Constructing a multi-ancestry polygenic risk score for uterine fibroids using publicly available data highlights need for inclusive genetic research

Jessica L.G. Winters, [1-3*†] Jacqueline A. Piekos,[1-3*] Jacklyn N. Hellwege,[1,4] Ozan Dikilitas,[6] Iftikhar J. Kullo,[6] Daniel J. Schaid,[7] Todd L. Edwards,[5] and Digna R. Velez Edwards[2-3‡]

[1]Vanderbilt Genetics Institute; [2]Department of Biomedical Informatics; [3]Division of Quantitative and Clinical Sciences, Department of Obstetrics and Gynecology; [4]Division of Genetic Medicine, [5]Division of Epidemiology, Department of Medicine; Vanderbilt University Medical Center, Nashville, TN 37203, USA

[6]Department of Cardiovascular Medicine; [7]Department of Health Sciences Research; Mayo Clinic, Rochester, MN 55905, USA

Email: todd.l.edwards@vumc.org, digna.r.velez.edwards@vumc.org

Uterine leiomyomata, or fibroids, are common gynecological tumors causing pelvic and menstrual symptoms that can negatively affect quality of life and child-bearing desires. As fibroids grow, symptoms can intensify and lead to invasive treatments that are less likely to preserve fertility. Identifying individuals at highest risk for fibroids can aid in access to earlier diagnoses. Polygenic risk scores (PRS) quantify genetic risk to identify those at highest risk for disease. Utilizing the PRS software PRS-CSx and publicly available genome-wide association study (GWAS) summary statistics from FinnGen and Biobank Japan, we constructed a multi-ancestry (META) PRS for fibroids. We validated the META PRS in two cross-ancestry cohorts. In the cross-ancestry Electronic Medical Record and Genomics (eMERGE) Network cohort, the META PRS was significantly associated with fibroid status and exhibited 1.11 greater odds for fibroids per standard deviation increase in PRS (95% confidence interval [CI]: 1.05 – 1.17, $p = 5.21 \times 10^{-5}$). The META PRS was validated in two BioVU cohorts: one using ICD9/ICD10 codes and one requiring imaging confirmation of fibroid status. In the ICD cohort, a standard deviation increase in the META PRS increased the odds of fibroids by 1.23 (95% CI: 1.15 – 1.32, $p = 9.68 \times 10^{-9}$), while in the imaging cohort, the odds increased by 1.26 (95% CI: 1.18 – 1.35, $p = 2.40 \times 10^{-11}$). We subsequently constructed single ancestry PRS for FinnGen (European ancestry [EUR]) and Biobank Japan (East Asian ancestry [EAS]) using PRS-CS and discovered a nominally significant association in the eMERGE cohort within fibroids and EAS PRS but not EUR PRS (95% CI: 1.09 – 1.20, $p = 1.64 \times 10^{-7}$). These findings highlight the strong predictive power of multi-ancestry PRS over single ancestry PRS. This study underscores the necessity of diverse population inclusion in genetic research to ensure precision medicine benefits all individuals equitably.

Keywords: Complex Traits; Health Disparities; Risk Assessment; Women's Health

---

[*] These authors contributed equally to this work.

[†] Presenting author.

[‡] Work supported by grants R01HD074711 and R03HD078567 to DRVE and R01HD093671 to DRVE/TLE.

1. Introduction

Uterine fibroids, or uterine leiomyomata, are benign tumors of the uterine smooth muscle that affect a substantial proportion of people with uteruses. While nearly all of these individuals will develop at least one fibroid in their lifetime, only about 50% will experience symptoms, leading to a condition with considerable variability in presentation.[1,2] Fibroids are recognized as a health disparity, with a higher prevalence reported among individuals identifying as Black compared to those identifying as White.[1,3] Additionally, fibroids impose a significant financial burden to the healthcare system, being the leading cause of hysterectomy and gynecological hospitalizations in the United States.[4]

Despite their common occurrence, the genetic factors contributing to fibroid development remain complex and multifactorial. Genome-wide association studies (GWAS) have enhanced our understanding of the genetic underpinnings of uterine fibroids, revealing that the condition is influenced by multiple genetic variants, each contributing a small amount to the overall risk.[5,6] This polygenic nature of fibroids means that identifying individual genes of interest through single-gene studies is insufficient. To better estimate genetic risk for polygenic diseases like fibroids, polygenic risk scores (PRS) have been developed. A PRS aggregates an individual's genetic risk across various loci, providing an overall estimate of their risk for the disease or other clinically relevant outcome.[7] In the context of uterine fibroids, PRS can refine diagnostic accuracy, help identify individuals at high genetic risk for fibroids, and predict the likelihood of treatment resistance or recurrence.[8] This personalized approach allows for more targeted interventions and pre-clinical monitoring, potentially leading to earlier and more effective management.

PRS development has traditionally relied on GWAS data from populations of European ancestry, which limits the applicability of these scores to populations of other ancestries.[9] The use of single ancestry GWAS also exacerbates issues with generalizability. There are several programs for PRS construction, and a review of the different programs and methodologies has been published elsewhere.[10] However, PRS-CSx is an approach which uses linkage disequilibrium (LD) reference panels matched to the ancestry of the GWAS population to perform continuous shrinkage across summary statistics.[11] This approach integrates multiple multi-ancestry GWAS summary statistics from different ancestry groups allowing for more genetic variability to be captured in the score. In 2022, our group published a PRS for fibroids using a European ancestry GWAS and validated it in a population of European ancestry.[12] Here, we aim to extend previous work by developing a multi-ancestry PRS for fibroids applicable to a diverse cohort. By using this method to construct a portable PRS, we hope to address and mitigate racial disparities in precision medicine by overcoming existing limitations in capturing polygenic traits.

2. Materials and Methods

2.1. *Study populations*

The Electronic Medical Records and Genomics (eMERGE) Network (2007 – present) is a national network of DNA repositories that are linked to electronic health records (EHRs). A detailed description of the organization of the eMERGE Network has been previously published.[13] Data contained in the EHR include International Classification of Disease (ICD) diagnostic and procedure codes, basic

demographics, discharge summaries, progress notes, health history, laboratory values, imaging reports, medication orders, and pathology details. Participants in the eMERGE network were genotyped separately, then imputed and merged. A detailed description of the genotyping, imputation, and quality control of the eMERGE phase III array dataset has been previously reported.[14]

The BioVU DNA Repository is a deidentified database of EHRs that are linked to patient DNA samples at Vanderbilt University Medical Center (VUMC). A detailed description about the database and its maintenance has been published elsewhere.[15] The EHR for BioVU contains the same information as stated above for eMERGE. This study also obtained Institutional Review Board (IRB) approval and was conducted in accordance with ethical standards.

While BioVU is a member of eMERGE, samples included in this study are unique to BioVU. BioVU participants were genotyped on a custom MEGA array with genotypes aligned to the forward strand. Initial quality control of both study populations excluded samples or variant sites with missingness above a 2% threshold. Samples were also excluded if consent had been withdrawn, if the sample was duplicated, if there was a failure in sex concordance, or if there was a discrepancy between reported race and genetically determined race. Genetic males were censored from analysis. Imputation was performed on the Michigan Imputation Server using Minimac4 and the 1000 Genomes Phase 3 combined reference panel.[16,17]

Phecodes within the EHR were based from ICD9 and ICD10 codes. Fibroid status in eMERGE was extracted based on phecodes recorded in EHR data.[18] Two cohorts were created in BioVU using different case and control definitions: BioVU-ICD and BioVU-imaged. The BioVU-ICD cohort classified fibroid status similarly to eMERGE, derived from phecodes, while the BioVU-imaged cohort used a previously published algorithm to identify cases or controls based on imaging records indicating the presence or absence of fibroids.[19] In the eMERGE and BioVU-ICD cohorts, cases had at least one code for fibroid diagnosis or a history of fibroid treatment, while controls had no such records. In the BioVU-imaged cohort, cases were identified by a history of fibroids or treatment procedures and at least one imaging procedure confirming fibroid presence. Controls in the BioVU-imaged cohort required two or more imaging events on separate dates without fibroid findings and no history of diagnosis or treatment. Race and ethnicity were determined via reporting through categorical options. The multi-ancestry group was comprised of all individuals that reported as White, Black, or Asian race and Hispanic or non-Hispanic ethnicity. The other two groups were based on either White or Black reported race and Hispanic or non-Hispanic ethnicity. The counts of each strata are given in Table 1.

2.2. *Polygenic risk score development*

Genetic effect weights for PRS construction were derived from uterine fibroid GWAS summary statistics from FinnGen r8 and BioBank Japan.[20,21] Both biobanks determined case and control status based on the presence or absence of ICD9/ICD10 codes or equivalent codes in their healthcare systems. For the multi-ancestry (META) PRS, posterior genetic effect weights were calculated using PRS-CSx, while weights for the single-ancestry scores, European (EUR) and East Asian (EAS) PRS, were calculated using PRS-CS.[11,22] We used linkage disequilibrium (LD) reference panels from the 1000 Genomes Project, with the EUR panel for the FinnGen cohort and the EAS panel for the BioBank Japan cohort. Both PRS-CS and PRS-CSx use a high-dimensional Bayesian framework that calculates a continuous shrinkage prior tailored to a target population, based on the selected LD reference panel.

This shrinkage prior is applied to the raw genetic weights from the source GWAS to derive posterior genetic effect weights, which are then summed to create the PRS. PRS-CS is designed for a single GWAS from a single population, whereas PRS-CSx integrates results from multiple GWAS summary statistics. The programs were applied to three target populations: eMERGE, BioVU-ICD, and BioVU-imaged. Posterior effect weights calculated for each population were summed to create a PRS using PLINK 2.0.[23,24]

2.3. *Statistical analysis*

All statistical analyses were performed using R Statistical Software (v4.2.2).[25] Samples remaining after exclusion in eMERGE and BioVU were used for ten-fold cross validation. Analysis of variance (ANOVA) test was used to determine if age and BMI differed within racial groups between all cohorts. These covariates were chosen because prior literature has revealed associations between uterine fibroid risk with both age and BMI.[1] Student's T-test was used to determine if mean META, EUR, and EAS PRS significantly differed between cases and controls for each racial group within the cohorts. Densities of each PRS stratified on case/control status, were visualized using 'ggplot2.'[26]

Table 1. Racial breakdown of cohorts and population characteristics. Listed below are total counts, mean and standard deviation (SD) of body mass index (BMI) and age, and numbers of cases and controls for each of the three groupings within all cohorts. Race consists of White reported race and non-Hispanic ethnicity (White), Black reported race and non-Hispanic ethnicity (Black), and all the above (All).

Cohort Reported Race	**N**	**BMI (SD)**	**Age (SD)**	**Controls (%)**	**Cases (%)**
eMERGE					
All	23,183	29.07 (7.49)	65.30 (18.69)	21,212 (91)	2,290 (9)
White	20,408	28.68 (7.19)	66.52 (18.52)	18,398 (91)	1,784 (9)
Black	2,775	32.44 (8.66)	56.94 (18.73)	2,306 (84)	439 (16)
BioVU - ICD					
All	33,391	29.27 (7.84)	52.53 (18.61)	32,764 (97)	1,076 (3)
White	27,141	28.69 (7.53)	54.64 (18.20)	25,812 (98)	596 (2)
Black	6,250	32.03 (8.72)	45.19 (18.19)	5,700 (93)	420 (7)
BioVU - imaged					
All	9,182	29.21 (8.08)	44.86 (17.33)	7,910 (84)	1,463 (16)
White	7,294	28.55 (7.69)	46.96 (17.45)	6,082 (86)	975 (14)
Black	1,888	31.90 (9.17)	38.02 (15.41)	1,464 (78)	410 (22)

2.4. *Ten-fold cross validation*

Ten-fold cross validation was performed using the R package 'caret.'[27] Each PRS (META, EUR, EAS) was tested for validation in each of the racial groups for every cohort, resulting in nine different

validation groups in total. Each of the nine groups was split into 80/20 training and testing sets. For each PRS, three models were applied to each of the nine validation groups. The adjusted model constructed the PRS as the main predictor with adjustments for age, BMI, and ten principal components (PCs). The unadjusted model estimated the PRS singularly, while the covariate model analyzed the model created by the covariates—age, BMI, and ten PCs—without the PRS. Odds ratios (OR) and 95% confidence intervals (CI) and pseudo-R^2 were calculated for each model. Area under receiver operator curve (AUROC) for the testing set was calculated using the 'pROC' R package.[28]

3. Results

3.1. *Population characteristics*

Out of 52,548 females in the eMERGE cohort, 23,502 samples passed quality control measures and exhibited fibroid status determinable by ICD codes (eMERGE). The average BMI of the overall group was 29.07 (standard deviation [SD] = 7.49), with 28.68 (SD = 7.19) for the White-reported race strata and 32.44 (SD = 8.66) for the Black-reported race strata. The overall average age was 65.30 (SD = 18.69), with 66.52 (SD = 18.52) for the White-reported race strata and 56.94 (SD = 18.73) for the Black-reported race strata. There were 2,290 fibroid cases in the multi-ancestry group. There were 1,784 cases in the White-reported race strata and 439 cases in the Black-reported race strata to make the prevalence of fibroids 9% and 16%, respectively (Table 1).

BioVU had 51,715 female samples of which 33,840 samples passed quality control and exhibited fibroid status determinable by ICD codes (BioVU-ICD). The average BMI of the multi-ancestry group was 29.27 (SD = 7.84). For the White-reported race strata, the average BMI was 28.69 (SD = 7.53), and for the Black-reported race strata, it was 32.03 (SD = 8.72). The average age of the overall group was 52.53 (SD = 18.61), while it was 54.64 (SD = 18.20) for the White-reported race strata and 45.19 (SD = 18.19) for the Black-reported race strata. There were 1,076 cases in the multi-ancestry group. In the White-reported race strata, there were 596 cases, and in the Black-reported race strata, there were 420 cases, for a fibroid prevalence of 2% and 7%, respectively (Table 1).

Of the 51,715 female individuals in BioVU, 9,373 samples passed quality control and had fibroid status as determined by the imaging algorithm (BioVU-imaged). The average BMI of the overall group was 29.21 (SD = 8.08). In the White-reported race strata, it was 28.55 (SD = 7.69), and in the Black-reported race strata it was 31.90 (SD = 9.17). The average age of the overall group was 44.86 (SD = 17.33). The White-reported race strata had an average age of 46.96 (SD = 17.45), and the Black-reported race strata had an average age of 38.02 (SD = 15.41). There was a fibroid prevalence of 16% out of 1,463 cases in the multi-ancestry group, whereas it was 14% of 975 cases in the White-reported race strata and 22% of 410 cases in the Black-reported race strata (Table 1).

3.2. *Polygenic risk score validation*

3.2.1. *Multi-ancestry (META) PRS*

The META PRS was validated in the multi-ancestry group of the eMERGE, BioVU-ICD, and BioVU-imaged cohorts. Student's T-tests for difference in means found mean META PRS to be significantly different between cases and controls in all multi-ancestry cohorts: $p = 9.85 \times 10^{-9}$ for eMERGE, p =

$2.50x10^{-10}$ for BioVU-ICD, and $p = 3.07x10^{-12}$ for BioVU-imaged (Table 2). For a one standard deviation increase in PRS, the OR for fibroid diagnosis was 1.11 (95% CI: 1.06 – 1.17, $p = 2.43x10^{-5}$) in eMERGE, 1.23 (95% CI 1.15 – 1.32, $p = 9.68x10^{-9}$) in BioVU-ICD, and 1.26 (95% CI: 1.18 – 1.35, $p = 2.4x10^{-12}$) in BioVU-imaged (Figure 1A). The META PRS performed best in the BioVU-imaged cohort with an AUROC of 0.74 (95% CI: 0.71 – 0.77), while the AUROC was 0.67 (95% CI: 0.64 – 0.69) in the eMERGE cohort and 0.66 (95% CI: 0.63 – 0.69) in the BioVU-ICD cohort (Figure 2A). The AUROCs for the covariate models were 0.73 (95% CI: 0.71 - 0.76), 0.66 (95% CI: 0.63 - 0.68), and 0.65 (95% CI: 0.62 - 0.69), respectively.

When the META PRS was applied to each reported race strata separately, it was validated in the White-reported race strata of each cohort but not in the Black-reported race strata (Figures 1B and 1C). The ORs for the White-reported race strata were 1.15 (95% CI 1.09 - 1.22, $p = 6.83x10^{-7}$) in eMERGE, 1.25 (95% CI: 1.15 – 1.39, $p = 5.63=10^{-7}$) in BioVU-ICD, and 1.34 (95% CI: 1.23 – 1.44, $p = 1.34x10^{-12}$) in BioVU-imaged. The META PRS performed best in the White-reported race strata of the BioVU-imaged cohort with an AUROC of 0.70 (95% CI: 0.66 - 0.73), while the AUROC was 0.63 (95% CI: 0.60 – 0.66) in eMERGE and 0.63 (95% CI: 0.58 – 0.68) in BioVU-ICD (Figure 2B). The AUROCs of the covariate model were 0.68 (95% CI: 0.65 – 0.72), 0.63 (95% CI: 0.60 – 0.65), and 0.58 (95% CI: 0.53 – 0.64), respectively. When the META PRS was modeled with covariates in the Black-reported race strata, the model itself had predictability, but the META PRS did not contribute any of the predictability (Figure 2C).

Table 2. Polygenic risk score (PRS) T-test results. Student's T-tests were used to determine if mean PRS was significantly different between cases and controls. Significance level is 0.002 (0.05/27 tests). Cases and controls in the multi-ancestry (META), European ancestry (EUR), and East Asian ancestry (EAS) PRS were stratified according to race: White reported race and non-Hispanic ethnicity (White), Black reported race and non-Hispanic ethnicity (Black), and all the above (All).

Cohort Reported Race	**META PRS**	**EUR PRS**	**EAS PRS**
eMERGE			
All	$9.85x10^{-9}$	$1.89x10^{-7}$	$7.67x10^{-17}$
White	$2.49x^{10-9}$	0.002	$9.90x10^{-12}$
Black	0.57	0.14	0.06
BioVU - ICD			
All	$2.50x10^{-10}$	$4.10x10^{-13}$	$4.64x10^{-6}$
White	$7.06x10^{-10}$	$1.06x10^{-7}$	0.00063
Black	0.07	0.21	0.06
BioVU - imaged			
All	$3.07x10^{-12}$	$6.91x10^{-11}$	$2.75x10^{-8}$
White	$1.77x10^{-13}$	$2.49x10^{-10}$	$2.21x10^{-6}$
Black	0.65	0.85	0.12

3.2.2. *European ancestry (EUR) PRS*

The EUR PRS was validated in the multi-ancestry and White-reported race strata but not in the Black-reported race strata for both BioVU cohorts. The EUR PRS was only validated in the multi-ancestry strata of the eMERGE cohort. Mean EUR PRS was significantly different between cases and controls for all multi-ancestry cohorts: $p = 1.89 \times 10^{-7}$ for eMERGE, $p = 4.10 \times 10^{-13}$ for BioVU-ICD, and $p = 6.91 \times 10^{-11}$ for BioVU-imaged (Table 2). In the multi-ancestry cohorts, the ORs were 1.18 in both BioVU-ICD (95% CI 1.09 – 1.26, $p = 8.94 \times 10^{-6}$) and BioVU-imaged (95% CI: 1.10 – 1.26, $p = 1.89 \times 10^{-6}$) (Figure 1D). The EUR PRS was not associated with the risk of fibroid diagnosis in the eMERGE cohort ($p = 0.30$). The EUR PRS performed best in the BioVU-imaged cohort with an AUROC of 0.74 (95% CI: 0.71 – 0.77), while the AUROC was 0.63 (95% CI: 0.60 – 0.66) in eMERGE and 0.67 (95%

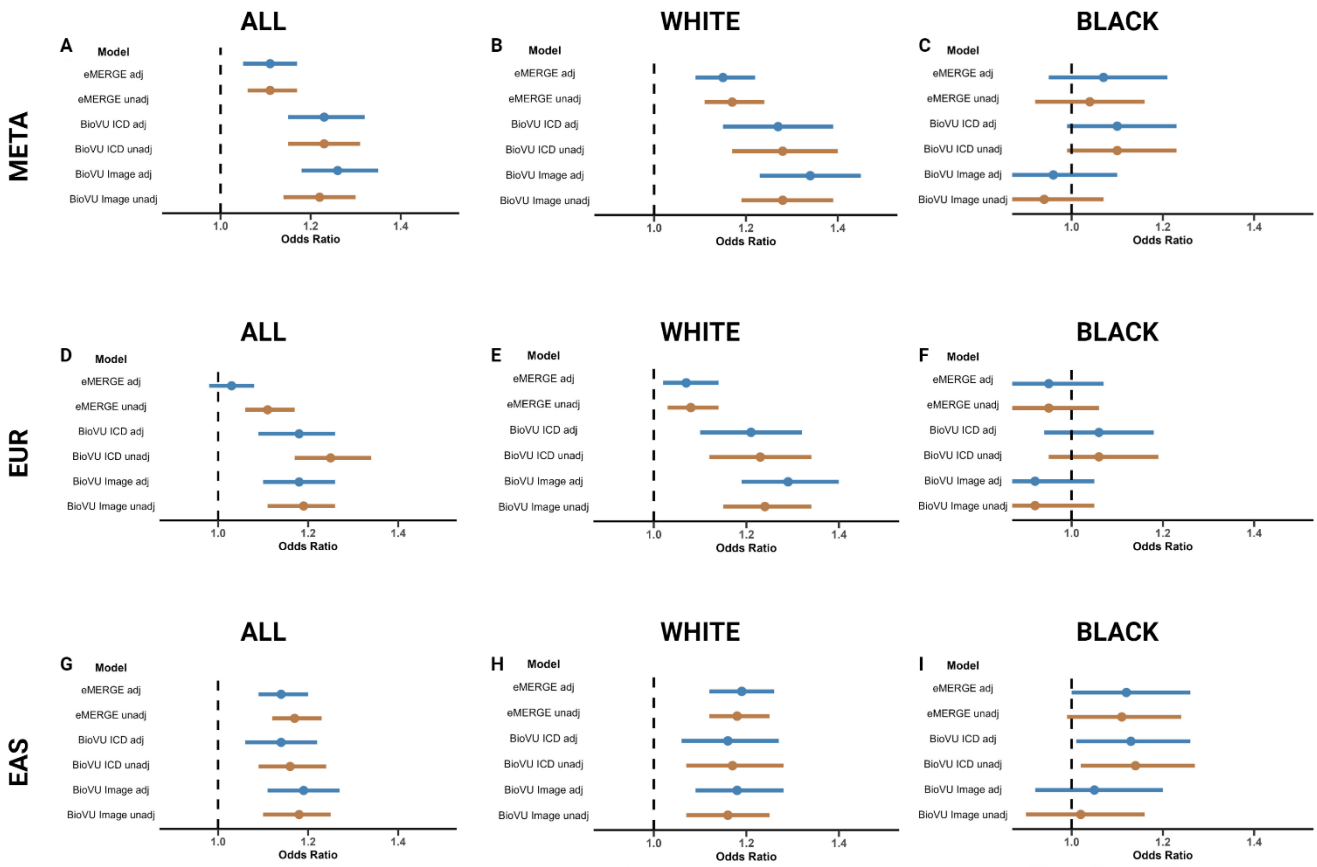

Fig. 1. Polygenic risk score (PRS) ten-fold cross validation results stratified by race for each cohort. Race refers to White reported race and non-Hispanic ethnicity (WHITE), Black reported race and non-Hispanic ethnicity (BLACK), and all the above (ALL). Odds ratios (ORs) are calculated for one standard deviation increase in PRS for adjusted and unadjusted models. **A/B/C)** ORs for all multi-ancestry (META) PRS cohorts. **D/E/F)** ORs for all European ancestry (EUR) PRS cohorts. **G/H/I)** ORs for all East Asian ancestry (EAS) PRS cohorts. Created with Biorender.com.

CI: 0.63 – 0.70) in BioVU-ICD (Figure 2D). The AUROCs for the covariate model were 0.73 (95% CI: 0.71 – 0.76), 0.66 (95% CI: 0.63 – 0.68), and 0.65 (95% CI: 0.62 – 0.69), respectively.

The EUR PRS was applied to the White-reported race strata of the cohorts, but it did not show an association with the risk of fibroid diagnosis in the eMERGE cohort ($p = 0.01$) because it did not reach the significance level of our ten-fold cross-validation for the EUR PRS ($p < 6.17 \times 10^{-4}$). The ORs for the EUR PRS in the BioVU cohorts were 1.21 (95% CI: 1.10 – 1.32, $p = 5.59 \times 10^{-5}$) in BioVU-ICD and 1.29 (95% CI: 1.19 – 1.40, $p = 4.69 \times 10^{-10}$) in BioVU-imaged (Figure 1E). The EUR PRS performed best in the BioVU-imaged cohort with an AUROC of 0.69 (95% CI: 0.66 – 0.72), while the AUROC was 0.63 (95% CI: 0.63 – 0.60 – 0.68) in eMERGE and 0.62 (95% CI: 057 – 0.67) in BioVU-ICD (Figure 2E). The AUROCs of the covariate model were 0.68 (95% CI: 0.65 – 0.72), 0.63 (95% CI: 0.60 – 0.65), and 0.58 (95% CI: 0.53 – 0.64), respectively. The EUR PRS did not associate with risk of

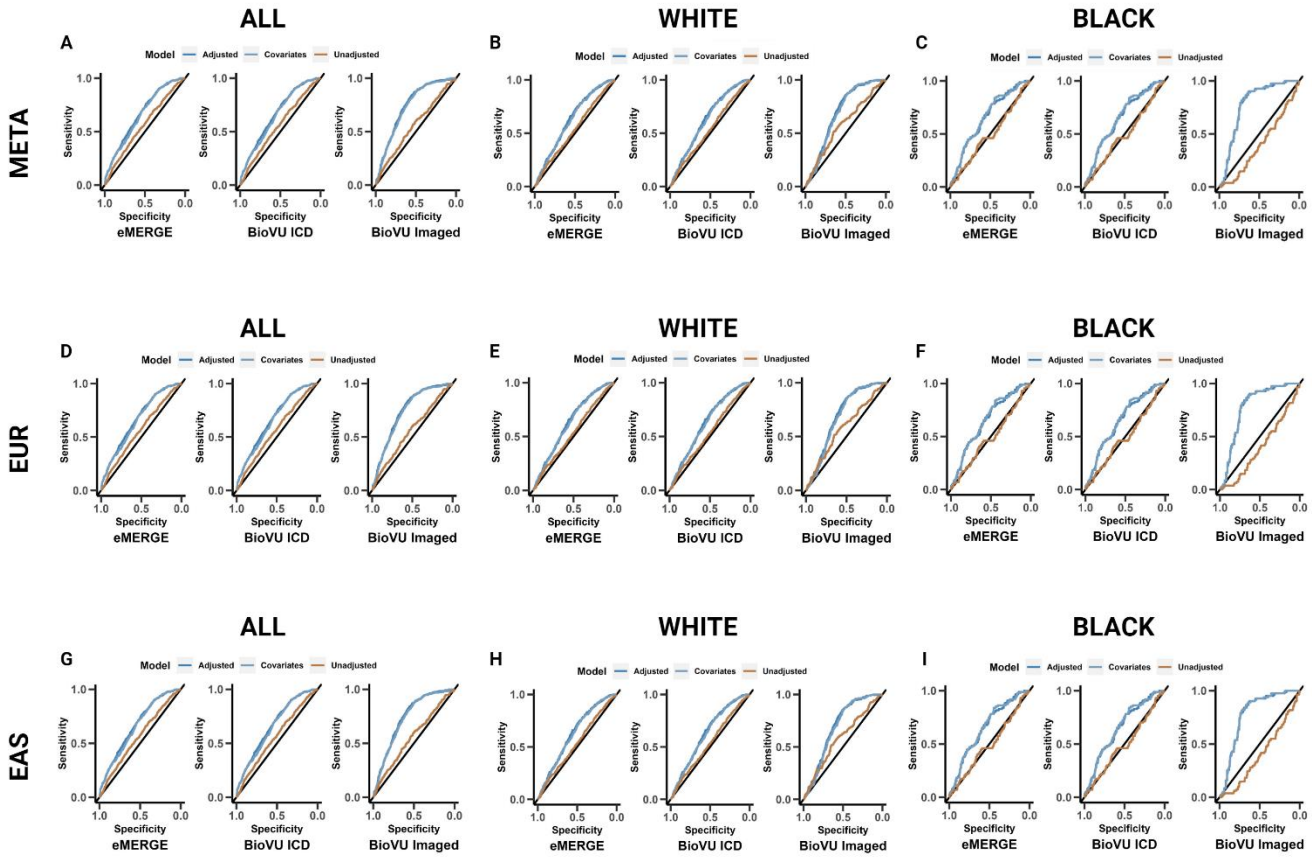

Fig. 2. Polygenic risk score (PRS) ten-fold cross validation results stratified by race for each cohort. Race refers to White reported race and non-Hispanic ethnicity (WHITE), Black reported race and non-Hispanic ethnicity (BLACK), and all the above (ALL). **A/B/C)** Area under receiver operator curve **(**AUROC) plots for each multi-ancestry (META) PRS cohort. **D/E/F)** AUROC plots for each European ancestry (EUR) PRS cohort. **G/H/I)** AUROC plots for each East Asian ancestry (EAS) PRS cohort. Created with Biorender.com.

fibroid diagnosis in the Black-reported race strata of any cohort nor did the models have predictability for fibroid status (Figures 1F and 2F).

3.2.3. *East Asian ancestry (EAS) PRS*

The EAS PRS was validated in the multi-ancestry and White-reported race strata but not the Black-reported race strata for all cohorts. There was a significant difference in mean EAS PRS between cases and controls for all multi-ancestry cohorts: $p = 7.67x10^{-17}$ for eMERGE, $p = 4.64x10^{-6}$ for BioVU-ICD, and $p = 2.75x10^{-8}$ for BioVU-imaged (Table 2). In the multi-ancestry cohorts, the ORs were 1.14 for both eMERGE (95% CI: 1.09 – 1.20, $p = 1.64x10^{-7}$) and BioVU-ICD (95% CI: 1.06 – 1.22, $p = 3.00x10^{-4}$) cohorts, while the BioVU-imaged cohort had a slightly larger OR of 1.19 (95% CI: 1.11 – 1.27, $p = 3.31x10^{-7}$) (Figure 1G). The EAS PRS performed best in the BioVU-imaged cohort with an AUROC of 0.73 (95% CI: 0.71 – 0.76), while the AUROC was 0.68 (95% CI: 0.65 – 0.70) in eMERGE and 0.66 (95% CI: 0.62 – 0.69) in BioVU-ICD (Figure 2G). The AUROCs for the covariate model were 0.73 (95% CI: 0.71 – 0.76), 0.66 (95% CI: 0.63 – 0.68), and 0.65 (95% CI: 0.62 – 0.69), respectively.

When the EAS PRS was applied to the White-reported race strata of each cohort, the ORs were similar: 1.19 (95% CI: 1.12 – 1.26, $p = 1.26x10^{-9}$) in eMERGE, 1.17 (95% CI: 1.07 – 1.28, $p = 1.00x10^{-4}$) in BioVU-ICD, and 1.18 (95% CI: 1.09 – 1.28, $p = 4.20x10^{-5}$) in BioVU-imaged. While the effect size of the EAS PRS was consistent across cohorts, the PRS had the most predictability in the BioVU-imaged cohort with an AUROC of 0.69 (95% CI: 0.66 – 0.72). Next was the eMERGE cohort with an AUROC of 0.64 (95% CI: 0.61 – 0.67) followed by the BioVU-ICD cohort with an AUROC of 0.60 95% CI: 0.54 – 0.65) (Figure 2H). The AUROCs of the covariate model were 0.68 (95% CI: 0.65 – 0.72), 0.63 (95% CI: 0.60 – 0.65), and 0.58 (95% CI: 0.53 – 0.64), respectively. The EAS PRS was not associated with risk of fibroids in the Black-reported race strata of any cohort, nor did it exhibit meaningful predictability (Figures 1I and 2I).

4. Discussion

Using current approaches to estimate PRSs and publicly available resources, we constructed and validated a multi-ancestry (META) PRS in two separate biobanks. META PRS performed better than the single ancestry PRSs, European ancestry (EUR) PRS and East Asian ancestry (EAS) PRS, in all cohorts. These findings show the utility of using a multi-ancestry approach over a single ancestry analysis for PRS. A PRS constructed from the same summary statistics may work in one target population but not others due to a variety of factors including differences in data structures, genotyped variants, and ancestry.[5] By enabling the use of two ancestries over one to construct a PRS, more genetic variation is included in the model, which is precisely what PRS-CSx was created to accomplish.[11] Including multiple different genetic ancestries in a PRS should enable the model to be transferrable to other racial groups, further attempting to answer a problem that has led to portability failures of past PRS models.

PRSs have suffered from an inability to transfer across racial and ethnic groups, resulting in concerns that use of PRS in precision medicine may further contribute to disparities observed in disease trends.[8] When our PRS was evaluated by Black-reporting and White-reporting racial strata, there were differences in validating the findings. The META PRS strongly associated with fibroid status in the

White-reported race strata among all cohorts but failed to validate in any Black-reported race strata. Yet, the AUROC of the modeled covariates in the Black-reported race strata was close to, and in some cases better than, the AUROC for the adjusted META PRS applied to the White-reported race strata. While the META PRS showed no association or predictability with fibroid diagnosis, adding the covariates of age, BMI, and ten PCs were sufficient for a prediction model. Additionally, the pseudo-R^2 was higher in the multi-ancestry group than in the racial strata, demonstrating how adding Black-reporting individuals to the overall model enhances the explained variation. We acknowledge that the smaller sample size of Black-reporting individuals may have limited statistical power, potentially affecting the precision of effect size estimates and the detection of significant associations. However, this limitation is common when studying underrepresented populations, underscoring the need for future efforts to increase sample sizes and improve cohort diversity to enhance the generalizability and accuracy of PRS in Black-reporting individuals. Excluding these populations from prediction modeling only serves to perpetuate health disparities among traditionally underrepresented populations. Thus, while META PRS does not hold any predictive power for Black-reporting individuals alone, their inclusion in the model remains essential for accurate risk assessment based upon clinical factors for all populations.

A major strength of this study is the use of publicly available resources to construct a multi-ancestry fibroid PRS, making it accessible for a broad audience. Utilizing large-scale biobank GWAS summary statistics from the FinnGen research project and the Biobank Japan, which have performed GWAS on thousands of traits, we demonstrated that these projects are sufficient for future PRS studies, sparing researchers from conducting their own on smaller populations. Despite this, we acknowledge the 'messiness' of clinical data used in these studies is often due to case-control definitions based on the presence or absence of a phenotype in an individual's EHR. In particular, case-control definitions based on EHRs are often reliant on the presence or absence of a clinical phenotype, which introduces potential inaccuracies. For example, fibroid cases may be underdiagnosed in individuals who are asymptomatic, resulting in the inclusion of false negatives among controls and subsequently impacting the accuracy and robustness of GWAS associations. A more stringent, precise set of case-control criteria, such as those incorporating diagnostic imaging, would likely improve both GWAS outcomes and PRS performance. This is demonstrated in the study, where the BioVU-imaged cohort, which confirmed fibroid diagnoses through imaging, showed improved AUROC and pseudo-R^2 compared to the ICD-defined cohort, demonstrating enhanced predictability and stability from more precise phenotyping.

Additionally, we observed significant heterogeneity across the populations studied. For instance, the Finnish population's unique genetic background, stemming from a founding bottleneck and relative isolation, may limit transferability to other populations, thereby affecting PRS-CSx program compatibility. This study primarily utilized European and East Asian ancestry data from the FinnGen research project and the Biobank Japan, but did not include African genetic ancestry, despite its known risk factor for fibroids. This highlights a broader issue in genetic research, where populations of European ancestry are often overrepresented, limiting the generalizability of findings. There has been one successful fibroid GWAS in individuals of African ancestry, which identified a unique locus associated with fibroids. This may indicate the genetic architecture of fibroids differs significantly across ancestries.[29,30] Expanding genetic studies into these underrepresented populations should help fill in this missing variance, thus increasing the predictability of PRS. We were unable to use the African

ancestry summary statistics, as that study was performed by our group with samples from BioVU. Removing the overlapping samples from the source population in the BioVU validation cohorts resulted in further insufficient sample sizes for the Black-reported race strata in this study.

In summary, we developed and validated a multi-ancestry (META) PRS in two biobanks, demonstrating superior performance compared to single ancestry PRSs (European and East Asian) across all cohorts. This underscores the advantage of a multi-ancestry approach, which incorporates a broader genetic variation and potentially increases model transferability across different racial groups. Despite the META PRS's strong association with fibroid status in White-reported race strata, it showed limited predictive power for Black-reported race strata, highlighting a persistent challenge in PRS models' applicability across racial groups. Nonetheless, including diverse ancestries in the PRS model improved overall prediction accuracy and addressed disparities in health risk assessment. Strengths of this study include the use of large-scale biobank data and imaging validation to enhance PRS robustness. However, limitations such as inaccurate case-control definitions and a lack of African genetic ancestry in the data underscore the need for more inclusive and precise research methodologies. Ultimately, while multi-ancestry PRS models hold promise for reducing health disparities, further efforts are needed to integrate diverse genetic ancestries and improve predictive accuracy for all populations.

References

1. Baird, D.D., Dunson, D.B., Hill, M.C., Cousins, D. & Schectman, J.M. High cumulative incidence of uterine leiomyoma in black and white women: ultrasound evidence. *Am. J. Obstet. Gynecol.* **188**, 1001–1007 (2003).
2. Wegienka, G., et al. Self-reported heavy bleeding associated with uterine leiomyomata. *Obstet. Gynecol.* **101**, 431–437 (2003).
3. Cramer, S.F. & Patel, A. The frequency of uterine leiomyomas. *Am. J. Clin. Pathol.* **94**, 435–438 (1990).
4. Varol, N., Healey, M., Tang, P., Sheehan, P., Maher, P. & Hill, D. Ten-year review of hysterectomy morbidity and mortality: can we change direction? *Aust. N. Z. J. Obstet. Gynaecol.* **41**, 295–302 (2001).
5. Lewis, C.M., & Vassos, E. Polygenic risk scores: from research tools to clinical instruments. *Genome Med* **12**, 1-11 (2020).
6. Visscher, P.M., Yengo, L., Cox, N.J., & Wray, N.R. Discovery and implications of polygenicity of common diseases. *Science* **373**, 1468-1473 (2021).
7. Wand, H., et al. Improving reporting standards for polygenic scores in risk prediction studies. *Nature* **591**(7849), 211-219 (2021).
8. Adeyemo, A., et al. Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. *Nat Med* **27**, 1876-1884 (2021).
9. Kachuri, L., et al. Principles and methods for transferring polygenic risk scores across global populations. *Nat Rev Genet* **1**, 1-18 (2023).
10. Osterman, M.D., Kinzy, T.G., & Bailey, J.N.C. Polygenic risk scores. *Curr Protoc* **1**, e126 (2021).
11. Ruan, Y. et al. Improving polygenic prediction in ancestrally diverse populations. *Nat. Genet.* **54**, 573–580 (2022).
12. Piekos, J. A., et al. Uterine fibroid polygenic risk score (PRS) associates and predicts risk for uterine fibroid. *Hum. Genet*. **141**, 1739–1748 (2022).
13. McCarty, C.A., et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics* **4**, 1-11 (2011).
14. Stanaway, I.B., et al. The eMERGE genotype set of 83,717 subjects imputed to ~40 million variants genome wide and association with the herpes zoster medical record phenotype. *Genet Epidemiol* **43**, 63-81 (2019).
15. Roden, D.M., et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther* **84**, 362-369 (2008).
16. Gallagher, C.S., et al. Genome-wide association and epidemiological analyses reveal common genetic origins between uterine leiomyomata and endometriosis. *Nat Commun* **10**, 4857 (2019).
17. Das, S., et al. Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284-1287 (2016).
18. Denny, J. C. et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26**, 1205-1210 (2010).
19. Feingold-Link, L., et al. Enhancing uterine fibroid research through utilization of biorepositories linked to electronic medical record data. *J Womens Health* **23**, 1027-1032 (2014).

20. Kurki, M.I., et al. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508-518 (2023).
21. Sakaue, S., et al. A cross-population atlas of genetic associations for 220 human phenotypes. *Nat Genet* **53**, 1415-1424 (2021).
22. Ge, T., Chen, C.Y., Ni, Y., Feng, Y.A., & Smoller, J.W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat Commun* **10**, 1776 (2019).
23. Purcell, S., et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575 (2007).
24. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2021).
25. Chang, C.C., et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
26. Whickam, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York, 2016).
27. Kuhn, M. Building predictive models in R using the caret package. *J Stat Softw* **28**(5), 1-26 (2008).
28. Robin, X., et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **7**, 77 (2011).
29. Hellwege, J.N., et al. A multi-stage genome-wide association study of uterine fibroids in African Americans. *Hum Genet* **136**, 1363-1373 (2017).
30. Edwards, T.L., et al. A trans-ethnic genome-wide association study of uterine fibroids. *Front Genet* **10**, 511 (2019).