenic risk score for body mass index with cardiometabolic health from childhood through midlife. *Circ Genomic Precision Med.* 2022;15(4):e003375. doi:10.1161/circgen.121.003375

7. Vilhjálmsson BJ, Yang J, Finucane HK, et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am J Hum Genet.* 2015;97(4):576-592.

8. Euesden J, Lewis CM, PF O'reilly. PRSice: polygenic risk score software. *Bioinformatics.* 2015;31(9):1466-1468.

9. Ge T, Chen C-Y, Ni Y, Feng Y-CA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat Commun.* 2019;10(1):1776.

10. Barness LA, Opitz JM, Gilbert-Barness E. Obesity: genetic, molecular, and environmental aspects. *Am J Med Genet Part A.* 2007;143(24):3016-3034.

11. Albuquerque D, Stice E, Rodríguez-López R, Manco L, Nóbrega C. Current review of genetics of human obesity: from molecular mechanisms to an evolutionary perspective. *Mol Genet Genomics.* 2015;290:1191-1221.

12. Locke AE, Kahali B, Berndt SI, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature.* 2015;518(7538):197-206. doi:10.1038/nature14177

13. Siva N. 1000 Genomes project. *Nat Biotechnol.* 2008;26(3):256-257.

14. Connolly JJ, Glessner JT, Li D, Sleiman PM, Hakonarson H. The Center for Applied Genomics at The Children's Hospital of Philadelphia–Pediatric Perspectives on Genomic Medicine adminpmls2016 2021-03-04T16: 43: 43-05: 00 (2020).

15. Reilly JJ. Diagnostic accuracy of the BMI for age in paediatrics. *Int J Obes.* 2006;30(4):595-597.

16. Dietz WH, Robinson TN. Overweight children and adolescents. *N Engl J Med.* 2005;352(20):2100-2109.

17. Skelton JA, Cook SR, Auinger P, Klein JD, Barlow SE. Prevalence and trends of severe obesity among US children and adolescents. *Acad Pediatr.* 2009;9(5):322-329.

18. Aquino EM, Barreto SM, Bensenor IM, et al. Brazilian longitudinal study of adult health (ELSA-Brasil): objectives and design. *Am J Epidemiol.* 2012;175(4):315-324.

19. Chor D, Pereira A, Pacheco AG, et al. Context-dependence of race self-classification: results from a highly mixed and unequal middle-income country. *PLoS One.* 2019;14(5):e0216653. doi:10.1371/journal.pone.0216653

20. Zheng W, Chow W-H, Yang G, et al. The Shanghai Women's Health Study: rationale, study design, and baseline characteristics. *Am J Epidemiol.* 2005;162(11):1123-1131.

21. Shu XO, Li H, Yang G, et al. Cohort Profile: the Shanghai Men's Health Study. *Int J Epidemiol.* 2015;44(3):810-818. 10.1093/ije/dyv013

22. Shu X-O, Li H, Yang G, et al. Cohort profile: the Shanghai men's health study. *Int J Epidemiol.* 2015;44(3):810-818.

23. Di Angelantonio E, Thompson SG, Kaptoge S, et al. Efficiency and safety of varying the frequency of whole blood donation (INTERVAL): a randomised trial of 45 000 donors. *Lancet.* 2017;390(10110):2360-2371. doi:10.1016/s0140-6736(17)31928-1

24. Welten M, Wijga AH, Hamoen M, et al. Dynamic prediction model to identify young children at high risk of future overweight: development and internal validation in a cohort study. *Pediatr Obes.* 2020;15(9):e12647.

25. Choi SW, O'Reilly PF. PRSice-2: polygenic risk score software for biobank-scale data. *GigaScience.* 2019;8(7):giz082. doi:10.1093/gigascience/giz082

26. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* 2015;1(6):417-425. doi:10.1016/j.cels.2015.12.004

27. Privé F, Arbel J, Vilhjálmsson BJ. LDpred2: better, faster, stronger. *Bioinformatics.* 2020;36(22-23):5424-3541. doi:10.1093/bioinformatics/btaa1029

28. Prugnolle F, Manica A, Charpentier M, Guégan JF, Guernier V, Balloux F. Pathogen-driven selection and worldwide HLA class I diversity. *Curr Biol.* 2005;15(11):1022-1027.

29. Meyer D, Aguiar VRC, Bitarello BD, Brandt DYC, Nunes K. A genomic perspective on HLA evolution. *Immunogenetics.* 2018;70(1):5-27. doi:10.1007/s00251-017-1017-3

30. Qu HQ, Qu J, Glessner J, et al. Improved genetic risk scoring algorithm for type 1 diabetes prediction. *Pediatr Diabetes.* 2022;23(3):320-323.

31. Wang Z, Choi SW, Chami N, et al. The value of rare genetic variation in the prediction of common obesity in european ancestry populations. *Front Endocrinol.* 2022;13:863893. doi:10.3389/fendo.2022.863893

32. de Toro-Martín J, Guénard F, Bouchard C, Tremblay A, Pérusse L, Vohl M-C. The challenge of stratifying obesity: attempts in the Quebec family study. *Front Genet.* 2019;10:994.

33. Filozof C, Gonzalez C, Sereday M, Mazza C, Braguinsky J. Obesity prevalence and trends in Latin-American countries. *Obes Rev.* 2001;2(2):99-106.

34. Jaime PC, Duran AC, Sarti FM, Lock K. Investigating environmental determinants of diet, physical activity, and overweight among adults in Sao Paulo, Brazil. *J Urban Health.* 2011;88:567-581.

35. Zhu Z, Li J, Si J, et al. A large-scale genome-wide association analysis of lung function in the Chinese population identifies

novel loci and highlights shared genetic aetiology with obesity. *Eur Respir J.* 2021;58(4). doi:10.1183/13993003.00199-2021

36. Akiyama M, Okada Y, Kanai M, et al. Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. *Nat Genet.* 2017;49(10):1458-1467. doi:10.1038/ng.3951

37. Nam K, Kim J, Lee S. Genome-wide study on 72,298 individuals in Korean biobank data for 76 traits. *Cell Genomics.* 2022;2(10):100189. doi:10.1016/j.xgen.2022.100189

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.