HGG Advances

ARTICLE

Polygenic risk scores and kidney traits in the Hispanic/Latino population: The Hispanic Community Health Study/Study of Latinos

Laura Y. Zhou,[1,8,*] Tamar Sofer,[2,3] Andrea R.V.R. Horimoto,[4] Gregory A. Talavera,[5] James P. Lash,[6] Jianwen Cai,[1] and Nora Franceschini[7]

Summary

Estimated glomerular filtration rate (eGFR) is used to evaluate kidney function and determine the presence of chronic kidney disease (CKD), a highly prevalent disease in the US[1–3] that varies among subgroups of Hispanic/Latino individuals.[4,5] The polygenic risk score (PRS) is a popular method that uses large genome-wide association studies (GWASs) to provide a strong estimate of disease risk.[7] However, due to the limited availability of summary statistics from GWAS meta-analyses based on Hispanic/Latino populations, PRSs can only be computed using different ancestry GWASs. The performance of eGFR PRSs derived from other GWAS reference populations for Hispanic/Latino population has not been examined. We compared PRS constructions for eGFR prediction in Hispanic/Latino individuals using GWAS-significant variants, clumping and thresholding (C&T),[8] and PRS-CS,[22] as well as a combination of PRSs calculated with different reference GWAS meta-analyses from European and multi-ethnic studies in Hispanic/Latino individuals from the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). All eGFR PRSs were highly associated with eGFR ($p < 1E-20$). Additionally, eGFR PRSs were significantly associated with lower risk of prevalent CKD at visit 1 or 2 and incident CKD at visit 2, with the combined PRSs having the best performance. These PRS findings were replicated in an additional dataset of Hispanic/Latino individuals using data from the Women's Health Initiative SNP Health Association Resource (WHI-SHARe).[17]

Introduction

Chronic kidney disease (CKD) is highly prevalent in the US (15%),[1–3] and is a cause and consequence of hypertension. Both CKD and hypertension prevalence vary among different subgroups of Hispanic/Latino individuals.[4,5] Estimated glomerular filtration rate (eGFR) is used to evaluate kidney function and to determine the presence of CKD. Genome-wide association studies (GWASs) have identified over 200 genetic variants associated with eGFR.[6] These GWASs were performed over populations of European ancestries or over multi-ethnic studies. Although each GWAS-identified variant has a small effect on eGFR, and therefore are not useful by themselves for quantifying disease risk, the combination of risk alleles together with their estimated risk via polygenic risk scores (PRSs) can provide a strong estimate of disease risk.[7]

Summary statistics from large GWAS meta-analyses based on Hispanic/Latino populations that can be used for obtaining risk estimates and construct Hispanic/Latino-specific PRSs are currently unavailable. Thus, to develop PRSs for the Hispanic/Latino population, we must use summary statistics from GWAS of other populations. However, PRSs derived from one population may not be generalizable to other populations. In order to examine the performance of PRSs for eGFR or CKD derived from other populations in the Hispanic/Latino population, we compared several PRS constructions for eGFR prediction in Hispanic/Latino individuals using published methods and reference GWAS meta-analyses from European and multi-ethnic studies. The different PRSs included the GWAS-significant variants, clumping and thresholding (C&T),[8] and PRS-CS,[22] as well as a combination of PRSs calculated with different reference GWASs. We compared the performance of different PRSs among Hispanic/Latino individuals from the Caribbean (Cuban, Dominican, and Puerto Rican) and Mainland (Central American, South American, and Mexican), as these groups differ by their ancestry admixture (high West African and high Native American ancestries, respectively).[9] We further examined the association of eGFR PRSs with incident and prevalent CKD and hypertension given that CKD and hypertension have interlinked pathophysiology. We also replicated our PRS findings in an additional dataset of Hispanic/Latino individuals using data from the Women's Health Initiative (WHI).

[1]Department of Biostatistics, University of North Carolina, Chapel Hill, NC, USA; [2]Division of Sleep and Circadian Disorders, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA; [3]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA; [4]Department of Biostatistics, University of Washington, Seattle, WA, USA; [5]Graduate School of Public Health, San Diego State University, San Diego, CA, USA; [6]Department of Medicine, University of Illinois at Chicago, Chicago, IL, USA; [7]Department of Epidemiology, University of North Carolina, Chapel Hill, NC, USA

[8]Lead contact

*Correspondence: lyzhou@email.unc.edu

https://doi.org/10.1016/j.xhgg.2023.100177.

Table 1. Descriptive statistics for covariates used in analysis in the HCHS/SOL study population overall and stratified by Mainland and Caribbean groups

Variable	Full (n = 12,461)	Mainland (n = 6,871)	Caribbean (n = 5,590)
Center, n (weighted %)			
Bronx	3,178 (28.3)	434 (10.9)	2,744 (45.0)
Chicago	2,976 (15.0)	2,326 (23.0)	650 (7.2)
Miami	3,414 (32.1)	1,275 (17.3)	2,139 (46.3)
San Diego	2,893 (24.7)	2,836 (48.8)	57 (1.5)
Gender, n (weighted %)			
Female	7,330 (51.1)	4,122 (51.1)	3,208 (51.0)
Male	5,131 (48.9)	2,749 (48.9)	2,382 (49.0)
Genetic group, n (weighted %)			
Central American	1,366 (7.9)	1,366 (16.1)	0
Cuban	2,238 (24.3)	0	2,238 (47.8)
Dominican	1,153 (10.0)	0	1,153 (19.6)
Mexican	4,608 (35.3)	4,608 (71.9)	0
Puerto Rican	2,199 (16.6)	0	2,199 (32.6)
South American	897 (5.8)	897 (11.9)	0
eGFR (mL/min/1.73m^2), mean ± SD	99.40 ± 0.31	103.83 ± 0.40	95.14 ± 0.40
Age (years), mean ± SD	41.55 ± 0.26	39.10 ± 0.32	43.91 ± 0.37

n (weighted %): total number of participants (weighted column percentage); eGFR: estimated filtration glomerular rate.

Material and methods

The Hispanic Community Health Study/Study of Latinos (HCHS/SOL)

The HCHS/SOL is a longitudinal study of 16,415 Hispanic/Latino individuals (aged 18–74 years at screening) recruited from households in predefined census-block groups from four US field centers (Chicago, Miami, the Bronx, and San Diego) between 2008 and 2011 and a second examination (2014–2017) performed at an average of follow-up of about 6 years.[10] HCHS/SOL individuals were sampled through a stratified multi-stage area probability sample design. A baseline clinical examination included clinical, behavioral, and sociodemographic assessments and the collection of fasting blood and spot urine samples. Serum creatinine was measured using a creatinase enzymatic method traceable to isotope dilution mass spectrometry (IDMS).[11] The study was approved by the institutional review boards at each field center and the coordinating center, and all subjects provided written informed consent. Individuals in this analysis also consented for genetic studies. There were 12,461 HCHS/SOL participants with complete visit 1 data. Of those, there were 11,534 individuals with visit 2 data.

The HCHS/SOL genotype data are annotated in the hg19 genome build. HCHS/SOL participants were genotyped using a Custom Illumina Omni2.5M array (HumanOmni2.5-8v.1-1, containing 2,536,661 SNVs), which was called using GenomeStudio v.2011.1, Genotyping Module v.1.9.4, and GenTrain v.2. In the reference GWAS, we removed duplicates, insertions or deletions (indels), and SNPs with minor allele frequencies (MAFs) <0.01. Further details of quality control are previously described in Laurie et al.[12] Principal components (PCs) were estimated in an unrelated subset of HCHS/SOL subjects, excluding 19 subjects with substantial Asian ancestry as previously described in Conomos et al.[9] Conomos et al. also estimated the genetic groups based on country of origin and genetic data, which were used as covariates in the statistical analyses. Untyped variants were imputed using the phased haplotypes from 1000 Genome Project phase 1 (more details in Conomos et al.).

HCHS/SOL outcomes and covariates

Our primary outcome was eGFR from visit 1 cross-sectional data. We computed eGFR based on the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) creatinine equation, noting that we do not use a race component for Hispanic/Latino individuals.[13] Secondary outcomes of interest were prevalent and incident CKD and hypertension. Prevalent CKD was defined by an eGFR <60 mL/min/1.73 m^2. Incident CKD was defined as eGFR <60 mL/min/1.73 m^2 at visit 2 among participants without CKD at visit 1 and an eGFR decline ≥1 mL/min/1.73 m^2/year between visits 1 and 2. Prevalent hypertension was defined, using the new ACC/AHA Guidelines definition, as having a systolic or diastolic blood pressure greater than or equal to 130/80 mm Hg or self-reported use of anti-hypertensive medications.[14] Incident hypertension was defined by a blood pressure ≥130/80 mm Hg or taking medications at visit 2 among participants with blood pressure <130/80 mmHg without anti-hypertensive medications at visit 1.

Covariates of interest were age at visit 1 or visit 2, field center, gender, genetic group, and the top fivePCs of genetic data, which was shown in Conomos et al. to account for the population substructure in the Hispanic/Latino population. Survey sampling variables used to account for the complex survey design (unequal probability of sampling, stratification, and clustering) were survey

Table 2. Number of variants used in PRS calculation for each threshold by reference GWAS for HCHS/SOL target data

Threshold (t)	5E−8	1E−7	1E−6	1E−5	1E−4	1E−3	1E−2	1E−1
EU	433	378	669	976	1,670	3,603	11,462	49,772
TE	239	364	397	646	1,340	3,886	15,808	76,420

Summary of number of variants used at each threshold for each reference GWAS for HCHS/SOL target data. C&T PRS (clumping with parameters distance = 250 kb, r2 = 0.1, and p < 1, then thresholding to SNPs with p < t). EU and TE denote the reference GWAS used for calculating the PRS.

sampling weights at HCHS/SOL visit 1 or 2, census block unit (primary sampling unit), and strata. Variable names are summarized in the supplemental information (Table S1).

Reference GWAS and validation data

The two reference summary statistics GWAS files are the European CKDGen[15] and the trans-ethnic Million Veteran Program[16](MVP). Summary statistics from these GWASs were cleaned and processed for quality control, including removing variant duplicates and indels.

To validate our eGFR primary analysis findings in HCHS/SOL, we used the WHI SNP Health Association Resource (WHI-SHARe)[17] Hispanic dataset, which includes 3,520 Hispanic women and overseen by an institutional review board. WHI-SHARe was genotyped on the Affimetrix Genome-wide Human SNP Array 6.0 using the hg19 build annotation and imputed to the 1000 Genomes reference dataset. Extensive quality control has been applied and are described in previous works.[18]

PRS calculation

We calculated PRSs for each individual as the sum of eGFR-associated alleles weighted by estimated effect sizes. Estimated effect sizes were obtained from two sets of GWAS summary statistics ("reference GWAS"): GWAS meta-analysis of 567,460 European ancestry individuals from CKDGen (EU GWAS) and a trans-ethnic GWAS from the MVP (TE GWAS) (dbGap phs001672) of 280,722 African American and European American participants. SNPs were selected for use in the PRS calculation using the C&T method.[8] The C&T PRS is calculated after clumping variants based on linkage disequilibrium (LD) and selecting variants that are below a chosen p-value significance level (or threshold). In the clumping stage, correlated variants are clumped by selecting the most significant variant and removing from consideration variants that are within 250 kb of this variant and those in LD ($R^2 > 0.1$) with variants estimated in HCHS/SOL. This pruned redundant correlated effects caused by LD between variants. Next, when calculating the PRS, we applied a threshold so that only variants with a p-value lower than a chosen level of significance are used. This step helped reduce noise by excluding null effects.[9] For individual i, the PRS is calculated as

$$\widetilde{\mathrm{PRS}_{\mathrm{i,t}}} = \sum_{\mathrm{j} \in \mathrm{M_t}} \frac{\beta_{\mathrm{j}}\, \mathrm{x_{ij}}}{\mathrm{S_i}} \qquad \text{(Equation 1)}$$

where $\mathrm{M_t}$ is the set of SNPs included in the PRS after C&T at t = 5E−8, 1E−7, 1E−6, 1E−5, 1E−4, 1E−3, 1E−2, and 1E−1; β_{i} is the estimated effect size of the effect allele at SNP i; $\mathrm{x_{ij}}$ is the genotype for individual i at SNP j; and $\mathrm{S_i}$ is the total number of alleles included in the PRS of individual i. We consider the GWAS-significant PRSs to be when t = 5E−8.

After calculating the PRS, we standardized all PRSs at each threshold across all individuals to adjust for different scales of PRS distributions caused by different number, frequencies, and weights of SNPs used. These standardized PRSs were used for all regression analyses. Let $\overline{\mu} = \sum_{\mathrm{i}=1}^{\mathrm{I}} \frac{\widetilde{\mathrm{PRS}_{\mathrm{i,t}}}}{\mathrm{I}}$ and $\sigma^2 = \frac{\sum_{\mathrm{i}=1}^{\mathrm{I}} (\widetilde{\mathrm{PRS}_{\mathrm{i,t}}} - \overline{\mu})^2}{\mathrm{I}}$, and then the standardized PRS is defined as

$$\mathrm{PRS_{i,t}} = \frac{\widetilde{\mathrm{PRS}_{\mathrm{i,t}}} - \overline{\mu}}{\sigma} \qquad \text{(Equation 2)}$$

Using the best performing PRS from each reference GWAS, defined by the PRS in the regression of eGFR with lowest mean-squared error (MSE), we calculated a combined PRS. Let $\mathrm{PR\,S_{EU,i}}$ denote the best PRS using the EU GWAS and $\mathrm{PR\,S_{TE,i}}$ denote the best PRS using the TE GWAS for individual i. We calculated the combined PRS as

$$\widetilde{\mathrm{PRS}_{\mathrm{comb,i}}} = \mathrm{PRS_{EU,i}} + \mathrm{PRS_{TE,i}} \qquad \text{(Equation 3)}$$

To make the combined PRS comparable, we standardized the combined PRS as

$$\mathrm{PRS_{comb,i}} = \frac{\widetilde{\mathrm{PRS}_{\mathrm{comb,i}}} - \overline{\mu}_{\mathrm{c}}}{\sigma_{\mathrm{c}}} \qquad \text{(Equation 4)}$$

where $\overline{\mu_{\mathrm{c}}} = \sum_{\mathrm{i}=1}^{\mathrm{I}} \frac{\widetilde{\mathrm{PRS}_{\mathrm{comb,i}}}}{\mathrm{I}}$ and $\sigma^2 = \frac{\sum_{\mathrm{i}=1}^{\mathrm{I}} (\widetilde{\mathrm{PRS}_{\mathrm{comb,i}}} - \overline{\mu_{\mathrm{c}}})^2}{\mathrm{I}}$.

For comparison, we also computed the PRS using PRS-CS,[22] which is a Bayesian approach that infers posterior effect sizes using CS prior on SNP effect sizes, using both the EU and TE GWAS summary statistics. The PRS-CS is computed using the R scripts provided by the authors in Ge et al.[22]

Primary analysis

To evaluate the association of the PRS with eGFR, we fit a linear regression model of eGFR as a function of PRS at visit 1. We used complex survey procedures to account for unequal probability of sampling, stratification, and clustering. We used visit 1 survey sampling weights and obtained unbiased effect estimates for the HCHS/SOL target population. A subgroup analysis was also performed fitting the same linear regression in individuals categorized as having Caribbean heritage (Cuban, Dominican, and Puerto Rican genetic groups) and Mainlander heritage (Central American, South American, and Mexican genetic groups). From the full data results, we selected the best performing PRS constructed based on each reference GWAS (EU and TE) determined by the regression model with the smallest MSE. We compared the association of the best performing PRS from each reference GWAS with the association of the PRS constructed by PRS-CS.

Secondary analysis

We performed additional analyses using the selected best PRS from each reference GWAS, the combined PRS using the best MSE, the combined PRS using the GWAS-significant threshold, and the PRS calculated by PRS-CS. Due to potential over-fitting through evaluation in the same dataset, we also used the GWAS-significant

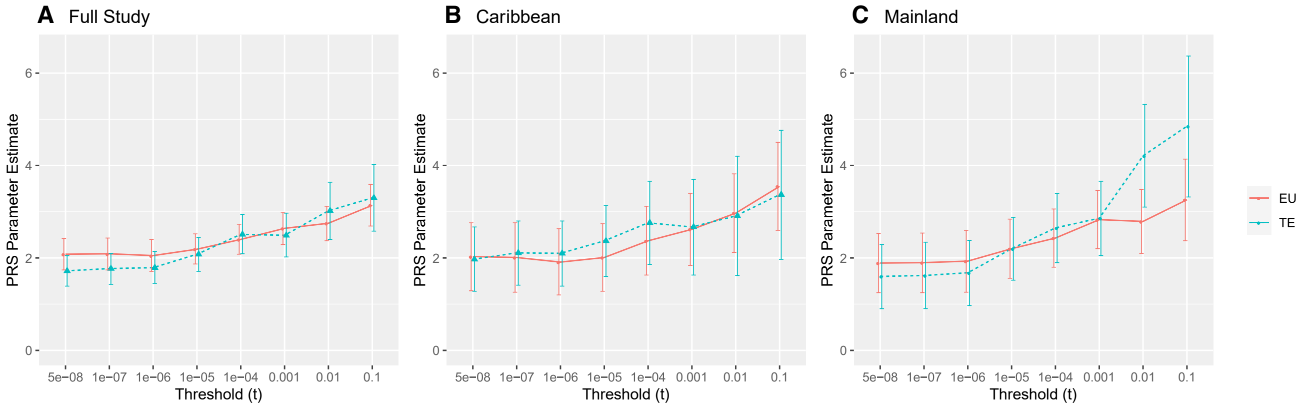

Figure 1. eGFR at visit 1 linear regression parameter estimate and 95% confidence interval for PRS at different thresholds by reference GWASs
(A) Full study population, (B) Caribbean group, and (C) Mainlander group. EU: European reference GWAS, TE: Trans-ethnic reference GWAS

threshold combined PRS, a conservative estimate, which was calculated using threshold $t = 5E-8$ in Equations 3 and 4.

We examined the association of the eGFR PRS with prevalent and incident CKD and hypertension. For modeling prevalence of CKD and hypertension at visit 1 or 2, we fit a logistic regression model with PRS and the covariates of interest using age and sampling variables at visit 1 or 2, respectively. For the incidence CKD and incidence hypertension outcomes, we fit a Poisson model with the same covariates as the linear regression with the years between visit 1 and 2 as an offset. Visit 2 survey sampling weights were used to account for non-responses for visit 2.

Validation analysis

The PRS using each reference GWAS for WHI was calculated with two different methods: (1) using the same methodology used in HCHS/SOL where the C&T is performed specifically in WHI-SHARe and (2) using the SNPs identified in the HCHS/SOL-specific analysis by the C&T algorithm. The PRSs were then standardized using the mean and standard deviation from the HCHS/SOL PRS using the same reference GWAS. We evaluated the two different methods with the combined PRS using the best threshold, the combined PRS using the GWAS-significant threshold, and the individual PRS using each reference GWAS at the best threshold to validate the eGFR linear regression results in HCHS/SOL. We also compared a weighted combined PRS using the parameter estimate for the PRS in the primary analysis linear regression using HCHS/SOL for the reference EU or TE GWAS, denoted by $\widehat{\beta_{EU}}$ and $\widehat{\beta_{TE}}$, respectively. This weighted combined PRS was calculated as follows:

$$\widetilde{\mathrm{PRS}_{\mathrm{wtcomb,i}}} = \widehat{\beta_{EU}}\mathrm{PRS}_{\mathrm{EU,i}} + \widehat{\beta_{TE}}\mathrm{PRS}_{\mathrm{TE,i}} \qquad \text{(Equation 5)}$$

This weighted combined PRS was standardized in the same form as Equation 4. The covariates used in the linear regression are the PRS, age, the first 10 PCs (PC1–PC10), and the region at randomization or enrollment (Table S4).

Software

PRSice-2[19] was used to calculate all PRSs by providing the post-quality control reference GWAS summary statistics file (with duplicated SNPs removed prior) and the reference genotype data in PLINK[20] format. All models were fit using the complex survey procedures in SAS and R-3.6.1.[21]

Results

The descriptive statistics of covariates used in the analysis for the individuals in HCHS/SOL are summarized in Table 1, computed for the full study population, and stratified to Mainland and Caribbean heritages only.

In EU GWAS and TE GWAS, there were 8,885,712 and 1,6089,081 total variants, respectively. After clumping, there were 199,282 variants in the EU GWAS and 300,812 variants in the TE GWAS. Table 2 summarizes the number of variants used to calculate the PRS at the different thresholds.

eGFR at visit 1

Results from association analysis of eGFR are provided in Figure 1, where 12,461 individuals had eGFR at visit 1. All PRSs were highly associated with eGFR at a 0.05 significance level ($p < 1E-20$). Figure 1A summarizes the PRS parameter estimate for each combination with corresponding 95% confidence intervals. A higher PRS was associated with a higher eGFR value. Additionally, as the PRS threshold increased, the parameter estimates also increased for both EU and TE GWASs. Across thresholds, the magnitude of the PRS parameter estimates calculated by different reference GWASs were similar or within a 95% confidence interval. For thresholds below 0.001, the parameter estimates were larger for the EU GWAS than the TE GWAS. However, for thresholds larger than 0.001, the parameter estimates using the TE GWAS were larger than the EU GWAS. The subgroup analysis results of PRS association with eGFR in the Caribbean group were similar to those observed in the full study population (Figure 1B). In the Mainland group, the PRS estimate using the TE GWAS had a larger magnitude than in the full study population or in the Caribbean group (Figure 1C). The eGFR PRS constructed by PRS-CS was similarly highly associated with eGFR (in unstratified analysis, parameter estimate = 2.113 and 1.943 for EU

Table 3. Association of eGFR PRS using EU and TE GWASs with prevalent and incident CKD in HCHS/SOL

Outcome	PRS type	PRS estimate	SE	p
Prevalent CKD at visit 1 (n = 12,461)	combined Best MSE	−0.435	0.084	4.02E−07[a]
	combined GWAS-sig	−0.296	0.081	0.0003[a]
	best EU GWAS	−0.335	0.067	7.05E−07[a]
	best TE GWAS	−0.410	0.081	6.07E−07[a]
	PRS-CS EU GWAS	−0.181	0.074	0.0151[a]
	PRS-CS TE GWAS	−0.275	0.083	0.0009[a]
Prevalent CKD at visit 2 (n = 8,969)	combined Best MSE	−0.346	0.111	0.0020[a]
	combined GWAS-sig	−0.237	0.099	0.0167[a]
	best EU GWAS	−0.334	0.093	0.0003[a]
	best TE GWAS	−0.253	0.089	0.0044[a]
	PRS-CS EU GWAS	−0.225	0.078	0.0043[a]
	PRS-CS TE GWAS	−0.360	0.086	2.98E−05[a]
Incident CKD at visit 2 (n = 8,663)	combined Best MSE	−0.162	0.053	0.0024[a]
	combined GWAS-sig	−0.109	0.050	0.0304[a]
	best EU GWAS	−0.117	0.043	0.0066[a]
	best TE GWAS	−0.132	0.046	0.0043[a]
	PRS-CS EU GWAS	−0.095	0.044	0.0306[a]
	PRS-CS TE GWAS	−0.118	0.046	0.0112[a]

Best EU GWAS: C&T PRS at threshold 0.001 using reference EU GWAS; best TE GWAS: C&T PRS at threshold 0.0001 using reference TE GWAS; combined best MSE: best EU GWAS + best TE GWAS, where "best" is defined as threshold with smallest MSE; combined GWAS-sig: C&T PRS at threshold 5E−8 using EU GWAS + C&T PRS at threshold 5E−8 using TE GWAS. eGFR, estimated filtration glomerular rate; CKD, chronic kidney disease.

[a] Significant at 0.05 level.

and TE GWASs, respectively, $p < 1E-16$ for both in HCHS/SOL).

Secondary analysis

The models fit in the full study population with the lowest MSE and highest R^2 for PRSs constructed by EU or TE GWASs were PRSs at thresholds $t = 1E-3$ and $1E-4$, respectively (Table S3). We used the PRS at these thresholds to calculate the combined best MSE PRS. For comparison, we also calculated the combined PRS using the GWAS-significant threshold in both EU and TE. The results for the CKD and hypertension prevalence and incidence analysis using the combined PRS are presented in Tables 3 and 4. The corresponding area under the receiver operating characteristic (ROC) curve are summarized in Table S4. All eGFR PRSs were significantly associated with a lower risk of prevalent CKD at visit 1 or 2 and incident CKD at visit 2. The combined PRS using the best MSE had the largest effect compared with the best PRS in EU or the best PRS in TE (Table 3). No statistically significant associations were detected for PRSs with prevalent hypertension at visit 1 and 2 or incident hypertension except for the best EU PRS for prevalent hypertension at visit 2 (Table 4). PRSs constructed with PRS-CS had similar significant associations as PRSs constructed with C&T. The effect size was smaller, i.e., the absolute value of the coefficient is closer to zero, using PRS-CS than C&T.

Validation results for eGFR

The description of the Hispanic participants in WHI-SHARe are shown in Table S4. The numbers of SNPs using method 1 at the best threshold for the EU and TE reference GWASs were 8,048 and 2,429, respectively. Using method 2, the numbers of SNPs used in the PRS calculation were 3,538 (EU reference GWAS) and 1,013 (TE reference GWAS) at the best threshold. At the GWAS-significant threshold, there were 754 (EU reference GWAS) and 394 (TE reference GWAS) SNPs used in the PRS calculation using method 1. 421 (EU reference GWAS) and 171 (TE reference GWAS) SNPs were used in the method 2 PRS calculation at the GWAS-significant threshold. As with the HCHS/SOL analysis, the combined PRS was significantly positively associated with eGFR ($\beta = 1.527$, $p < 2E-16$) in covariate-adjusted analyses for the WHI-SHARe Hispanic validation dataset (Table 5). When restricting analysis to the SNPs identified by HCHS/SOL at the threshold of the best PRS, the linear regression analysis with the PRS was significant, using the best PRS constructed with reference EU GWAS or TE GWAS. The combined PRS built upon the EU and TE GWASs using the HCHS/SOL SNPs was also significant, though the association was smaller in magnitude.

Table 4. Association of eGFR PRS using EU and TE GWASs with prevalent and incident hypertension in HCHS/SOL

Outcome	PRS type	PRS estimate	SE	p
Prevalent hypertension at visit 1 (n = 12,461)	combined best MSE	0.010	0.041	0.7958
	combined GWAS-sig	0.033	0.034	0.3264
	best EU GWAS	0.013	0.031	0.6623
	best TE GWAS	0.002	0.038	0.9525
	PRS-CS EU GWAS	0.053	0.031	0.0825
	PRS-CS TE GWAS	0.026	0.032	0.4208
Prevalent hypertension at visit 2 (n = 9,029)	combined Best MSE	0.069	0.048	0.1331
	combined GWAS-sig	0.033	0.042	0.4282
	best EU GWAS	0.078	0.037	0.0348[a]
	best TE GWAS	0.045	0.051	0.3801
	PRS-CS EU GWAS	0.146	0.041	0.0004[a]
	PRS-CS TE GWAS	0.297	0.046	0.5175
Incident hypertension at visit 2 (n = 4,806)	combined Best MSE	0.078	0.058	0.1819
	combined GWAS-sig	−0.018	0.055	0.7420
	best EU GWAS	0.005	0.052	0.9291
	best TE GWAS	−0.074	0.064	0.2486
	PRS-CS EU GWAS	0.006	0.055	0.9161
	PRS-CS TE GWAS	−0.063	0.061	0.2986

Best EU GWAS: C&T PRS at threshold 0.001 using reference EU GWAS; best TE GWAS: C&T PRS at threshold 0.0001 using reference TE GWAS; combined best MSE: best EU GWAS + best TE GWAS, where "best" is defined as threshold with smallest MSE; combined GWAS-sig: C&T PRS at threshold 5E−8 using EU GWAS + C&T PRS at threshold 5E−8 using TE GWAS. eGFR, estimated filtration glomerular rate.

[a] Significant at 0.05 level.

Discussion

The goal of this study was to compare the performance of various PRS approaches for eGFR prediction in the Hispanic/Latino population. The Hispanic/Latino population is admixed, i.e., the individuals have recent admixture from three populations, and the performance of differently constructed PRSs has not been previously compared in this population. We were interested in whether there was an effect on performance based on the type of GWAS used to construct the PRS, EU or TE, as well as which PRS performs better for prediction, combined or individual PRS.

From our primary analysis for eGFR at visit 1, we expected that the PRS constructed with TE would have a higher estimate. However, across different thresholds, there was no difference in PRS performance based on reference GWAS used. The trend of the effect estimate was similar in models evaluated in the Caribbean and Mainland Hispanic/Latino groups, which did not support a differential effect by genetic ancestry admixture, though there is a large proportion of EU ancestry in Hispanic/Latino populations. At thresholds below 0.001, PRSs constructed with TE had slightly lower parameter estimates than PRSs constructed with the EU GWAS. However, for thresholds above 0.001, PRSs constructed with the TE GWAS had slightly higher estimates than PRSs constructed with EU. This result may be driven by the larger sample size of the EU dataset used to select the SNPs for the PRS. We did confirm our hypothesis that a combined PRS would perform better than the individual PRSs when using the best MSE. This is also noted in both the prevalent and incident CKD analysis (Table 3), where the combined PRS has a larger magnitude of effect than either PRSs constructed with EU or TE GWASs and a lower standard deviation. However, when compared with the conservative combination using the GWAS-significant threshold, we see that the combined PRS parameter is generally smaller in magnitude than the individual PRS parameter.

C&T PRS has been the most commonly used PRS in the literature. As seen in Figure 1, as the p-value threshold used for selecting SNPs into the PRS becomes less conservative (i.e., higher), the PRS effect estimate increases. This finding supports the importance of the clumping step, which prunes redundant correlated effects caused by LD between variants, and suggests that the use of C&T PRS should be prioritized over a thresholding-only PRS.

We validated our results in WHI using two different groups of SNPs: (1) using the same methodology as HCHS/SOL where the C&T is performed specifically in WHI-SHARe and (2) using the same set of SNPs identified

Table 5. Different forms of PRS association with eGFR in the WHI-SHARe target dataset to validate findings in the HCHS/SOL dataset

Method	Unweighted estimate	SD	Weighted estimate	SD
1: Combined best threshold	2.724[a]	0.2265	2.726[a]	0.2263
2: combined best threshold with HCHS/SOL SNPs	2.766[a]	0.2266	2.766[a]	0.2262
1: combined GWAS-sig threshold	2.324[a]	0.2161	2.332[a]	0.2151
2: combined GWAS-sig threshold with HCHS/SOL SNPs	2.1484[a]	0.2180	2.764[a]	0.2254
1: EU PRS @ t = 0.001	3.287[a]	0.2894	–	–
1: TE PRS @ t = 0.0001	3.360[a]	0.3433	–	–
2: EU PRS using HCHS/SOL SNPs @ t = 0.001	2.555[a]	0.2256	–	–
2: TE PRS using HCHS/SOL SNPs @ t = 0.0001	2.449[a]	0.2573	–	–
1: EU PRS @ GWAS-sig threshold	2.380[a]	0.2283	–	–
1: TE PRS @ GWAS-sig threshold	2.388[a]	0.2569	–	–
2: EU PRS using HCHS/SOL SNPs @ GWAS-sig threshold	2.029[a]	0.2138	–	–
2: TE PRS using HCHS/SOL SNPs @ GWAS-sig threshold	1.626[a]	0.2062	–	–

Covariates included age, first 10 principal components, and region. Combined PRS: sum of C&T PRS constructed with reference EU GWAS and C&T PRS constructed with reference TE GWAS at best threshold (t = 0.001 and 0.0001 for EU and TE, respectively) or GWAS-significant threshold (t = 5E−08). Using HCHS/SOL SNPs: PRS calculated using set of SNPs identified in the HCHS/SOL dataset for the reference GWAS. Weighted denotes a combined PRS weighted by the corresponding estimated parameter for PRS in HCHS/SOL. EU, European; TE, trans-ethnic; eGFR, estimated filtration glomerular rate.

[a] $p < 0.0001$.

in the HCHS/SOL dataset after C&T for each reference GWAS. Though our validation set from WHI was females only, we expected to see a significant association in eGFR PRS and eGFR due to the highly significant association in the HCHS/SOL dataset, which consisted of males and females. Both the weighted and unweighted eGFR PRSs were significantly associated with eGFR at visit 1 regardless of whether it was constructed using method 1 or 2. Even with a smaller subset of overlapping SNPs used in method 2, the eGFR PRS association was significant. The weighted combined PRS and the unweighted combined PRS had very similar parameter estimates. In contrast, the parameter estimates for method 1 for both EU and TE PRSs at the best thresholds were larger in magnitude compared with method 2. The overestimation could be due to a difference in the WHI subjects LD structure or other different factors from the sample, such as it being only females. In method 1, the same threshold value that was identified best in the HCHS/SOL data. However, in method 2, we are using the SNPs identified in HCHS/SOL, which is essentially using the C&T process in HCHS/SOL.

An additional finding from this study was the lack of association of PRS for eGFR with hypertension outcomes. All eGFR PRSs were not associated with prevalent or incident hypertension; however, this may be due to the smaller sample size and therefore lower power. The current study cannot also distinguish if the null findings were due to low predictive effect of eGFR PRSs in the Hispanic/Latino population or due to a true lack of genetic association between eGFR PRSs and hypertension.

This work has shown that the eGFR PRS is associated with eGFR and CKD in the Hispanic/Latino population, which has not been shown in the current literature. This work highlights the need of developing novel methods to address PRS prediction and health disparities in the Hispanic/Latino population. A limitation of this work is that PRSs are not yet strong enough for prediction, as they do not result in substantial predictive measures. In this article, we have done a preliminary evaluation of the extent of using genetics to learn about the relationship of eGFR. Though C&T PRS is the most commonly used form of PRS used to date, a disadvantage of this method is that it discards more information compared with more recently developed PRS methods that include correlated SNPs while accounting for LD by recomputed SNP weights. However, for a smaller training size and number of causal SNPs, the performance of C&T was expected to be similar to other PRS methods.[22] Nevertheless, we directly compared it with PRS constructed with PRS-CS to verify the performance. This is the first step toward understanding the use of genetics with a relationship to eGFR, and in future studies, we expect the use of other developed PRS methods could show a stronger relationship. Additionally, none of the studies used for selecting the SNPs for the PRS had representation of Native American ancestry.[9] Therefore, further expansion could be to utilize American Indian GWASs, though these types of GWASs are currently very limited. Furthermore, we could expand our work to incorporate the local ancestry data of Hispanic/Latino individuals, which we expect to improve the performance.

Data and code availability

There are restrictions to the availability of HCHS/SOL and WHI data. These data are available through application.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2023.100177.

Acknowledgments

We would like to thank Brian Spitzer with his help on the implementation of PRS-CS for comparison. The baseline examination of HCHS/SOL was carried out as a collaborative study supported by contracts from the National Heart, Lung, and Blood Institute (NHLBI) to the University of North Carolina (N01-HC65233); University of Miami (N01-HC65234); Albert Einstein College of Medicine (N01-HC65235); Northwestern University (N01-HC65236); and San Diego State University (N01-HC65237). The following Institutes/Centers/Offices contributed to the first phase of HCHS/SOL through a transfer of funds to the NHLBI: National Institute on Minority Health and Health Disparities; National Institute on Deafness and Other Communication Disorders; National Institute of Dental and Craniofacial Research (NIDCR); National Institute of Diabetes and Digestive and Kidney Diseases; and National Institute of Neurological Disorders and Stroke, NIH Institution-Office of Dietary Supplements. The Genetic Analysis Center at University of Washington was supported by NHLBI and NIDCR contracts (HHSN268201300005C AM03 and MOD03). Genotyping efforts were supported by NHLBI HSN 26220/20054C, NCATS CTSI grant UL1TR000124, and NIDDK Diabetes Research Center (DRC) grant DK063491. This study is supported by the following grants: National Institutes of Health DK117445, MD012765, HL140385, and HL123677 to N.F. The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts 75N92021D00001, 75N92021D00002, 75N92021D00003, 75N92021D00004, and 75N92021D00005.

Declaration of interests

The authors declare no competing interests.

Received: July 5, 2022
Accepted: January 9, 2023

References

1. Sarnak, M.J. (2003). Cardiovascular complications in chronic kidney disease. Am. J. Kidney Dis. *41*, 11–17.
2. Go, A.S., Chertow, G.M., Fan, D., McCulloch, C.E., and Hsu, C.Y. (2004). Chronic kidney disease and the risks of death, cardiovascular events, and hospitalization. N. Engl. J. Med. *351*, 1296–1305.
3. Collins, A.J., Foley, R.N., Herzog, C., Chavers, B., Gilbertson, D., Herzog, C., Ishani, A., Johansen, K., Kasiske, B., Kutner, N., et al. (2013). US renal data system 2012 annual data report. Am. J. Kidney Dis. *61*. A7, e1-476.
4. Daviglus, M.L., Talavera, G.A., Avilés-Santa, M.L., Allison, M., Cai, J., Criqui, M.H., Gellman, M., Giachello, A.L., Gouskova, N., Kaplan, R.C., et al. (2012). Prevalence of major cardiovascular risk factors and cardiovascular diseases among Hispanic/Latino individuals of diverse backgrounds in the United States. JAMA *308*, 1775–1784.
5. Ricardo, A.C., Flessner, M.F., Eckfeldt, J.H., Eggers, P.W., Franceschini, N., Go, A.S., Gotman, N.M., Kramer, H.J., Kusek, J.W., Loehr, L.R., et al. (2015). Prevalence and correlates of CKD in hispanics/latinos in the United States. Clin. J. Am. Soc. Nephrol. *10*, 1757–1766.
6. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47*, D1005–D1012.
7. Choi, S.W., Mak, T.S.H., and O'Reilly, P.F. (2020). Tutorial: a guide to performing polygenic risk score analyses. Nat. Protoc. *15*, 2759–2772.
8. Privé, F., Vilhjálmsson, B.J., Aschard, H., and Blum, M.G.B. (2019). Making the most of clumping and thresholding for polygenic scores. Am. J. Hum. Genet. *105*, 1213–1221.
9. Conomos, M.P., Laurie, C.A., Stilp, A.M., Gogarten, S.M., McHugh, C.P., Nelson, S.C., Sofer, T., Fernández-Rhodes, L., Justice, A.E., Graff, M., et al. (2016). Genetic diversity and association studies in US hispanic/latino populations: applications in the hispanic community health study/study of Latinos. Am. J. Hum. Genet. *98*, 165–184.
10. Lavange, L.M., Kalsbeek, W.D., Sorlie, P.D., Avilés-Santa, L.M., Kaplan, R.C., Barnhart, J., Liu, K., Giachello, A., Lee, D.J., Ryan, J., et al. (2010). Sample design and cohort selection in the hispanic community health study/study of Latinos. Ann. Epidemiol. *20*, 642–649.
11. Thyagarajan, B., Howard, A.G., Durazo-Arvizu, R., Eckfeldt, J.H., Gellman, M.D., Kim, R.S., Liu, K., Mendez, A.J., Penedo, F.J., Talavera, G.A., et al. (2016). Analytical and biological variability in biomarker measurement in the hispanic community health study/study of Latinos. Clin. Chim. Acta *463*, 129–137.
12. Laurie, C.C., Doheny, K.F., Mirel, D.B., Pugh, E.W., Bierut, L.J., Bhangale, T., Boehm, F., Caporaso, N.E., Cornelis, M.C., Edenberg, H.J., et al. (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. Genet. Epidemiol. *34*, 591–602.
13. Inker, L.A., Schmid, C.H., Tighiouart, H., Eckfeldt, J.H., Feldman, H.I., Greene, T., Kusek, J.W., Manzi, J., Van Lente, F., Zhang, Y.L., et al. (2012). Estimating glomerular filtration rate from serum creatinine and cystatin C. N. Engl. J. Med. *367*, 20–29.
14. Whelton, P.K., Carey, R.M., Aronow, W.S., Casey, D.E., Jr., Collins, K.J., Dennison Himmelfarb, C., DePalma, S.M., Gidding, S., Jamerson, K.A., Jones, D.W., et al. (2018). 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the American College of cardiology/American Heart association task force on clinical practice Guidelines. Hypertension *71*, e13–e115.
15. Wuttke, M., Li, Y., Li, M., Sieber, K.B., Feitosa, M.F., Gorski, M., Tin, A., Wang, L., Chu, A.Y., Hoppmann, A., et al. (2019). A

catalog of genetic loci associated with kidney function from analyses of a million individuals. Nat. Genet. *51*, 957–972.

16. Giri, A., Hellwege, J.N., Keaton, J.M., Park, J., Qiu, C., Warren, H.R., Torstenson, E.S., Kovesdy, C.P., Sun, Y.V., Wilson, O.D., et al. (2019). Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. Nat. Genet. *51*, 51–62.
17. Hays, J., Hunt, J.R., Hubbell, F.A., Anderson, G.L., Limacher, M., Allen, C., and Rossouw, J.E. (2003). The Women's Health Initiative recruitment methods and results. Ann. Epidemiol. *13*, S18–S77.
18. Chen, C.T.L., Fernández-Rhodes, L., Brzyski, R.G., Carlson, C.S., Chen, Z., Heiss, G., North, K.E., Woods, N.F., Rajkovic, A., Kooperberg, C., and Franceschini, N. (2012). Replication of loci influencing ages at menarche and menopause in Hispanic women: the Women's Health Initiative SHARe Study. Hum. Mol. Genet. *21*, 1419–1432.
19. Choi, S.W., and O'Reilly, P.F. (2019). PRSice-2: Polygenic Risk Score software for biobank-scale data. GigaScience *8*, giz082.
20. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
21. R Core Team (2017). R: A Language and Environment for Computing (R Foundation for Statistical Computing). https://wwwR-projectorg/.
22. Ge, T., Chen, C.Y., Ni, Y., Feng, Y.C.A., and Smoller, J.W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat. Commun. *10*, 1776.