Polygenic Risk Scores for Prediction of Breast Cancer Risk in Women of African Ancestry: a Cross-Ancestry Approach

Guimin Gao[1], Fangyuan Zhao[1], Thomas U. Ahearn[2], Kathryn L. Lunetta[3], Melissa A. Troester[4], Zhaohui Du[5], Temidayo O. Ogundiran[6], Oladosu Ojengbede[7], William Blot[8], Katherine L. Nathanson[9], Susan M. Domchek[9], Barbara Nemesure[10], Anselm Hennis[10, 11], Stefan Ambs[12], Julian McClellan,[1] Mark Nie,[1] Kimberly Bertrand[13], Gary Zirpoli[13], Song Yao[14], Andrew F. Olshan[4], Jeannette T. Bensen[4], Elisa V. Bandera[15], Sarah Nyante[16], David V. Conti[17], Michael F. Press[18], Sue A. Ingles[17], Esther M. John[19], Leslie Bernstein[20], Jennifer J. Hu[21], Sandra L. Deming-Halverson[8], Stephen J. Chanock[2], Regina G. Ziegler[2], Jorge L. Rodriguez-Gil[22], Lara E. Sucheston-Campbell[23], Dale P. Sandler[24], Jack A. Taylor[24], Cari M. Kitahara[25], Katie M. O'Brien[24], Manjeet K. Bolla[26], Joe Dennis[26], Alison M. Dunning[27], Douglas F. Easton[26, 27], Kyriaki Michailidou[28], Paul D.P. Pharoah[26, 27], Qin Wang[26], Jonine Figueroa[29, 30], Richard Biritwum[31], Ernest Adjei[32], Seth Wiafe[33], GBHS Study Team, Christine B. Ambrosone[14], Wei Zheng[8], Olufunmilayo I. Olopade[34], Montserrat García-Closas[2], Julie R. Palmer[13], Christopher A. Haiman[17,*], Dezheng Huo[1, 34,*]

[1] Department of Public Health Sciences, The University of Chicago, Chicago, IL, USA
[2] Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA
[3] Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA
[4] Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[5] Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
[6] Department of Surgery, College of Medicine, University of Ibadan, Ibadan, Nigeria
[7] Centre for Population & Reproductive Health, College of Medicine, University of Ibadan, Ibadan, Nigeria
[8] Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN, USA
[9] Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[10] Department of Family, Population and Preventive Medicine, Stony Brook University, Stony Brook, NY, USA
[11] University of the West Indies, Bridgetown, Bardados
[12] Laboratory of Human Carcinogenesis, National Cancer Institute, Bethesda, MD, USA
[13] Slone Epidemiology Center, Boston University, Boston, MA, USA
[14] Department of Cancer Prevention and Control, Roswell Park Comprehensive Cancer Center, Buffalo, NY, USA
[15] Cancer Prevention and Control Program, Rutgers Cancer Institute of New Jersey, New Brunswick, NJ, USA
[16] Department of Radiology, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[17] Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

[18] Department of Pathology, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA
[19] Departments of Epidemiology & Population Health and of Medicine (Oncology) and Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA, USA
[20] Biomarkers of Early Detection and Prevention, Department of Population Sciences, Beckman Research Institute, City of Hope Comprehensive Cancer Center, Duarte, CA, USA
[21] Department of Public Health Sciences, University of Miami Miller School of Medicine, Miami, FL, USA
[22] Genomics, Development and Disease Section, Genetic Disease Research Branch, National Human Genome Research Institute, Bethesda, MD, USA
[23] Department of Veterinary Biosciences, College of Veterinary Medicine, The Ohio State University, Columbus, OH, USA
[24] Epidemiology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC, USA
[25] Radiation Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA
[26] Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK
[27] Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK
[28] Biostatistics Unit, The Cyprus Institute of Neurology & Genetics, Nicosia, Cyprus
[29] Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh Medical School, Edinburgh, UK
[30] Cancer Research UK Edinburgh Centre, Edinburgh, UK
[31] University of Ghana, Accra, Ghana
[32] Komfo Anokye Teaching Hospital, Kumasi, Ghana
[33] School of Public Health, Loma Linda University, Loma Linda, CA, USA
[34] Center for Clinical Cancer Genetics & Global Health, The University of Chicago, Chicago, IL, USA

* Corresponding authors:

Dezheng Huo
Department of Public Health Sciences
University of Chicago
5841, South Maryland Avenue, MC 2000
Chicago, IL, 60637
Email: dhuo@health.bsd.uchicago.edu
Telephone: 773-834-0843

Christopher A. Haiman
Department of Preventative Medicine,
Keck School of Medicine,
University of Southern California, Los Angeles, CA, 90007
Email: haiman@usc.edu

Abstract

Polygenic risk scores (PRSs) are useful for predicting breast cancer risk, but the prediction accuracy of existing PRSs in women of African ancestry (AA) remains relatively low. We aim to develop optimal PRSs for prediction of overall and estrogen receptor (ER) subtype-specific breast cancer risk in AA women. The AA dataset comprised 9,235 cases and 10,184 controls from four genome-wide association study (GWAS) consortia and a GWAS study in Ghana. We randomly divided samples into training and validation sets. We built PRSs using individual level AA data by a forward stepwise logistic regression and then developed joint PRSs that combined 1) the PRSs built in the AA training dataset, and 2) a 313-variant PRS previously developed in women of European ancestry. PRSs were evaluated in the AA validation set. For overall breast cancer, the odd ratio (OR) per standard deviation of the joint PRS in the validation set was 1.34 (95% CI: 1.27-1.42) with area under receiver operating characteristic curve (AUC) of 0.581. Compared to women with average risk (40th-60th PRS percentile), women in the top decile of the PRS had a 1.98-fold increased risk (95% CI: 1.63-2.39). For PRSs of ER-positive and ER-negative breast cancer, the AUCs were 0.608 and 0.576, respectively. Compared to existing methods, the proposed joint PRSs can improve prediction of breast cancer risk in AA women.

Introduction

Breast cancer is the most common cancer in women in the United States and worldwide. It is a complex genetic disorder caused by high-penetrance genes, multiple common variants, and non-genetic factors. In the last 10 years, genome-wide association studies (GWAS) have identified more than 180 breast cancer susceptibility loci (1-4). A polygenic risk score (PRS) is an additive linear combination of the effects of multiple single nucleotide polymorphisms (SNPs) from GWAS, and can achieve a degree of risk stratification that is useful for risk-based programs of breast cancer screening and early detection. PRSs have been developed to predict breast cancer risk in non-Hispanic white, Asian, and Latin American women (5-10). Recently, a large study has developed a 313-variant PRS for breast cancer risks in women of European ancestry (5). This PRS model distinguished breast cancer cases from controls (area under receiver operating characteristic curve, AUC = 0.630 overall), with a better discriminating capacity for ER-positive breast cancer (AUC = 0.641) than for ER-negative breast cancer (AUC = 0.601).

African Americans have higher risk of developing early-onset breast cancer and about 40% higher breast cancer mortality than other racial/ethnic groups in the United States (11), so it is very important to have risk-stratified screening in this population, especially for women age 40 to 49 years. Currently, however, reliable PRS models do not exist for women of African ancestry (AA), including native Africans living in Sub-Saharan Africa and Africa diaspora. Most GWASs of breast cancer were conducted in women of European ancestry, and given the distinct allele frequencies and linkage disequilibrium (LD) structures across populations, PRSs developed in European ancestry populations have an attenuated, though statistically significant, predictive value when applied to African ancestry populations (12, 13). Recently, we showed that the 313-

variant PRS exhibits reduced discriminating accuracy in AA, with AUC being 0.571, 0.588, and 0.562 for overall, ER-positive, and ER-negative breast cancer, respectively (14).

To effectively use genetic information, such as allele frequencies and LD, in AA data, we adopted a forward stepwise logistic regression approach (5) to select genetic variants and then construct PRSs for AA women by using individual genotypic and phenotypic data. The stepwise approach can retain SNPs significantly associated with the phenotype at a given threshold and effectively control the number of noise SNPs used for PRSs. Since the sample sizes of existing AA datasets are much smaller than those from European-ancestry studies, using only AA data to develop a PRS may have limited accuracy. To further increase the prediction accuracy, we adopted the method of Márquez-Luna et al. (15) to develop joint PRSs by combining two components: 1) optimal PRS trained in women of African ancestry by the stepwise logistic regression method, and 2) the 313-variant PRS that was previously developed in women of European ancestry. We used data in women of African ancestry from four breast cancer GWAS consortia and the Ghana Breast Health Study (GBHS); the four consortia were: ROOT (The GWAS of Breast Cancer in the African Diaspora consortium), AMBER (The African American Breast Cancer Epidemiology and Risk consortium), BCAC (Breast Cancer Association Consortium), and AABC (African American Breast Cancer consortium) (see Supplemental Table S1).

Results

We have evaluated the three types of PRS methods described in Materials and Methods: 1) PRSs built by using genome-wide data in women of African ancestry (PRS_{AFR}), 2) the 313-variant PRS using effect sizes directly from previous European ancestry studies (PRS_{EUR}), and 3) the joint

and hybrid PRSs (PRS_{Joint}). The evaluation was performed in an African ancestry validation dataset.

PRSs Built by Using African Ancestry Data Only (PRS_{AFR})

We built PRS models using preset p-value thresholds for filtering SNPs and selecting SNPs by a "hard-thresholding" forward stepwise logistic regression in the African ancestry training set (see Materials and Methods). Table 1 shows the comparison of the performance of these PRS models developed using AA data only and evaluated in independent validation set. Using the forward stepwise regression approach, the prediction accuracy of PRSs increased as the p value threshold increased from 10^{-5} to 0.1. The accuracy increased only slightly when the p value cutoff changed from 0.05 to 0.1, while the numbers of SNPs selected for PRSs for three phenotypes increased about 1.6-fold. Therefore, we used the PRS models with the p value threshold of 0.05 for further analysis. The covariate-adjusted AUCs of PRS_{AFR}, $PRS_{AFR.ERp}$, $PRS_{AFR.ERn}$ were 0.535, 0.546, and 0.548 for overall, ER-positive and ER-negative breast cancer, respectively (16); PRS_{AFR}, $PRS_{AFR.ERp}$, and $PRS_{AFR.ERn}$ denote the PRSs for overall, ER-positive, and ER-negative using 29569, 29004, and 28100 SNPs, respectively, selected by stepwise forward regression in the African ancestry training dataset.

Table 1. Comparison of the performance of PRS models developed using genome-wide approach in AA data: Results in the validation set

P Value Cutoff [a]	SNPs Entering Model (n)	SNPs Selected (n)	OR (95% CI)[b]	AUC (95% CI)[b]
Overall Breast Cancer				
$< 10^{-5}$	288	62	1.04 (0.99-1.10)	0.509 (0.495-0.524)
$< 10^{-4}$	2,053	428	1.03 (0.98-1.09)	0.506 (0.489-0.522)
$< 10^{-3}$	19,067	2,351	1.07 (1.01-1.13)	0.521 (0.507-0.535)
$< 10^{-2}$	175,161	10,647	1.12 (1.06-1.18)	0.535 (0.519-0.551))
< 0.05	829,335	29,569	**1.13 (1.07-1.19)**	**0.535 (0.519-0.551)**
< 0.1	1,615,762	46,854	1.15 (1.09-1.22)	0.541 (0.527-0.556)
ER-positive				
$< 10^{-5}$	201	79	1.06 (0.99-1.13)	0.517 (0.499-0.536)
$< 10^{-4}$	2026	408	1.04 (0.97-1.12)	0.512 (0.491-0.534)
$< 10^{-3}$	20,186	2,339	1.10 (1.03-1.18)	0.529 (0.508-0.550)
$< 10^{-2}$	178,697	10,493	1.19 (1.10-1.27)	0.543 (0.523-0.562)
< 0.05	832,622	29,004	**1.22 (1.13-1.31)**	**0.546 (0.527-0.566)**
< 0.1	1,624,378	45,997	1.22 (1.13-1.31)	0.546 (0.527-0.565)
ER-negative				
$< 10^{-5}$	209	50	1.13 (1.04-1.22)	0.531 (0.508-0.554)
$< 10^{-4}$	1872	419	1.08 (0.99-1.17)	0.528 (0.506-0.550)
$< 10^{-3}$	16,751	2,230	1.03 (0.95-1.11)	0.506 (0.482-0.531))
$< 10^{-2}$	160,097	10,138	1.14 (1.05-1.23)	0.535 (0.510-0.559)
< 0.05	784,928	28,100	**1.20 (1.11-1.31)**	**0.548 (0.525-0.572)**
< 0.1	1,552,045	44,889	1.23 (1.13-1.33)	0.551 (0.527-0.575)

[a] The p value cut off used for selecting SNPs based on their marginal associations with cancer risk and then in stepwise regression in the training set;

[b] Odds ratio (OR) per 1 standard deviation (SD) for the PRS. OR for association with breast cancer in the validation set was derived using logistic regression adjusting for age, consortium/study, and ten PCs. Area under receiver operating characteristic curve (AUC) of PRSs were calculated under the covariate-adjusted ROC model adjusting for age, consortium/study, and ten PCs of genotype data.

The PRS previously developed in women of European ancestry (PRS_{EUR})

Directly applying the PRS developed in data on women of European ancestry (PRS_{EUR}) to our study sample of African ancestry, we found that it was significantly associated with breast cancer risk, with varying prediction accuracy for the three breast cancer phenotypes (Table 2). We noticed that the PRSs trained in women of European ancestry (PRS_{EUR}) had almost no correlation with the PRS developed with "hard-thresholding" approach (PRS_{AFR}) that used AA

data only, suggesting that additional predictive power could be gained if combining these PRSs together (Supplemental Table S3).

The Joint and Hybrid PRS Models

A joint PRS is a weighted linear combination of the two components PRSs, i.e., $PRS_{Joint} = \alpha_1 PRS_{AFR} + \alpha_2 PRS_{EUR}$ (see Materials and Methods). Table 3 shows the prediction performance of the joint and hybrid PRS models in the validation set. For each phenotype, the two-component joint PRS model performed better than individual PRSs. For overall breast cancer, adding the PRS developed in European ancestry population (PRS_{EUR}) to the base model developed using "hard-thresholding" stepwise regression approach (PRS_{AFR}), the AUC increased from 0.535 to 0.577. Similar results were observed for ER-positive and ER-negative breast cancer. Interestingly, the PRSs developed in European ancestry population contributed more to the two-component joint PRS model for overall (69%) and ER-positive breast cancer (65%). By contrast, the PRS developed using AA data (47%) has similar contribution to the joint PRS of ER-negative disease as the PRS developed in European ancestry population (53%). The ORs per unit standard deviation was 1.49 (95% confidence interval, CI: 1.39-1.60) for the joint PRS of ER-positive breast cancer and 1.31 (95% CI: 1.21-1.43) for the joint PRS of ER-negative disease.

The joint PRS for overall breast had lower prediction accuracy (AUC = 0.577) than the joint PRSs for ER-positive (AUC = 0.608) and almost the same accuracy for ER-negative disease (AUC = 0.576). Therefore, we calculated the hybrid PRS for overall breast cancer that combines the PRSs of ER-positive and ER-negative diseases weighted by subtype proportions. The OR per standard deviation of the hybrid PRS was 1.34 (95% CI: 1.27-1.42) with an AUC of 0.581. The SNPs and corresponding joint effect sizes used for the final joint and hybrid PRSs for the three phenotypes are listed in Supplemental Tables S4, S5, and S6.

The contributing weights α_k (k=1,2) of the two component PRSs (PRS_{AFR} and PRS_{EUR}) in the joint PRS models (Table 2) were estimated in the validation set with a logistic regression model including the two components PRSs, so there might be an overfitting problem. For the two-component joint PRS of overall breast cancer, the liability scale adjusted R^2 was 1.86%, which was very similar to the raw R^2 of 1.91%. For ER-positive joint PRS, the adjusted and raw R^2 were 3.60% and 3.66%, respectively. For ER-negative joint PRS, the adjusted and raw R^2 were 1.13% and 1.21%, respectively. These analyses suggested that the bias due to overfitting is minimal.

Table 2. Performance of ancestry-specific and joint prediction PRS models in the validation set

	Weight (α_k) for each predictor [a]	OR (95% CI)[a]	*P*	AUC (95% CI)[a]
Overall Breast Cancer				
PRS_{AFR} (genome-wide threshold P<0.05)		1.13 (1.07-1.19)	$7.8x10^{-06}$	0.535 (0.519-0.551)
PRS from European ancestry (PRS_{EUR})[b]		1.30 (1.23-1.37)	$2.8x10^{-21}$	0.571 (0.557-0.585)
$\alpha_1 PRS_{AFR} + \alpha_2 PRS_{EUR307}$	α_1=0.31, α_2=0.69	1.34 (1.27-1.41)	$3.4x10^{-25}$	0.577 (0.561-0.593)
PRS_{hybrid}[c]		1.34 (1.27-1.42)	$3.0x10^{-26}$	0.581 (0.566-0.597)
ER-positive				
$PRS_{AFR.ERp}$ (genome-wide threshold P<0.05)		1.22 (1.13-1.31)	$2.7x10^{-7}$	0.546 (0.527-0.566)
PRS from European ancestry ($PRS_{EUR.ERp}$)[b]		1.43 (1.33-1.53)	$6.1x10^{-24}$	0.597 (0.577-0.617)
$\alpha_1 PRS_{AFR.ERp} + \alpha_2 PRS_{EUR.ERp}$	α_1=0.35, α_2=0.65	1.49 (1.39-1.60)	$1.1x10^{-28}$	0.608 (0.588-0.627)
ER-negative				
$PRS_{AFR.ERn}$ (genome-wide threshold P<0.05)		1.20 (1.11-1.31)	$1.1x10^{-5}$	0.548 (0.525-0.572)
PRS from European ancestry ($PRS_{EUR.ERn}$)[b]		1.23 (1.13-1.34)	$8.7x10^{-7}$	0.557 (0.534-0.581)
$\alpha_1 PRS_{AFR.ERn} + \alpha_2 PRS_{EUR.ERn}$	α_1=0.47, α_2=0.53	1.31 (1.21-1.43)	$1.1x10^{-10}$	0.576 (0.553-0.598)

[a] Weight (α_k) in the joint PRSs was estimated in validation set with a logistic regression model including two component PRSs (PRS_{AFR}, and PRS_{EUR307}) as predictors, and adjusting for age, consortium/study, and ten PCs; Odds ratio (OR) per 1 SD. Area under receiver operating characteristic curve (AUC) of PRSs were calculated under the covariate-adjusted ROC model adjusting for age, consortium/study, and ten PCs of genotype data.

[b] For the 313 SNPs reported by Mavaddat et al. (5) for PRS in women of European ancestry, 307 SNPs appeared in our data of African ancestry.

[c] PRS_{hybrid} for overall cancer risk is a linear combination of the two joint PRSs for ER-positive and ER-negative breast cancer, with weight of 0.62 for ER-positive and 0.38 for ER-negative cancer.

Table 3 showed associations between breast cancer risk and percentiles of the joint and hybrid PRSs. Women in the top 10% and 5% of the hybrid PRS had a 1.98-fold (95% CI: 1.63-2.39) and a 2.12-fold (95% CI: 1.67-2.69) elevated overall breast cancer risk compared to women at average risk (PRS in 40th-60th percentiles), respectively. For ER-positive breast cancer, compared to the population average, women in the top 10% and 5% of the joint PRS had a 2.20-fold (95% CI: 1.74-2.77) and a 2.58-fold (95% CI: 1.95-3.42) increased risk, respectively. For ER-negative breast cancer, those in the top 10% and 5% of the joint PRS had a 1.80-fold (95% CI: 1.37-2.38) and a 2.13-fold (95% CI: 1.52-3.00) increased risk, respectively, compared to women at average risk.

The joint and hybrid PRSs were significantly associated with breast cancer risk in women with and without family history of breast cancer (Table 4). We did not see any significant interaction between PRS and family history of breast cancer. In addition, family history was associated with about 1.76 to 2.05-fold increased risk of overall or subtype-specific breast cancer. We only observed slight attenuation of the association of family history with overall breast cancer and ER-negative cancer risk after adjusting for PRS (Table 4).

Table 3. Associations between PRS percentiles and breast cancer risk in the validation set

PRS Category	No. Control	Overall Breast Cancer		ER-positive		ER-negative	
		No. Case	OR (95% CI) [a]	No. Case	OR (95% CI) [a]	No. Case	OR (95% CI) [a]
< 5%	156	100	0.79 (0.59-1.05)	35	0.61 (0.41-0.92)	26	0.74 (0.47-1.18)
5% - 10%	155	102	0.82 (0.62-1.09)	28	0.52 (0.34-0.81)	39	1.09 (0.73-1.63)
0% - 10%	311	202	0.81 (0.65-1.00)	63	0.57 (0.42-0.78)	65	0.92 (0.67-1.27)
10% - 20%	313	180	0.73 (0.58-0.91)	77	0.72 (0.53-0.97)	58	0.84 (0.60-1.18)
20% - 40%	624	422	0.85 (0.72-1.02)	185	0.82 (0.66-1.04)	111	0.80 (0.61-1.06)
40% - 60% (ref.)	624	486	1 (ref.)	222	1 (ref.)	141	1 (ref.)
60% - 80%	624	595	1.22 (1.03-1.44)	266	1.18 (0.95-1.46)	184	1.36 (1.06-1.74)
80% - 90%	312	350	1.45 (1.19-1.76)	192	1.64 (1.29-2.09)	94	1.39 (1.03-1.87)
90% - 100%	311	467	1.98 (1.63-2.39)	256	2.20 (1.74-2.77)	127	1.80 (1.37-2.38)
90% - 95%	155	216	1.83 (1.44-2.34)	107	1.82 (1.35-2.45)	55	1.61 (1.12-2.32)
>95%	156	251	2.12 (1.67-2.69)	149	2.58 (1.95-3.42)	72	2.13 (1.52-3.00)

[a] Odds ratio (95% confidence intervals) were adjusted for age, consortium and 10 principal components.

Table 4. Associations between polygenic risk scores (PRS) and breast cancer risk by family history of breast cancer in the validation set

Model	Overall Breast Cancer	ER-positive	ER-negative
	OR (95% CI) [a]	OR (95% CI) [a]	OR (95% CI) [a]
Association of PRS and cancer risk by family history			
PRS unadjusted for family history	1.31 (1.24-1.39)	1.45 (1.35-1.56)	1.31 (1.20-1.44)
PRS in women without family history	1.30 (1.22-1.39)	1.45 (1.33-1.57)	1.31 (1.19-1.45)
PRS in women with family history	1.34 (1.15-1.56)	1.45 (1.21-1.74)	1.29 (1.04-1.60)
P for testing interaction between PRS and family history	*0.829*	*0.965*	*0.779*
Association of family history and cancer risk			
Family history unadjusted for PRS	1.79 (1.52-2.11)	2.05 (1.70-2.49)	1.76 (1.39-2.23)
Family history adjusted for PRS	1.76 (1.49-2.08)	2.05 (1.68-2.49)	1.72 (1.35-2.18)

[a] For PRS, odds ratios (95% confidence intervals) per 1 SD were presented. For family history, the odds ratio comparing women with versus without family history of breast cancer. In all logistic regression models, age, consortium and 10 principal components were adjusted for.

We did not observe a statistically significant interaction between the joint/hybrid PRSs and age at diagnosis for overall or subtype-specific breast cancer risk (Supplementary Figure S1), although the association between PRS and overall or ER-positive breast cancer risk was weak for women 70 years or older.

We examined association of PRSs and breast cancer risk in two populations: Africans vs. African Americans & African Barbadians. In both populations, PRSs were associated with breast cancer risk and there was no statistically significant interaction (Supplementary Table S7). There was no significant interaction between ancestry groups (<80% African ancestry vs. >80% African ancestry) and PRSs. There was a marginally significant heterogeneity effects of the PRSs for overall breast cancer and ER-negative breast cancer across the five consortia/studies, but not for ER-positive PRS (Supplementary Figure S2). For overall breast cancer, the PRS has a moderate association in the ROOT and AABC consortia, and a stronger association in the AMBER consortium.

Absolute Risk of Developing Breast Cancer According to the PRS

Figure 1 shows the estimated life-time and 10-year absolute risks of breast cancer for African Americans according to percentile of the PRSs. The absolute risk of overall breast cancer by age 80 years was 18.8% for women in the 99th percentile of the hybrid PRS and 4.3% for women in the lowest 1st percentile. The absolute risk of ER-positive breast cancer by age 80 ranged from 2.3% in the lowest percentile of PRS to 17.6% in the highest percentile of PRS. For ER-negative breast cancer, the absolute risk by age 80 ranged from 1.3% to 4.8%. By contrast, the absolute risk of overall breast cancer by age 80 ranged from 3.2% to 31.3% for European American women in lowest and highest percentiles of the 313-variant PRS of European ancestry (5) (Supplementary Figure S4). The absolute risk by age 80 ranged from 2.4% to 31.6% for ER-

positive and from 0.5% to 3.3% for ER-negative breast cancer among European Americans. The dotted line in Figure 1D illustrates the age at which women at different categories of the PRS reach a threshold of 10-year risk of 2%, which corresponds to the average risk for women age 45 years in the U.S. This threshold was reached at 35, 38, and 39 years for women whose PRS is >99th, 95-99th, and 90-95th percentiles, respectively.

Discussion

In this study, we developed and validated joint PRSs of breast cancer among women of African ancestry by pooling multiple studies and leveraging an existing polygenic risk score developed in European ancestry population. We adopted the method of Márquez-Luna et al. (15) to develop the joint PRSs that combined the PRS developed with only data from African ancestry and the 313-variant PRS developed in women of European ancestry (5). With AUCs of 0.581, 0.608, and 0.576 for overall, ER-positive, and ER-negative breast cancer, the joint PRSs provide a better predictive value than previous PRS models in African ancestry women. Allman et al evaluated a 77-variant PRS in African Americans and reported an AUC of 0.55 for overall breast cancer risk (12). Wang et al reported an AUC of 0.531 for a 34-variant recalibrated PRS in women of African ancestry (13). Recently, Du et al evaluated the 313-variant PRS using the same dataset as the current study, and reported an AUC of 0.571, 0.588, and 0.562 for overall, ER-positive, and ER-negative breast cancer, respectively (14). Although comparing with previous models, the improvements in AUCs are not large, the current PRSs can provide better risk stratification, making them suitable for clinical use.

The improved prediction value of the joint PRS models in women of African ancestry may be because it has leveraged the strengths of two types of PRSs. The 313-variant PRS was developed with very large sample size of 94,075 breast cancer cases and 75,017 controls of

European descent in BCAC (5), so it achieves high precision. The PRS model developed using "hard-thresholding" genome-wide approach in AA datasets has the advantage that the training and validation dataset have the similar LD patterns. Of note, the contribution of the individual PRSs to the joint PRSs varied by breast cancer phenotypes. The 313-variant PRS has a better performance in predicting ER-positive than ER- negative breast cancer in both European and African ancestry populations (5, 14). Consistently, it also contributed more to the ER-positive joint PRS in this study. This may reflect that about 80% of breast cancer patients of European ancestry has ER-positive disease, so GWAS data in the BCAC contains more genetic information on ER-positive disease. By contrast, women of African descent patients have higher proportion of ER-negative disease than other populations. Probably because of this, the PRS trained in our combined AA dataset had about half contribution to the joint PRS for ER-negative risk.

We also observed that the subtype-specific PRSs performed better than the PRS for overall breast cancer risk. This is probably because of breast cancer etiology heterogeneity; many genetic variants have different effects on ER-positive and ER-negative breast cancers (4, 17, 18). Therefore, we generated a hybrid PRS for overall breast cancer risk that is a weighted average of ER-positive and ER-negative joint PRSs. We found that the hybrid PRS had higher prediction accuracy than the corresponding joint PRS for overall breast cancer risk. If the finding that "the sum of the parts is greater than the whole" can be confirmed in future studies, it could be a good strategy to estimate omnibus risk of breast cancer (19). While an overall breast cancer risk model and an ER-negative model may be useful for clinical decision making regarding timing and frequency of breast cancer screening, an ER-positive model has the additional advantage of

potentially identify high risk women who may benefit from chemoprevention with endocrine agents.

Although the joint PRS models have a better predictive performance than previous PRS models in African ancestry women, the prediction accuracy is still lower than models reported for other racial/ethnic populations. Mavaddat et al reported AUCs of 0.63 and 0.64 for their 313-variant and 3820-variant PRSs, respectively, for predicting overall breast cancer in women of European ancestry (5). Shieh et al examined the performance of 71- and 180-variant PRS for overall breast cancer in a large Latino study and reported AUCs of 0.61 to 0.63 (10). Wen et al examined a 67-variant PRS for overall breast cancer in East Asians and reported an AUC of 0.61 (9). In another PRS study of Asians, Ho et al examined a 287-variant PRS and reported an AUC of 0.613 for overall breast cancer (20). The weaker performance of PRS in people of African ancestry has been observed in other disease phenotypes (21). One study found that the prediction accuracy was 4.9-fold lower in Africans on average compared with that in European populations for 17 phenotypes, while the reduction in accuracy was 1.6-fold in Hispanic/Latino Americans, 1.7-fold in South Asians, and 2.5-fold in East Asians (21). These observations are consistent with previous studies which showed that poorer PRS performance is related to genetic divergences between training and target populations (22, 23). Therefore, several factors could account for this disparity, including relatively limited sample size, different LD patterns, allele frequencies, and possible heterogeneity in effect sizes between populations.

To further improve prediction accuracy of PRS in people of African ancestry, it is important to include more racially/ethnically diverse individuals in medical genomic research. The ongoing Confluence project led by U.S. National Cancer Institute has prioritized large-scale genotyping for diverse populations (https://dceg.cancer.gov/research/cancer-types/breast-

cancer/confluence-project), so it could improve the prediction accuracy of breast cancer PRS. Advances in methodologies in statistical genetics could also help to develop a better PRS utilizing information hidden in the existing GWAS datasets. For example, sophisticated methods that integrate additional biological information, genetic architecture, and LD information can be promising to apply to diverse populations (24-26). For African Americans, an admixed population, global admixture proportion could help to predict cancer risk (15, 27). We found the proportion of European ancestry was not associated with overall and ER-negative breast cancer ($p>0.3$) but marginally significantly associated with ER-positive breast cancer (odds ratio=1.14 per a 25% increase in European ancestry, $p=0.011$). Global admixture is essentially the same as the first principal component ($r=0.996$), which was used to control for population stratification, so we did not use global admixture in our risk prediction model building. However, local ancestry, which is robust to population stratification, could also be tapped in future studies to gain statistical power to improve accuracy of genetic risk prediction (28-30).

The AUC, a discriminating accuracy metric, of the new PRS model is moderate, but the model could still provide meaningful risk stratification in the population. Women in the top 5^{th} percentile of the new PRS have more than 2-fold elevated breast cancer risk compared to women at average risk. For women at average risk, the American Cancer Society strongly recommends to initiate regular screening mammography at age 45 years, whose 10-year risk of developing breast cancer is about 2% (31). Based on the PRS, we estimated that about 10% of African American women have 10-year risk of 2% before they reach age 40. These women could start breast cancer screening earlier than age 40 and are possibly eligible for intensive screening programs or chemoprevention trials.

In summary, we proposed joint breast cancer PRSs in women of African ancestry, which has moderate prediction value, but are still not optimal. We found that the joint model can gain more information on ER-positive breast cancer prediction from the existing PRS developed in European ancestry population, while GWAS data from African ancestry contributes more information to the prediction of ER-negative breast cancer.

Materials and Methods

Study Participants and Genotyping

This study includes women of African ancestry from four breast cancer GWAS consortia and a study in Ghana, with a combined sample size of 19,419 participants including 9235 breast cancer cases and 10184 controls. Data collection for individual studies of these consortia have been described previously (32-36). Sample size and selected characteristics for each consortium and study are summarized in Supplemental Tables S1. Women in the study sites in United States and Barbados were self-identified as African American or African Barbadian, while women in the African study sites were implied to be of African ancestry. African ancestry was confirmed using GWAS data. For each consortium/study in this project, individual protocols were approved by the relevant Institutional Review Boards at participating centers. All participants provided written informed consent in accordance with the local institutional review boards.

Each consortium/study utilized a different GWAS array. Genotyping and quality control (QC) procedures have been described in details in Supplemental Table S1. The GWAS of Breast Cancer in the African Diaspora consortium (ROOT) consists of study participants from six studies (32), and samples were genotyped using the Illumina HumanOmni 2.5-8v1 array. After quality control (QC), 1,657 cases (404 ER-positive, 374 ER-negative) and 2,028 controls from the ROOT consortium remained in the analysis. The African American Breast Cancer

consortium (AABC) consists of nine epidemiological studies (33, 37, 38). Samples in AABC were genotyped using the Illumina Human 1M-Duo BeadChip. After QC, a total of 3,005 cases (1,517 ER-positive, 986 ER-negative) and 2,713 controls remained in the analysis. The African American Breast Cancer Epidemiology and Risk consortium (AMBER) consists of three studies (34). The AMBER samples were genotyped using the Illumina MEGA array, and after QC, 1406 cases (951 ER-positive, 385 ER-negative) and 2,407 controls remained in the analysis. Nine studies with cases and controls of African ancestry contributed samples to the Breast Cancer Association Consortium (BCAC). Genotyping for BCAC was performed using Illumina OncoArray (with 260K GWAS backbone) (39). After removing overlapped samples between BCAC (OncoArray) with AABC, AMBER and ROOT, a total of 2,268 cases (1,127 ER-positive, 613 ER-negative) and 1,406 controls remained for the analysis. The Ghana Breast Health Study (GBHS) includes 899 cases (296 ER-positive, 277 ER-negative) and 1,630 controls (35, 36). Samples in GBHS were genotyped using Illumina Global Screening Array.

Training Set and Validation Set

In order to pool the samples from these studies, we conducted uniformed imputation using the cosmopolitan reference panel in the 1000 Genomes Project (1KGP) (Phase III release) within each consortium/study by the software IMPUTE2 (http://mathgen.stats.ox.ac.uk/impute/impute_v2.html) (40). After imputation, we filtered in variants (~15 million SNP or indel) with average minor allele frequency (MAF) > 0.01 and average imputation information score > 0.85. The distribution of imputation info across GWAS array was described in Supplemental Table S1. We pooled datasets from the four African ancestry consortia and the Ghana study into a combined dataset. Principal components (PCs) of genotype data were estimated using EIGENSTRAT in the pooled dataset (41, 42). As shown in

the scatter plots of the top five eigenvectors from the principal component analysis (Supplementary Figures S3A and S3B), the first PC can distinguish participants from different continents (Africa vs. North America) and indicates essentially the global proportion of African ancestry. The third and fifth PCs can distinguish countries in Africa. We then randomly split the combined dataset into a training set (n=13,598; 70%) and a validation set (n=5,821; 30%). Model development was conducted in the training set, while the performance of the PRS models were evaluated in the validation set.

Development of PRSs using Genome-wide Data in Women of African Ancestry

A PRS can be expressed as

$$\text{PRS} = \beta_1 G_1 + \beta_2 G_2 + \ldots \beta_k G_k + \ldots + \beta_K G_K \qquad (1)$$

where β_k is the per-allele log odds ratio (OR) for breast cancer associated with SNP k and serves as the weight in PRS calculation, G_k is the allele dosage for SNP k, and K is the total number of SNPs included in the PRS. This form of PRS assumes a log-additive genetic model for individual SNPs, which was considered appropriate in previous PRS development (5-10). To find an optimal PRS, we need to determine which SNPs among all genome-wide variants should be included in the PRS according to association test results from the training dataset. We used a modified version of the model selection strategy outlined by Mavaddat and colleagues (5), which used a "hard-thresholding" forward stepwise logistic regression. First, we performed single SNP-based association tests using multivariable logistic regression in the training set, adjusting for age, consortium/study, and the top ten PCs of genotype data. The per allele log-odds ratios estimated in the single SNP-based analyses are called "marginal" effect sizes. We estimated the association for each of the three phenotypes (overall, ER-positive, and ER-negative breast cancer) in parallel. The model development was also separately for each phenotype. In the "hard-

thresholding" approach, we selected SNPs in three steps. In step 1, we split each chromosome into 5Mb bins and sorted SNPs by p value within each bin. To avoid collinear problem in logistic regression, we filtered SNPs based on LD such that highly correlated SNPs (LD $r^2 > 0.9$) with larger p values were removed. In step 2, we selected SNPs by a series of stepwise forward logistic regression in 5 Mb bin. Only SNPs passing the pre-specified p value thresholds were included in the multivariable models. The SNP with the smallest (conditional) p value was added sequentially to the model, until no further SNPs could be added. We set p value thresholds to be 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.05, and 0.1. In step 3, bins of the same chromosome were combined. SNPs on the boundary of two bins (2 Mb boundary) were filtered using LD and stepwise logistic regressions as described in steps 1 and 2. Finally, marginal beta coefficients for all selected SNPs across the genome were compiled together to calculate a PRS according to Equation 1. We labeled this PRS as PRS_{AFR}. For a high p value threshold (e.g. 0.05), there are many (uncorrelated) SNPs on one chromosome and our sample size is limited, so the logistic model including all SNPs cannot be fit reliably.

The 313-variant PRS using Effect Sizes from European Ancestry Population (PRS_{EUR})

The 313-variant PRS was developed previously using data of European ancestry (5). Although its performance in African ancestry populations is not optimal, it still offers moderate discriminatory ability (14). Therefore, we directly applied the weights (beta coefficients) from the 313-variant PRS in the validation set. Of the 313 variants, 6 variants were removed because of low minor allele frequency or imputation score and the remaining 307 variants are shown in Supplemental Table S2. Here, we use PRS_{EUR}, $PRS_{EUR.ERp}$, and $PRS_{EUR.ERn}$ to denote the PRSs for overall, ER-positive, and ER-negative phenotypes, respectively, where subscript "EUR" indicates the weights are from European ancestry population.

Joint and Hybrid PRS Models

To improve risk prediction in diverse populations, Márquez-Luna et al (15) proposed a multiethnic PRS method. The method combines PRS based on European training data with PRS based on training data from the target population (such as African Americans). Márquez-Luna and colleagues showed that the derived multiethnic PRS significantly improve prediction accuracy in the target population and is robust to overfitting (15). Here, we adapted this method to construct a joint PRS as a weighted linear combination of two PRSs:

$$PRS_{Joint} = \alpha_1 PRS_{AFR} + \alpha_2 PRS_{EUR} \quad (2)$$

where PRS_{AFR} and PRS_{EUR} are polygenic risk scores described above, and the weights α_1 and α_2 were estimated in the validation set using a logistic regression model including PRS_{AFR} and PRS_{EUR} as predictors, and adjusting for age, consortium/study, and ten PCs of genotypes. If we let $\alpha_1 + \alpha_2 = 1$, the weights represent the proportional contribution of the two PRSs on the joint PRS.

Since prediction accuracy of the joint PRS for overall breast cancer was relatively low compared to that of the joint PRS for ER-positive and very close to that of the joint PRS for ER-negative breast cancer, we also developed a hybrid PRS as a linear combination of the joint PRSs for ER-positive and for ER-negative breast cancer: $PRS_{hybrid} = \eta\ PRS_{Joint.\ ERp} + (1 - \eta)\ PRS_{Joint.ERn}$, where $\eta = 0.62$ was the proportion of ER-positive cases in our study samples.

Model Evaluation in the Validation Set

For each PRS model described above, we evaluated its performance in the validation set. As the measure of the discriminating accuracy of a PRS, we calculated adjusted AUC using covariate-adjusted receiver operating characteristic (ROC) regression (16), in which age, consortium, and

the top 10 PCs were adjusted for. The adjusted AUC quantifies the pure discriminating accuracy of a PRS without confounding from other covariates. In the evaluation of joint PRSs, we calculated liability scale adjusted R^2 (43), which roughly corrects for overfitting problem from estimating the contributing weights α_1 and α_2 in the validation set.

To estimate the strength of association, we fit multivariable logistic regression models and calculated OR and 95% CI per unit standard deviation of PRS, adjusting for age, consortium, and the top 10 PCs. We also categorized PRSs by percentile (<5%, 5-10%, 10-20%, 20-40%, 40-60%, 60-80%, 80-90%, 90-95%, >95%) in controls, and calculated adjusted OR for each category with 40-60% as the reference group. All analyses were done for overall, ER-positive, and ER-negative breast cancer, separately.

We examined whether age or first degree family history of breast cancer modified the association between PRS and breast cancer risk by adding interaction terms in logistic regression models. We further examined whether the effect of PRS varied between Africans and African Americans/African Barbadians, between groups defined by African ancestry (<80% vs. >80%), and between the 5 consortium/study.

Calculation of Absolute Risks

We calculated the lifetime and 10-year absolute risks of developing breast cancer (overall and subtype-specific disease), based on population incidence rates and relative risk estimates for different PRS categories after taking into account the competing risk of dying from causes other than breast cancer, as described previously (6). The theoretical ORs for women in different PRS categories versus women in the 40th-60th percentiles were calculated using the method of Wen et al (9), in which PRS was modeled as continuous predictor of breast cancer risk. Other inputs included age-specific breast cancer incidence rates in African Americans from Surveillance,

Epidemiology and End Results (SEER, 2000-2017) (44) and the non-breast cancer mortality rates from Centers for Disease Control and Prevention (CDC 1999-2018) in United States (45). Similarly, we calculated absolute risk of ER-positive and ER-negative breast cancer, using subtype-specific incidence rates from SEER (44) and without accounting for the competing risk of other subtype. As a contrast, we also calculated the lifetime and 10-year absolute risks of developing breast cancer (overall and subtype-specific disease) for European Americans using existing PRS model in women of European ancestry (5) and breast cancer incidence rates in European Americans (44). Further details are provided in the Supplemental Material and Methods.

We conducted the analyses using R v.3.6.0 and Stata v.16. All tests of statistical significance were two-sided.

Acknowledgements

The ROOT investigators were supported by National Cancer Institute grants R01-CA228198, R01-CA142996, R01-CA89085, and P20-CA233307. DH and OIO were also supported by Breast Cancer Research Foundation (BCRF-21-071). DH and GG were also partially supported by the National Cancer Institute (R03-CA227357 and R01-CA242929). KLN and SD were supported by Basser Center for BRCA. FZ was supported by the Susan G. Komen Foundation (TREND21675016).

AABC was supported by a Department of Defense Breast Cancer Research Program Era of Hope Scholar Award to CAH [W81XWH-08-1-0383] and the Norris Foundation. Each of the participating AABC studies was supported by the following grants: MEC (National Institutes of Health grants R01-CA63464 and R37-CA54281); CARE (National Institute for Child Health and Development grant NO1-HD-3-3175, K05 CA136967); WCHS (U.S. Army Medical Research and Material Command (USAMRMC) grant DAMD-17-01-0-0334, the National Institutes of Health grant R01-CA100598, and the Breast Cancer Research Foundation; SFBCS (National Institutes of Health grant R01-CA077305 and United States Army Medical Research Program grant DAMD17-96-6071); NC-BCFR (National Institutes of Health grant U01-CA069417); CBCS (National Institutes of Health Specialized Program of Research Excellence in Breast Cancer, grant number P50-CA58223, and Center for Environmental Health and Susceptibility National Institute of Environmental Health Sciences, National Institutes of Health, grant number P30-ES10126); PLCO (Intramural Research Program, National Cancer Institute, National Institutes of Health); NBHS (National Institutes of Health grant R01-CA100374). The Breast Cancer Family Registry (BCFR) was supported by the National Cancer Institute, National Institutes of Health under RFA-CA-06-503 and through cooperative agreements with members of the Breast Cancer Family Registry and Principal Investigators. The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the BCFR, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government or the BCFR. MP was supported by Breast Cancer Research Foundation, Tower Cancer Research Foundation, and a gift from Dr. Richard Balch.

AMBER was supported by the National Cancer Institute grants P01-CA151135, R01-CA098663, R01-CA058420, UM1-CA164974, R01-CA100598, P50-CA58223, R01CA202981, U01-CA164974, R01-CA228357, and the University Cancer Research Fund of North Carolina. JRP was supported by the Susan G. Komen Foundation and the Karin Grunebaum Foundation. Pathology data were obtained from numerous state cancer registries (Arizona, California, Colorado, Connecticut, Delaware, District of Columbia, Florida, Georgia, Hawaii, Illinois, Indiana, Kentucky, Louisiana, Maryland, Massachusetts, Michigan, New Jersey, New York, North Carolina, Oklahoma, Pennsylvania, South Carolina, Tennessee, Texas, Virginia). For the studies included in AMBER, individual protocols were approved by the relevant Institutional Review Boards (IRBs) and by the IRBs of participating cancer registries as required. The results reported do not necessarily represent the views of the National Institutes of Health, or the state cancer registries.

BCAC is funded by Cancer Research UK [C1287/A16563, C1287/A10118], the European Union's Horizon 2020 Research and Innovation Programme (grant numbers 634935 and 633784 for BRIDGES and B-CAST respectively), and by the European Community´s Seventh Framework Programme under grant agreement number 223175 (grant number HEALTH-F2-2009-223175) (COGS). The EU Horizon 2020 Research and Innovation Programme funding source had no role in study design, data collection, data analysis, data interpretation or writing of the report. The Sister Study was funded by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (Z01-ES044005).

GBHS authors acknowledge the research contributions of the Cancer Genomics Research Laboratory for their expertise, execution, and support of this research in the areas of project planning, wet laboratory processing of specimens, and bioinformatics analysis of generated data. This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under NCI Contract No. 75N910D00024. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. The success of this investigation would not have been possible without exceptional teamwork and the diligence of the field staff who oversaw the recruitment, interviews and collection of data from study subjects. Special thanks are due to the following individuals: Korle Bu Teaching Hospital,Accra—Dr Adu-Aryee, Obed Ekpedzor, Angela Kenu, Victoria Okyne, Naomi Oyoe Ohene Oti, Evelyn Tay; Komfo Anoyke Teaching Hospital, Kumasi— Marion Alcpaloo, Bernard Arhin, Emmanuel Asiamah, Isaac Boakye, Samuel Ka-chungu and; Peace and Love Hospital, Kumasi—Samuel Amanama, Emma Abaidoo, Prince Agyapong, Thomas Agyei, Debora Boateng-Ansong, Margaret Frempong, Bridget Nortey Mensah, Richard Opoku, and Kofi Owusu Gyimah. The study was further enhanced by surgical expertise provided by Dr Lisa Newman of the University of Michigan and by pathological expertise provided by Drs. Stephen Hewitt and Petra Lenz of the National Cancer Institute and Dr. Maire A. Duggan from the Cumming School of Medicine, University of Calgary, Canada. Study management assistance was received from Ricardo Diaz, Shelley Niwa, and Usha Singh. Appreciation is also expressed to the many women who agreed to participate in the study and to provide information and biospecimens in hopes of preventing and improving outcomes of breast cancer in Ghana.

Conflict of Interest Statement

The authors declare no conflicts of interest.

References:

1. Lilyquist J, Ruddy KJ, Vachon CM, Couch FJ. Common Genetic Variation and Breast Cancer Risk-Past, Present, and Future. Cancer Epidemiol Biomarkers Prev. 2018;27(4):380-94.
2. Shu X, Long J, Cai Q, Kweon SS, Choi JY, Kubo M, et al. Identification of novel breast cancer susceptibility loci in meta-analyses conducted among Asian and European descendants. Nature communications. 2020;11(1):1217.
3. Zhang H, Ahearn TU, Lecarpentier J, Barnes D, Beesley J, Qi G, et al. Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. Nat Genet. 2020;52(6):572-81.
4. Amos CI, Dennis J, Wang Z, Byun J, Schumacher FR, Gayther SA, et al. The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. Cancer Epidemiol Biomarkers Prev. 2017;26(1):126-35.
5. Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. Am J Hum Genet. 2019;104(1):21-34.
6. Mavaddat N, Pharoah PD, Michailidou K, Tyrer J, Brook MN, Bolla MK, et al. Prediction of breast cancer risk based on profiling with common genetic variants. J Natl Cancer Inst. 2015;107(5).
7. Pharoah PD, Antoniou AC, Easton DF, Ponder BA. Polygenes, risk prediction, and targeted prevention of breast cancer. N Engl J Med. 2008;358(26):2796-803.
8. Wacholder S, Hartge P, Prentice R, Garcia-Closas M, Feigelson HS, Diver WR, et al. Performance of common genetic variants in breast-cancer risk models. N Engl J Med. 2010;362(11):986-93.
9. Wen W, Shu XO, Guo X, Cai Q, Long J, Bolla MK, et al. Prediction of breast cancer risk based on common genetic variants in women of East Asian ancestry. Breast Cancer Res. 2016;18(1):124.
10. Shieh Y, Fejerman L, Lott PC, Marker K, Sawyer SD, Hu D, et al. A Polygenic Risk Score for Breast Cancer in US Latinas and Latin American Women. J Natl Cancer Inst. 2020;112(6):590-8.
11. DeSantis CE, Ma J, Gaudet MM, Newman LA, Miller KD, Goding Sauer A, et al. Breast cancer statistics, 2019. CA Cancer J Clin. 2019;69(6):438-51.
12. Allman R, Dite GS, Hopper JL, Gordon O, Starlard-Davenport A, Chlebowski R, et al. SNPs and breast cancer risk prediction for African American and Hispanic women. Breast Cancer Res Treat. 2015;154(3):583-9.
13. Wang S, Qian F, Zheng Y, Ogundiran T, Ojengbede O, Zheng W, et al. Genetic variants demonstrating flip-flop phenomenon and breast cancer risk prediction among women of African ancestry. Breast Cancer Res Treat. 2018;168(3):703-12.
14. Du Z, Gao G, Adedokun B, Ahearn T, Lunetta KL, Zirpoli G, et al. Evaluating Polygenic Risk Scores for Breast Cancer in Women of African Ancestry. J Natl Cancer Inst. 2021.
15. Marquez-Luna C, Loh PR, South Asian Type 2 Diabetes C, Consortium STD, Price AL. Multiethnic polygenic risk scores improve risk prediction in diverse populations. Genet Epidemiol. 2017;41(8):811-23.
16. Janes H, Pepe MS. Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. Biometrika. 2009;96(2):371-82.
17. Milne RL, Kuchenbaecker KB, Michailidou K, Beesley J, Kar S, Lindstrom S, et al. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. Nat Genet. 2017;49(12):1767-78.
18. Huo D, Feng Y, Haddad S, Zheng Y, Yao S, Han YJ, et al. Genome-wide association studies in women of African ancestry identified 3q26.21 as a novel susceptibility locus for oestrogen receptor negative breast cancer. Human molecular genetics. 2016;25(21):4835-46.
19. Gierach GL, Yang XR, Figueroa JD, Sherman ME. Emerging Concepts in Breast Cancer Risk Prediction. Curr Obstet Gynecol Rep. 2013;2(1):43-52.

20. Ho WK, Tan MM, Mavaddat N, Tai MC, Mariapun S, Li J, et al. European polygenic risk score for prediction of breast cancer shows similar performance in Asian women. Nat Commun. 2020;11(1):3833.
21. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019;51(4):584-91.
22. Scutari M, Mackay I, Balding D. Using Genetic Distance to Infer the Accuracy of Genomic Prediction. PLoS Genet. 2016;12(9):e1006288.
23. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am J Hum Genet. 2017;100(4):635-49.
24. Vilhjalmsson BJ, Yang J, Finucane HK, Gusev A, Lindstrom S, Ripke S, et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. Am J Hum Genet. 2015;97(4):576-92.
25. Hu Y, Lu Q, Powles R, Yao X, Yang C, Fang F, et al. Leveraging functional annotations in genetic risk prediction for human complex diseases. PLoS computational biology. 2017;13(6):e1005589.
26. Lloyd-Jones LR, Zeng J, Sidorenko J, Yengo L, Moser G, Kemper KE, et al. Improved polygenic prediction by Bayesian multiple regression on summary statistics. Nature communications. 2019;10(1):5086.
27. Fejerman L, John EM, Huntsman S, Beckman K, Choudhry S, Perez-Stable E, et al. Genetic ancestry and risk of breast cancer among U.S. Latinas. Cancer Res. 2008;68(23):9723-8.
28. Guan Y. Detecting structure of haplotypes and local ancestry. Genetics. 2014;196(3):625-42.
29. Chen W, Ren C, Qin H, Archer KJ, Ouyang W, Liu N, et al. A Generalized Sequential Bonferroni Procedure for GWAS in Admixed Populations Incorporating Admixture Mapping Information into Association Tests. Human heredity. 2015;79(2):80-92.
30. Atkinson EG, Maihofer AX, Kanai M, Martin AR, Karczewski KJ, Santoro ML, et al. Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. Nat Genet. 2021;53(2):195-204.
31. Oeffinger KC, Fontham ET, Etzioni R, Herzig A, Michaelson JS, Shih YC, et al. Breast Cancer Screening for Women at Average Risk: 2015 Guideline Update From the American Cancer Society. Jama. 2015;314(15):1599-614.
32. Huo D, Feng Y, Haddad S, Zheng Y, Yao S, Han YJ, et al. Genome-wide association studies in women of African ancestry identified 3q26.21 as a novel susceptibility locus for oestrogen receptor negative breast cancer. Hum Mol Genet. 2016;25(21):4835-46.
33. Chen F, Chen GK, Stram DO, Millikan RC, Ambrosone CB, John EM, et al. A genome-wide association study of breast cancer in women of African ancestry. Human genetics. 2013;132(1):39-48.
34. Palmer JR, Ambrosone CB, Olshan AF. A collaborative study of the etiology of breast cancer subtypes in African American women: the AMBER consortium. Cancer Causes Control. 2014;25(3):309-19.
35. Brinton LA, Awuah B, Nat Clegg-Lamptey J, Wiafe-Addai B, Ansong D, Nyarko KM, et al. Design considerations for identifying breast cancer risk factors in a population-based study in Africa. International journal of cancer. 2017;140(12):2667-77.
36. Nyante SJ, Biritwum R, Figueroa J, Graubard B, Awuah B, Addai BW, et al. Recruiting population controls for case-control studies in sub-Saharan Africa: The Ghana Breast Health Study. PLoS One. 2019;14(4):e0215347.
37. Feng Y, Rhie SK, Huo D, Ruiz-Narvaez EA, Haddad SA, Ambrosone CB, et al. Characterizing Genetic Susceptibility to Breast Cancer in Women of African Ancestry. Cancer Epidemiol Biomarkers Prev. 2017;26(7):1016-26.
38. Feng Y, Stram DO, Rhie SK, Millikan RC, Ambrosone CB, John EM, et al. A comprehensive examination of breast cancer risk loci in African American women. Hum Mol Genet. 2014;23(20):5518-26.

39. Michailidou K, Lindstrom S, Dennis J, Beesley J, Hui S, Kar S, et al. Association analysis identifies 65 new breast cancer risk loci. Nature. 2017;551(7678):92-4.
40. The 1000 Genomes Project Consortium, Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, et al. A global reference for human genetic variation. Nature. 2015;526:68.
41. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS genetics. 2006;2(12):e190.
42. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38(8):904-9.
43. Lee SH, Goddard ME, Wray NR, Visscher PM. A better coefficient of determination for genetic profile analysis. Genetic epidemiology. 2012;36(3):214-24.
44. SEER*Stat Database: Incidence - SEER Research Data, 9 Registries, Nov 2019 Sub (1975-2017) - Linked To County Attributes - Time Dependent (1990-2017) Income/Rurality, 1969-2017 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, released April 2020, based on the November 2019 submission [Internet]. 2020.
45. Underlying Cause of Death 1999-2018 on CDC WONDER Online Database, released in 2020 [Internet]. 1999-2018 [cited Nov 2, 2020]. Available from: http://wonder.cdc.gov/ucd-icd10.html.

Legends

Figure 1. Cumulative life-time and 10-Year Absolute Risk of Developing Breast Cancer among African Americans

Table 1. Comparison of the performance of PRS models developed using genome-wide approach in AA data: Results in the validation set

Table 2. Performance of ancestry-specific and joint prediction PRS models in the validation set

Table 3. Associations between PRS percentiles and breast cancer risk in the validation set

Table 4. Associations between PRS and breast cancer risk by family history of breast cancer in the validation set

Abbreviations

Polygenic risk scores (PRSs); African ancestry (AA); estrogen receptor (ER); genome-wide association study (GWAS); odd ratio (OR); area under receiver operating characteristic curve (AUC); single nucleotide polymorphisms (SNPs); linkage disequilibrium (LD); the Ghana Breast Health Study (GBHS); Breast Cancer Association Consortium (BCAC); The African American Breast Cancer Epidemiology and Risk consortium (AMBER); African American Breast Cancer consortium (AABC); The GWAS of Breast Cancer in the African Diaspora consortium (ROOT); Receive Operating Characteristic (ROC); principal components (PCs); confidence interval (CI); standard deviation (SD); quality control (QC); minor allele frequency (MAF).

Supporting Information

Supplementary Authorship

The GBHS Study Team

Florence Dedey[1], Richard Biritwum[2], Lawrence Edusei[1], Verna Vanderpuye[1], Ernest Adjei[3], Francis Aitpillah[3], Joseph Oppong[3], Margaret Frempong[4], Jonine Figueroa[5], Louise Brinton[6], Thomas U. Ahearn[6], Ernest Osei-Bonsu[3], Nicholas Titiloye[3], Michelle Brotzman[6], Ann Truelove[7], Evelyn Tay[1], Naomi Oyoe Ohene Oti[1], Victoria Okyne[1], Isaac Boakye[3], Bernard Arhin[3], Marion Alcpaloo[3], Emma Abaidoo[4], Prince Agyapong[4], Joe Nat Clegg-Lamptey[1], Joel Yarney[1], Kofi Nyarko[2], Daniel Ansong[3], Baffour Awuah[3], Seth Wiafe[4], Beatrice Addai Wiafe[4], Montserrat Garcia-Closas[6]

Affiliations:
1. Korle Bu Teaching Hospital, Accra, Ghana
2. University of Ghana, Accra, Ghana
3. Komfo Anoyke Teaching Hospital, Kumasi, Ghana
4. Peace and Love Hospital, Kumasi, Ghana
5. University of Edinburgh, Edinburgh, Scotland

6. U.S. National Cancer Institute, Bethesda, MD
7. Westat, Inc., MD, USA

Supplemental Material and Methods (including supplementary Figures S1-4)

Supplemental Table S1. Descriptive characteristics of study samples.

Supplemental Table S2. Beta coefficients of the 313-variant polygenic risk score model

Supplemental Table S3. Correlation coefficients among selected polygenic risk scores in the validation set.

Supplemental Table S4. Beta coefficients for hybrid polygenic risk score for overall breast cancer risk prediction.

Supplemental Table S5. Beta coefficients of joint polygenic risk score for ER-positive breast cancer risk prediction.

Supplemental Table S6. Beta coefficients of joint polygenic risk score for ER-negative breast cancer risk prediction.

Supplemental Table S7. Associations between polygenic risk scores and breast cancer risk by race in the validation set.

Supplemental Figure S1. Association of the polygenic risk score and breast cancer risk in different age categories (in years) in the validation set.

Supplemental Figure S2. Association of the polygenic risk score and breast cancer risk in different consortium/study in the validation set.

Supplemental Figure S3. Scatter plots of the top 5 eigenvectors from principal component (PC) analysis according to consortium/study (A) and country (B).

Supplemental Figure S4. Cumulative life-time and 10-year absolute risk of developing breast cancer among European Americans.

Supplemental Material and Methods

Calculation of lifetime and 10-year absolute risks

We estimated the lifetime absolute risk as well as the 10-year absolute risk of developing breast cancer by age (20-80 years old) for each risk category defined by polygenic risk score (PRS), and further estimated the absolute risk for ER-positive and ER-negative breast cancer respectively. Besides relative risk estimates from the PRS model, we also used data on the breast cancer incidence rates in the United States (Surveillance, Epidemiology and End Results, SEER 2000-2017)[1] and the cause-specific mortality rates in the United States (Centers for Disease Control and Prevention, CDC 1999-2018).[2]

First of all, the overall incidence of breast cancer at age t can be expressed as:

$$i(t) = \frac{\sum_g \tau_g \mu_g(t) S_g(t-1)}{\sum_g \tau_g S_g(t-1)} = \frac{\sum_g \tau_g \mu_o(t) exp(\beta_g) S_g(t-1)}{\sum_g \tau_g S_g(t-1)}$$

g - PRS risk category

τ_g - frequency of the PRS category in the population

β_g - coefficient estimate of breast cancer risk in PRS category g compared with the baseline PRS category, i.e. log odds ratio

$\mu_0(t)$ - breast cancer incidence in the baseline PRS category at age t

$\mu_g(t)$ - breast cancer incidence in PRS category g at age t, $\mu_g(t) = \mu_0(t) exp(\beta_g)$

$S_g(t)$ - the probability of being free of breast cancer to age t in PRS category g:

If $t < 20$:

$S_g(t) = 1$ and $\mu_g(t) = 0$

If $20 \leq t \leq 80$:

$$S_g(t) = S_g(t-1) \cdot \left[1 - \mu_g(t-1)\right] = S_g(t-1) \cdot \left[1 - \mu_0(t-1) exp(\beta_g)\right]$$

Because the overall breast cancer incidence rate in the United States (SEER 2000-2017) for each age group t, $i(t)$ is known, we can use the expression above to iteratively estimate the probability of being free of breast cancer to age t in PRS category g, $S_g(t)$. The theoretical $OR_g = exp(\beta_g)$ for women in PRS category g versus women in the 40th-60th percentiles was calculated using Equation 7 in Wen et al,[3] in which PRS was modelled as continuous predictor of breast cancer risk. The standard deviation (SD) was 0.31, 0.41, and 0.33 for overall, ER-positive, and ER-negative breast cancer PRS, respectively.

Next, we can estimate the lifetime absolute risk of developing breast cancer $AR_g(t)$ considering the competing risk of death from causes other than breast cancer through:

$$AR_g(t) = \sum_{u=0}^{t} \mu_g(u) S_g(u) S_m(u)$$

$\mu_g(t)$ - breast cancer incidence in PRS category g at age t

$S_g(t)$ - the probability of being free of breast cancer to age t in PRS category g

$S_m(t)$ - the probability of surviving to age t

If $t < 20$:

$S_m(t) = 1$

If $20 \leq t \leq 80$:

$S_m(t) = S_m(t-1) \cdot [1 - m(t-1)]$, $m(t)$ - mortality rate from causes other than breast cancer at age t (CDC, 1999-2018)

The 10-year absolute risk of developing breast cancer at age t ($20 \leq t \leq 70$) can be estimated through:

$$\frac{AR_g(t+10) - AR_g(t)}{S_m(t) \cdot S_g(t)}$$

Finally, we repeat the previous steps to estimate the absolute risk of developing ER-positive and ER-negative breast cancer, using the subtype-specific incidence rates in the population as well as the subtype-specific PRS risk categories. When estimating the absolute risk of developing ER-positive or ER-negative breast cancer, we assume that the individual is free of any type of breast cancer at age t.

References

1. Surveillance, E., and End Results (SEER) Program (www.seer.cancer.gov). (2020). SEER*Stat Database: Incidence - SEER Research Data, 21 Registries, Nov 2019 Sub (2000-2017), National Cancer Institute, DCCPS, Surveillance Research Program, released April 2020, based on the November 2019 submission.
2. Centers for Disease Control and Prevention, National Center for Health Statistics. (1999-2018). Underlying Cause of Death 1999-2018 on CDC WONDER Online Database, released in 2020.
3. Wen, W., Shu, X.O., Guo, X., Cai, Q., Long, J., Bolla, M.K., Michailidou, K., Dennis, J., Wang, Q., Gao, Y.T., et al. (2016). Prediction of breast cancer risk based on common genetic variants in women of East Asian ancestry. Breast Cancer Res 18, 124.

Supplemental Figures

Figure S1. Association of the PRS and breast cancer risk in different age categories (in years) in the validation set. (The *P* for test for heterogeneity was 0.13 for overall breast cancer risk, 0.13 for ER-positive, and 0.50 for ER-negative breast cancer).

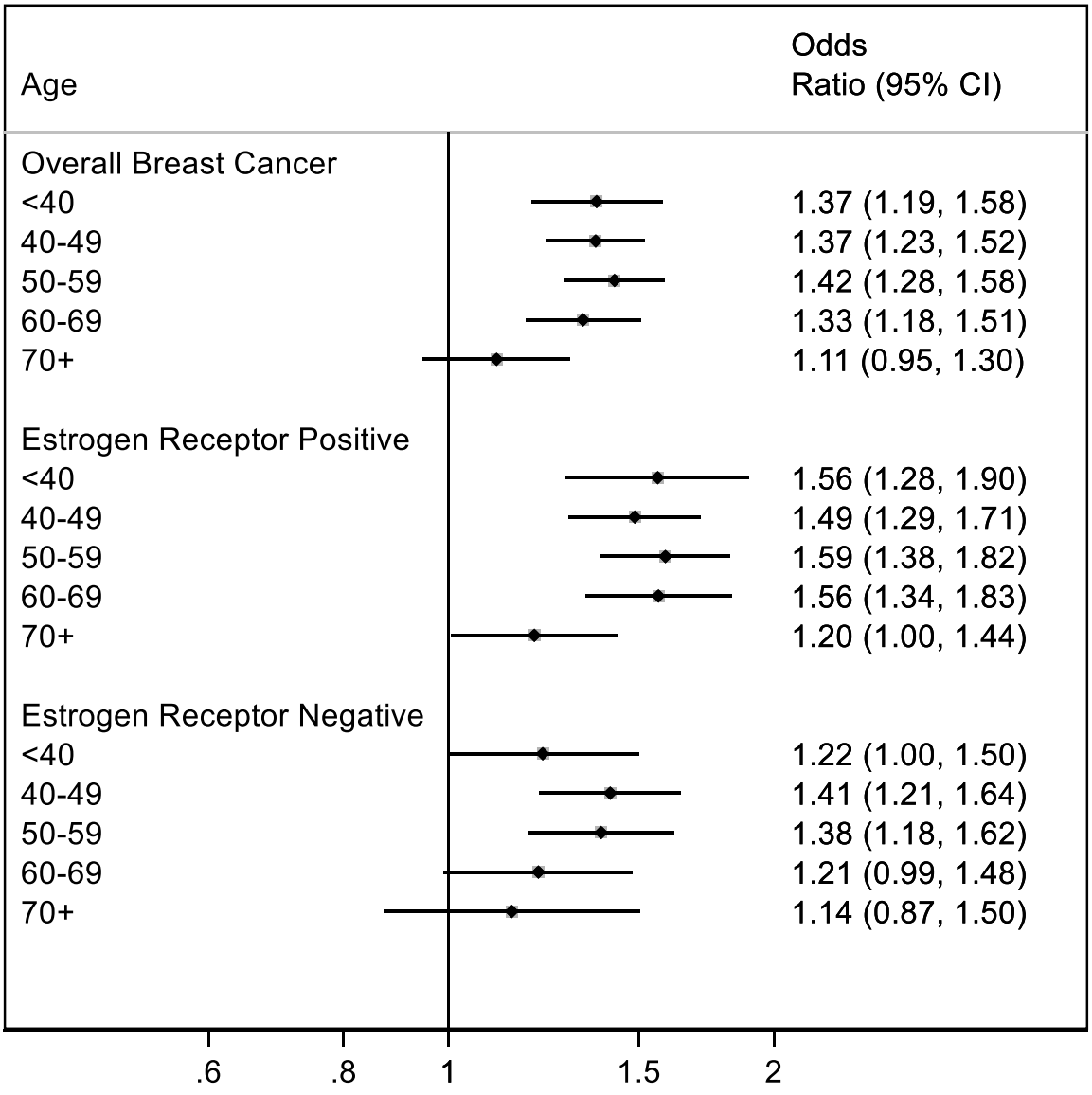


Figure S2. Association of the PRS and breast cancer risk in different consortium/study in the validation set. (The *P* for test for heterogeneity was 0.021 for overall breast cancer risk, 0.52 for ER-positive, and 0.013 for ER-negative breast cancer).

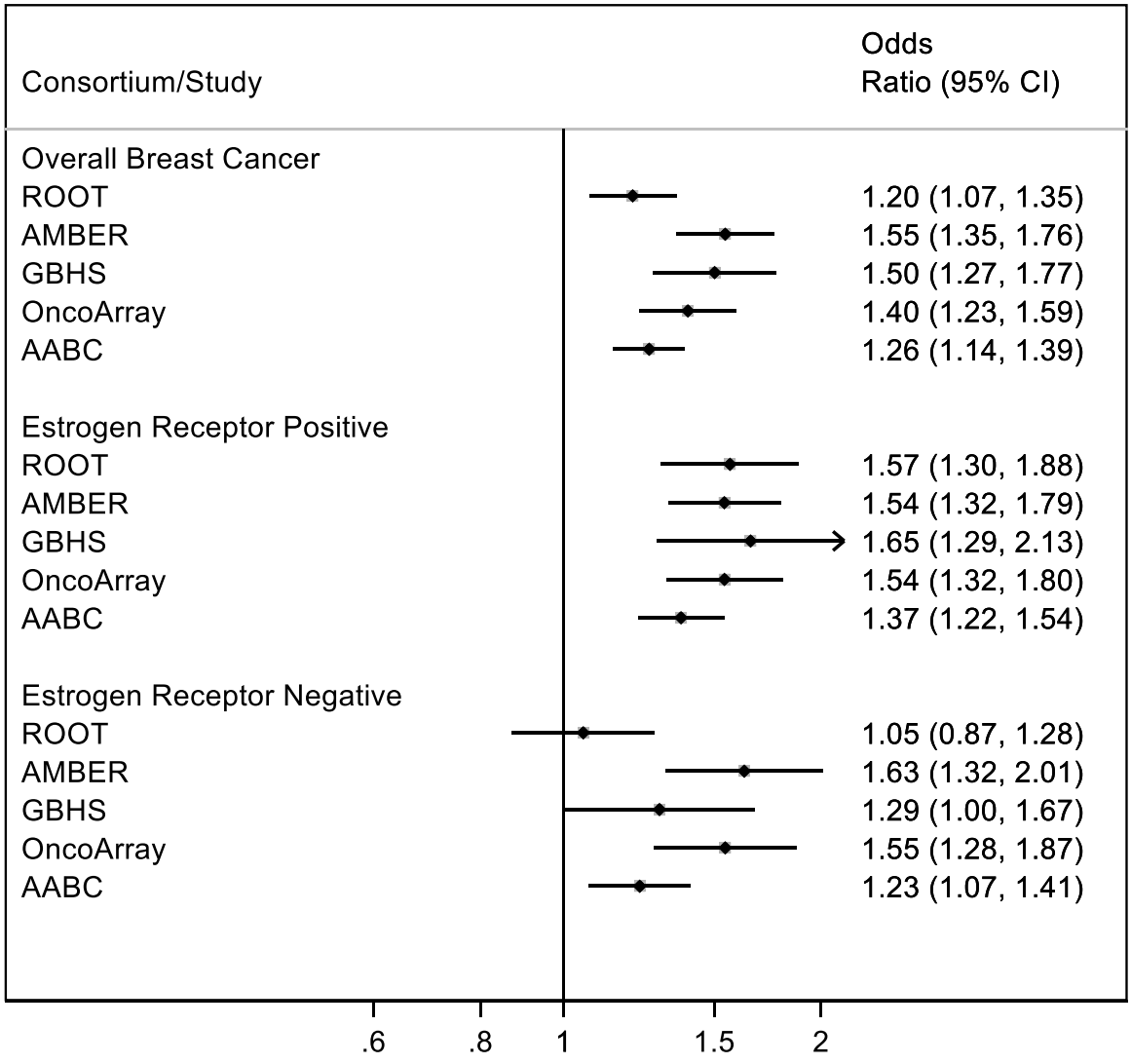
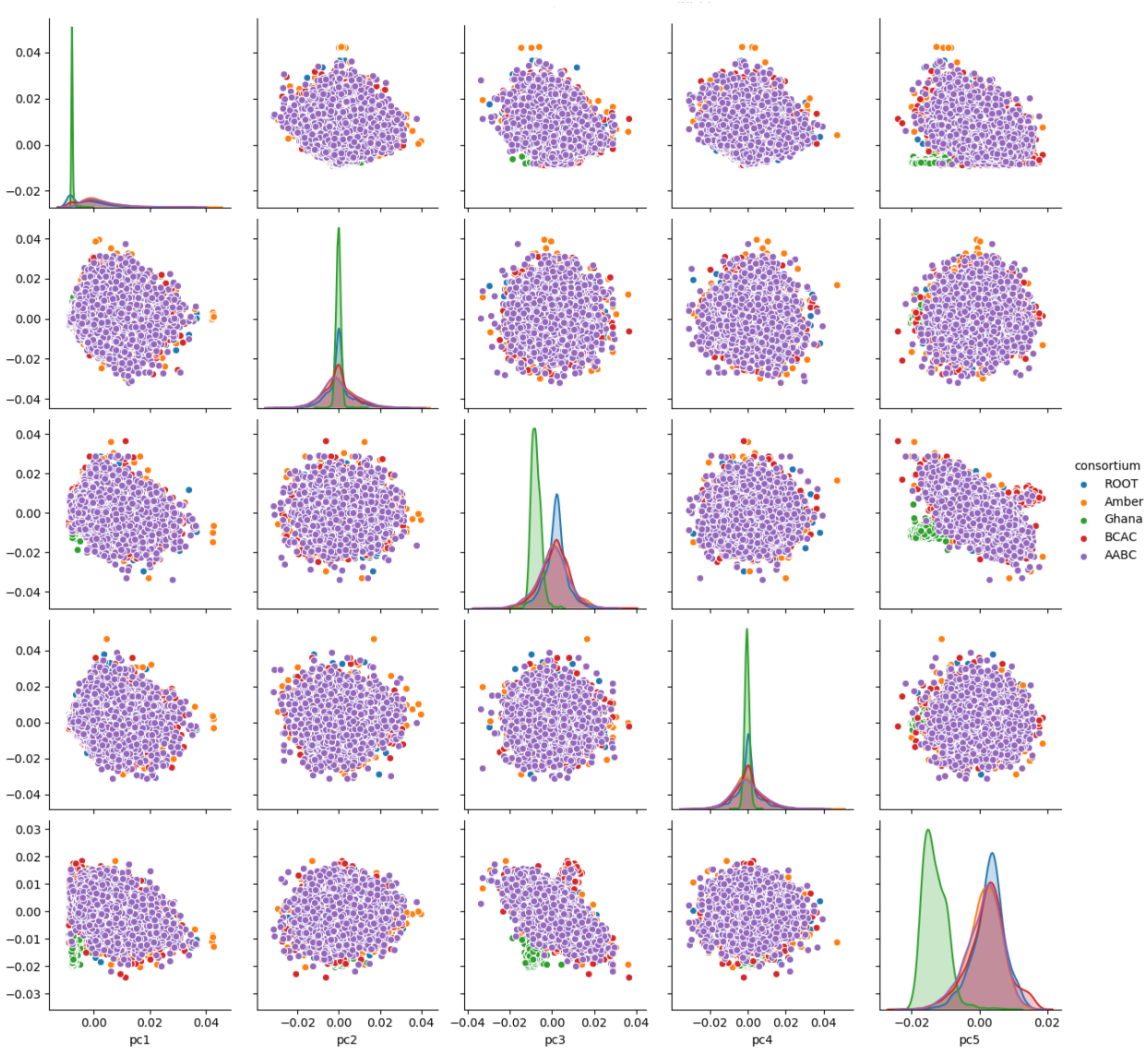


Figure S3. Scatter plots of the top 5 eigenvectors from principal component (PC) analysis according to study consortium/study (A) and country (B).

A



B

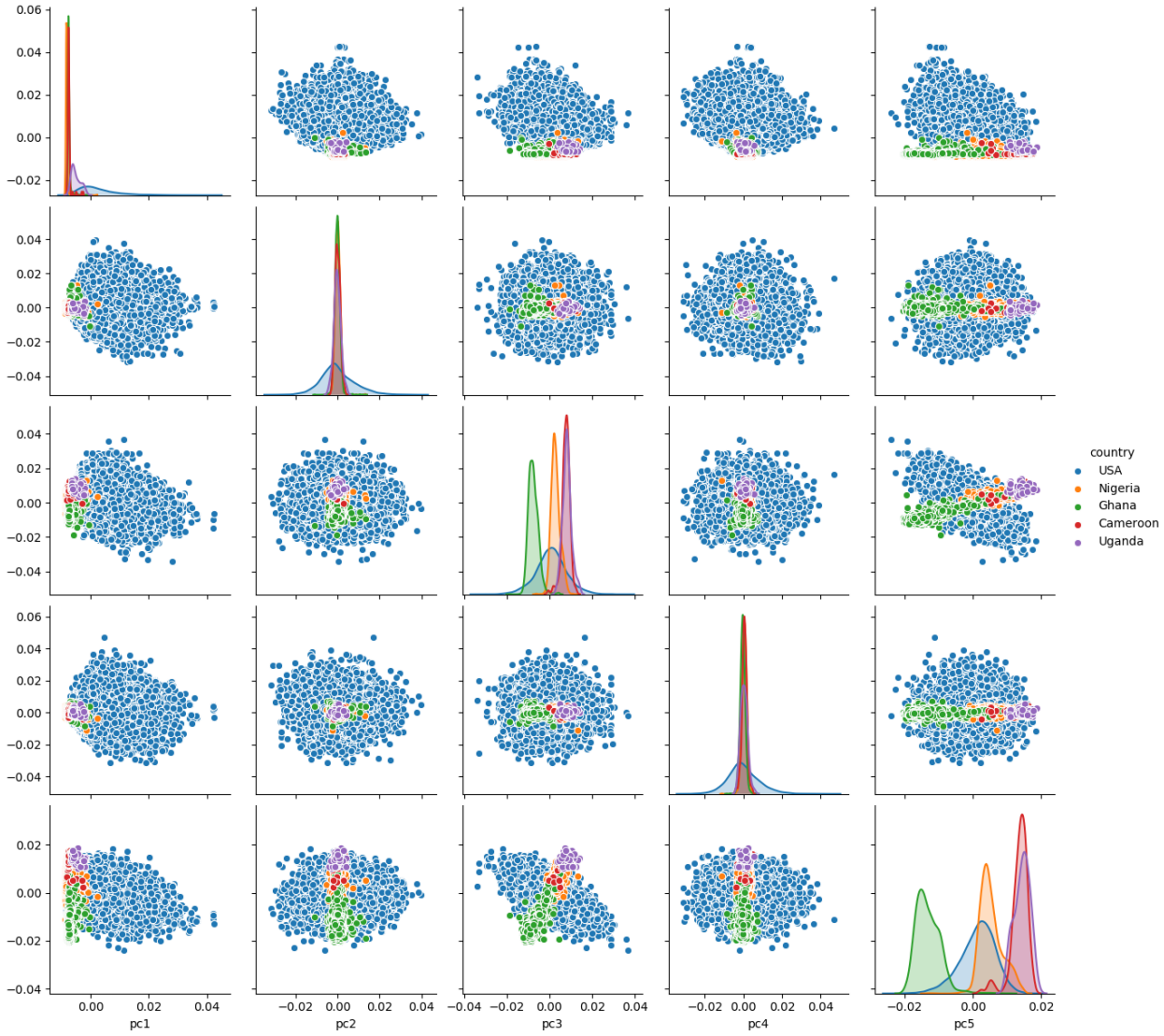
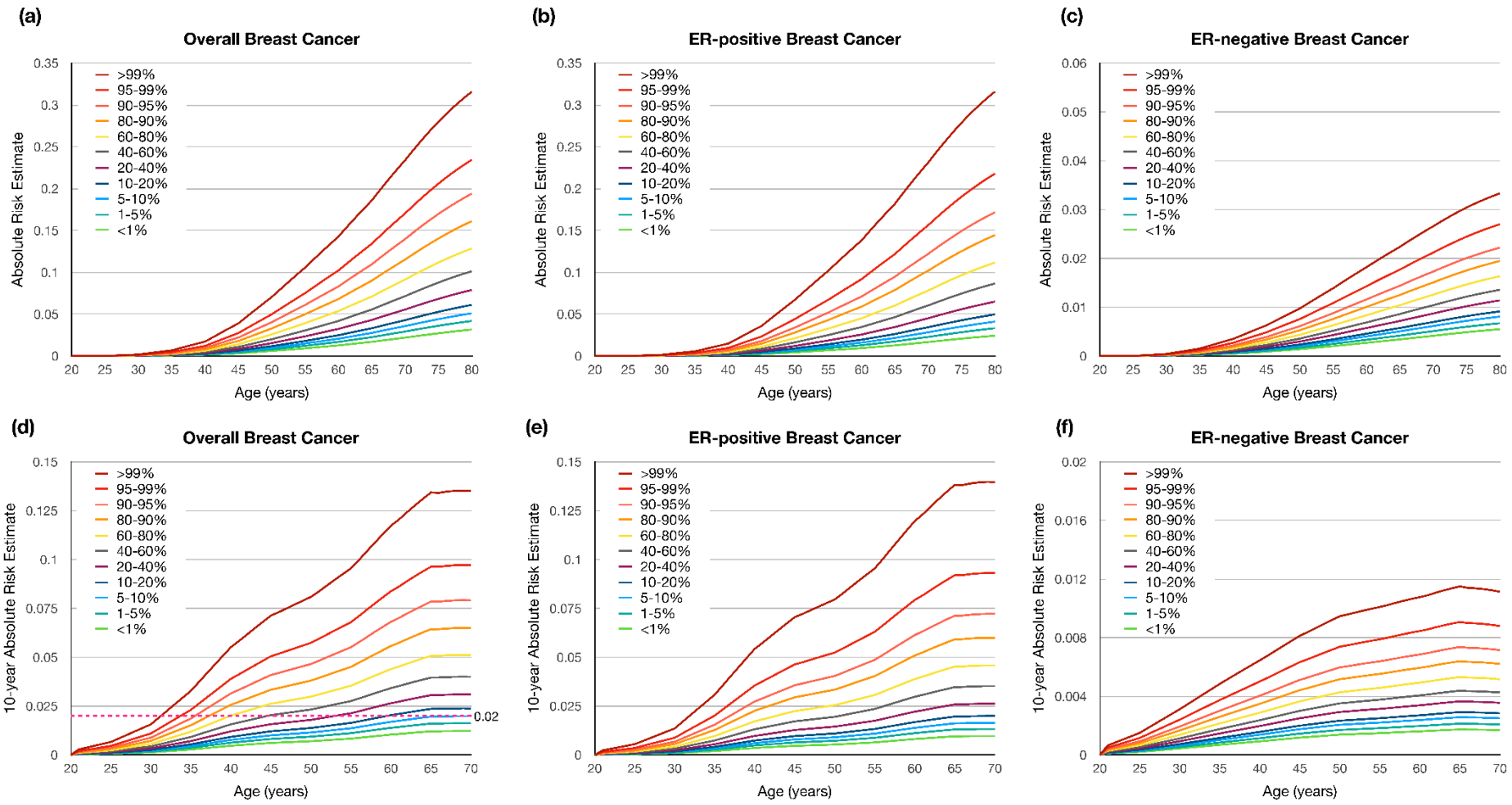

country
USA
Nigeria
Ghana
Cameroon
Uganda
pc1
pc2
pc3
pc4
pc5

Figure S4. Cumulative Life-time and 10-Year Absolute Risk of Developing Breast Cancer among European Americans.



Supplemental Table S1. Descriptive characteristics of analysis population.

Consortium/Study	Total	Controls	Cases	ER positive	ER negative	Mean age (SD), year	Family history of breast cancer, no. (%)	Genotyping platform	Quality control	Number of genotyped variants after initial QC	Imputation R-square after filtering, median (IQR)
ROOT (The GWAS of Breast Cancer in the African Diaspora consortium)	3685	2028	1657	404	374	48.8 (12.7)	439 (13.8)	Illumina HumanOmni 2.5-8v1 array	Individuals were excluded based on call rate (CR) ≤ 98%, ancestry outlier and relativeness. SNPs were removed based on CR <98%, MAF=0, discordant calls, Mendelian errors or with HWE P <10-4. All samples had African ancestry > 12%	1837433	0.993 (0.979 - 0.999)
AMBER (The African American Breast Cancer Epidemiology and Risk consortium)	3813	2407	1406	951	385	51.9 (11.4)	423 (11.1)	Illumina Human Multi-Ethnic Genotyping Array (MEGA)	Individuals were removed based on genotypic sex not female and missing call rate >3%. SNPs with missing call rate >=2%, >1 discordant call in 86 study duplicates, > 1 Mendelian error in 17 HapMap trios, HWE p<1e-4 in a homogenous subset of samples, and MAF <0.01 were excluded prior to imputation. The African ancestry confimation was based on self-reported African Americans or blacks.	1104770	0.983 (0.959 - 0.995)
GBHS (Ghana Breast Health Study)	2529	1630	899	296	277	47.4 (12.7)	N/A	Infinium Global Screening Array-24	Individuals were removed based on a CR < 95%, extreme heterozygosity and relativeness. Variants with CR < 95% were excluded. Deviation from Hardy-Weinberg proportions was assessed and no significant deviation was detected. Ancestry was assessed using the GLU struct.admix module, and participants with <80% African ancestry were excluded.	388652	0.876 (0.795 - 0.937)
BCAC Oncoarray (The BCAC and GAME-ON OncoArray consortium)	3674	1406	2268	1127	613	54.9 (11.5)	783 (23.1)	Illumina Infinium OncoArray-500K BeadChip	Individuals were removed based on CR < 95%, ancestry outlier and relativeness. Variants with CR < 98%, MAF < 1%, not in HWE, concordance < 98% or with poor cluster in visual inspection were excluded. Overlapping samples between ONCO with AABC, AMBER and ROOT were further removed. Genetic ancestry was checked using STRUCTURE and all samples were with African ancestry >20%.	390084	0.953 (0.905 - 0.982)
AABC (African American Breast Cancer consortium)	5718	2713	3005	1517	986	56.6 (12.7)	849 (15.3)	IlluminaHuman1M-Duo BeadChip	Individuals were removed based on ancestry outlier (samples with ≤5% African ancestry), call rate (CR) < 95%, 1st degree relatedness, gender/sex mismatches. SNPs were excluded based on CR < 95%, minor allele frequency (MAF) < 1%, and concordance rate < 98%.	1022969	0.974 (0.938 - 0.993)
Total	19419	10184	9235	4295	2635	52.7 (12.7)	2494 (16.2)				
ROOT Consortium contributing studies											
NBCS - Nigerian Breast Cancer Study	1335	624	711	42	99	46.5 (12.0)	68 (5.1)				
BNCS - Barbados National Cancer Study	321	229	92			55.4 (13.6)	31 (9.7)				
RVGBC - Racial Variability in Genotypic Determinants of Breast Cancer Study	402	257	145	27	25	42.6 (11.9)	59 (14.9)				
CCPS - Chicago Cancer Prone Study	780	386	394	172	140	45.8 (11.6)	178 (50.0)				
BBCS - Baltimore Breast Cancer Study	197	102	95	45	44	52.9 (13.8)	28 (15.6)				
SCCS - Southern Community Cohort Study	650	430	220	118	66	56.8 (9.0)	75 (12.5)				
AABC Consortium contributing studies											
NC-BCFR - Northern California site of the Breast Cancer Family Registry	469	49	420	218	121	50.1 (9.3)	136 (29.0)				
CARE - Women's Contraceptive and Reproductive Experiences Study, Los Angeles component	567	211	356	183	129	48.6 (8.0)	55 (10.1)				
CBCS - Carolina Breast Cancer Study	1222	587	635	272	317	51.5 (11.7)	159 (13.5)				
MEC - Multiethnic Cohort Study	1665	974	691	407	176	67.0 (9.3)	269 (17.1)				
NBHS - Nashville Breast Health Study	483	181	302	142	64	53.2 (10.9)	75 (15.5)				
PLCO - Prostate, Lungs, Colorectal and Ovarian Cancer Screening Trial Cohort	172	116	56	14	6	68.2 (6.3)	17 (10.2)				
SFBCS - San Francisco Bay Area Breast Cancer Study	382	218	164	84	50	55.5 (11.9)	42 (11.0)				
WCHS - Women's Circle of Health Study	497	236	261	131	80	49.8 (9.6)	61 (12.3)				
WFBC - Wake Forest University Breast Cancer Study	261	141	120	66	43	55.1 (11.4)	35 (13.4)				
AMBER Consortium											
BWHS - Black Women's Health Study	2464	2153	311	205	64	51.6 (11.6)	211 (8.6)				
CBCS - Carolina Breast Cancer Study	607	1	606	407	186	51.0 (11.3)	102 (17.0)				
WCHS - Women's Circle of Health Study	742	253	489	339	135	53.5 (11.0)	110 (14.9)				
BCAC (OncoArray) African ancestry											
2SISTER - The Two Sister Study	45	0	45	28	16	45.0 (4.2)	9 (20.9)				
NC-BCFR, The Northern California Breast Cancer Family Registry	80	1	79	40	28	46.9 (10.7)	45 (56.3)				
CBCS - Carolina Breast Cancer Study	974	64	910	503	299	51.5 (10.5)	183 (19.6)				
MEC - Multiethnic Cohort Study	1329	700	629	413	138	60.1 (10.2)	183 (16.5)				
NBHS - The Nashville Breast Health Study	146	59	87	17	40	53.0 (10.9)	18 (12.3)				
PLCO - Prostate, Lungs, Colorectal and Ovarian Cancer Screening Trial Cohort	93	68	25	13	2	65.5 (6.3)	9 (9.9)				
SISTER - The Sister Study	317	166	151	92	27	55.4 (8.8)	272 (88.3)				
USRT - The U.S. Radiologic Technologists	68	39	29			56.7 (9.4)	7 (12.5)				
WAABCS - West African Ancestry Breast Cancer Study	622	309	313	21	63	49.0 (12.2)	57 (9.1)				

Supplemental Table S1. Descriptive characteristics of analysis population.

Consortium/Study	Total	Controls	Cases	ER positive	ER negative	Mean age (SD), year	Family history of breast cancer, no. (%)	Genotyping platform	Quality control	Number of genotyped variants after initial QC	Imputation R-square after filtering, median (IQR)
ROOT (The GWAS of Breast Cancer in the African Diaspora consortium)	3685	2028	1657	404	374	48.8 (12.7)	439 (13.8)	Illumina HumanOmni 2.5-8v1 array	Individuals were excluded based on call rate (CR) ≤ 98%, ancestry outlier and relativeness. SNPs were removed based on CR <98%, MAF=0, discordant calls, Mendelian errors or with HWE P <10-4. All samples had African ancestry > 12%	1837433	0.993 (0.979 - 0.999)
AMBER (The African American Breast Cancer Epidemiology and Risk consortium)	3813	2407	1406	951	385	51.9 (11.4)	423 (11.1)	Illumina Human Multi-Ethnic Genotyping Array (MEGA)	Individuals were removed based on genotypic sex not female and missing call rate >3%. SNPs with missing call rate >=2%, >1 discordant call in 86 study duplicates, > 1 Mendelian error in 17 HapMap trios, HWE p<1e-4 in a homogenous subset of samples, and MAF <0.01 were excluded prior to imputation. The African ancestry confimation was based on self-reported African Americans or blacks.	1104770	0.983 (0.959 - 0.995)
GBHS (Ghana Breast Health Study)	2529	1630	899	296	277	47.4 (12.7)	N/A	Infinium Global Screening Array-24	Individuals were removed based on a CR < 95%, extreme heterozygosity and relativeness. Variants with CR < 95% were excluded. Deviation from Hardy-Weinberg proportions was assessed and no significant deviation was detected. Ancestry was assessed using the GLU struct.admix module, and participants with <80% African ancestry were excluded.	388652	0.876 (0.795 - 0.937)
BCAC Oncoarray (The BCAC and GAME-ON OncoArray consortium)	3674	1406	2268	1127	613	54.9 (11.5)	783 (23.1)	Illumina Infinium OncoArray-500K BeadChip	Individuals were removed based on CR < 95%, ancestry outlier and relativeness. Variants with CR < 98%, MAF < 1%, not in HWE, concordance < 98% or with poor cluster in visual inspection were excluded. Overlapping samples between ONCO with AABC, AMBER and ROOT were further removed. Genetic ancestry was checked using STRUCTURE and all samples were with African ancestry >20%.	390084	0.953 (0.905 - 0.982)
AABC (African American Breast Cancer consortium)	5718	2713	3005	1517	986	56.6 (12.7)	849 (15.3)	IlluminaHuman1M-Duo BeadChip	Individuals were removed based on ancestry outlier (samples with ≤5% African ancestry), call rate (CR) < 95%, 1st degree relatedness, gender/sex mismatches. SNPs were excluded based on CR < 95%, minor allele frequency (MAF) < 1%, and concordance rate < 98%.	1022969	0.974 (0.938 - 0.993)
Total	19419	10184	9235	4295	2635	52.7 (12.7)	2494 (16.2)				
ROOT Consortium contributing studies											
NBCS - Nigerian Breast Cancer Study	1335	624	711	42	99	46.5 (12.0)	68 (5.1)				
BNCS - Barbados National Cancer Study	321	229	92			55.4 (13.6)	31 (9.7)				
RVGBC - Racial Variability in Genotypic Determinants of Breast Cancer Study	402	257	145	27	25	42.6 (11.9)	59 (14.9)				
CCPS - Chicago Cancer Prone Study	780	386	394	172	140	45.8 (11.6)	178 (50.0)				
BBCS - Baltimore Breast Cancer Study	197	102	95	45	44	52.9 (13.8)	28 (15.6)				
SCCS - Southern Community Cohort Study	650	430	220	118	66	56.8 (9.0)	75 (12.5)				
AABC Consortium contributing studies											
NC-BCFR - Northern California site of the Breast Cancer Family Registry	469	49	420	218	121	50.1 (9.3)	136 (29.0)				
CARE - Women's Contraceptive and Reproductive Experiences Study, Los Angeles component	567	211	356	183	129	48.6 (8.0)	55 (10.1)				
CBCS - Carolina Breast Cancer Study	1222	587	635	272	317	51.5 (11.7)	159 (13.5)				
MEC - Multiethnic Cohort Study	1665	974	691	407	176	67.0 (9.3)	269 (17.1)				
NBHS - Nashville Breast Health Study	483	181	302	142	64	53.2 (10.9)	75 (15.5)				
PLCO - Prostate, Lungs, Colorectal and Ovarian Cancer Screening Trial Cohort	172	116	56	14	6	68.2 (6.3)	17 (10.2)				
SFBCS - San Francisco Bay Area Breast Cancer Study	382	218	164	84	50	55.5 (11.9)	42 (11.0)				
WCHS - Women's Circle of Health Study	497	236	261	131	80	49.8 (9.6)	61 (12.3)				
WFBC - Wake Forest University Breast Cancer Study	261	141	120	66	43	55.1 (11.4)	35 (13.4)				
AMBER Consortium											
BWHS - Black Women's Health Study	2464	2153	311	205	64	51.6 (11.6)	211 (8.6)				
CBCS - Carolina Breast Cancer Study	607	1	606	407	186	51.0 (11.3)	102 (17.0)				
WCHS - Women's Circle of Health Study	742	253	489	339	135	53.5 (11.0)	110 (14.9)				
BCAC (OncoArray) African ancestry											
2SISTER - The Two Sister Study	45	0	45	28	16	45.0 (4.2)	9 (20.9)				
NC-BCFR, The Northern California Breast Cancer Family Registry	80	1	79	40	28	46.9 (10.7)	45 (56.3)				
CBCS - Carolina Breast Cancer Study	974	64	910	503	299	51.5 (10.5)	183 (19.6)				
MEC - Multiethnic Cohort Study	1329	700	629	413	138	60.1 (10.2)	183 (16.5)				
NBHS - The Nashville Breast Health Study	146	59	87	17	40	53.0 (10.9)	18 (12.3)				
PLCO - Prostate, Lungs, Colorectal and Ovarian Cancer Screening Trial Cohort	93	68	25	13	2	65.5 (6.3)	9 (9.9)				
SISTER - The Sister Study	317	166	151	92	27	55.4 (8.8)	272 (88.3)				
USRT - The U.S. Radiologic Technologists	68	39	29			56.7 (9.4)	7 (12.5)				
WAABCS - West African Ancestry Breast Cancer Study	622	309	313	21	63	49.0 (12.2)	57 (9.1)				

Supplemental Table S3. Correlation coefficients among selected PRSs in the validation set

Overall Breast Cancer						
	PRS_{EUR307}	PRS_{AFR62}	PRS_{AFR428}	$PRS_{AFR2351}$	$PRS_{AFR10647}$	$PRS_{AFR29569}$
Threshold P< 1E-5, PRS_{AFR62}*	0.075					
Threshold P< 1E-4, PRS_{AFR428}	0.034	0.451				
Threshold P< 1E-3, $PRS_{AFR2351}$	0.017	0.252	0.512			
Threshold P< 1E-2, $PRS_{AFR10647}$	0.034	0.141	0.326	0.596		
Threshold P< 5E-2, $PRS_{AFR29569}$	**0.031**	0.113	0.265	0.469	0.733	
Threshold P< 1E-1, $PRS_{AFR46854}$	0.034	0.111	0.250	0.433	0.668	0.858
ER-positive						
	$PRS_{EUR307.ERp}$	$PRS_{AFR79.ERp}$	$PRS_{AFR408.E}$	$PRS_{AFR2339.ERp}$	$PRS_{AFR10493.ERp}$	$PRS_{AFR29004.ERp}$
Threshold P< 1E-5, $PRS_{AFR79.ERp}$	0.024					
Threshold P< 1E-4, $PRS_{AFR408.ERp}$	0.019	0.462				
Threshold P< 1E-3, $PRS_{AFR2339.ERp}$	0.012	0.247	0.507			
Threshold P< 1E-2, $PRS_{AFR10493.ERp}$	-0.018	0.194	0.326	0.598		
Threshold P< 5E-2, $PRS_{AFR29004.ERp}$	**-0.017**	0.178	0.248	0.448	0.756	
Threshold P< 1E-1, $PRS_{AFR45997.ERp}$	-0.007	0.168	0.221	0.410	0.705	0.887
ER-negative						
	$PRS_{EUR307.ERn}$	$PRS_{AFR50.ERn}$	$PRS_{AFR419.E}$	$PRS_{AFR2230.ERn}$	$PRS_{AFR10138.ERn}$	$PRS_{AFR28100.ERn}$
Threshold P< 1E-5, $PRS_{AFR50.ERn}$	0.072					
Threshold P< 1E-4, $PRS_{AFR419.ERn}$	0.016	0.434				
Threshold P< 1E-3, $PRS_{AFR2230.ERn}$	0.010	0.221	0.494			
Threshold P< 1E-2, $PRS_{AFR10138.ERn}$	0.005	0.110	0.274	0.556		
Threshold P< 5E-2, $PRS_{AFR28100.ERn}$	**-0.004**	0.075	0.205	0.410	0.707	
Threshold P< 1E-1, $PRS_{AFR44889.ERn}$	-0.007	0.069	0.184	0.361	0.639	0.852

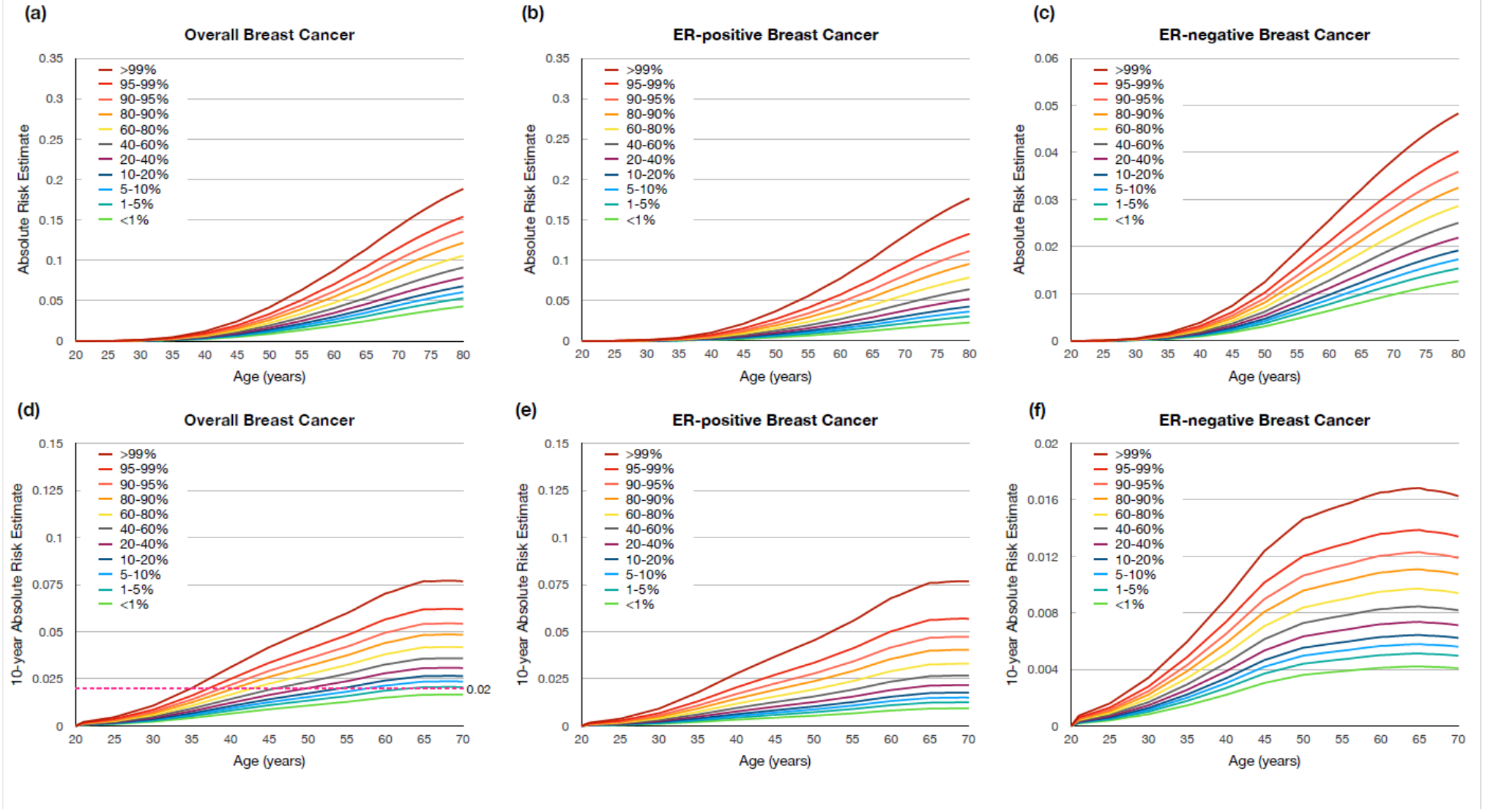
*PRS_{AFR62} denotes PRS_{AFR} using 62 SNPs selected by the threshold

Supplemental Table S7. Associations between polygenic risk scores (PRS) and breast cancer risk by populations the validation set

	No.	Overall Breast Cancer	ER-positive	ER-negative
Stratified by Population				
PRS in Africans	1361			
Mean (SD) in controls	792	0.348 (0.276)	-0.118 (0.378)	0.780 (0.253)
Mean (SD) in cases	569	0.399 (0.308)	0.138 (0.403)	0.819 (0.290)
OR (95% CI)*		1.34 (1.19-1.51)	1.60 (1.28-2.00)	1.18 (0.97-1.45)
PRS in African Americans & African Barbadians	4460			
Mean (SD) in controls	2327	0.412 (0.293)	0.069 (0.394)	0.793 (0.267)
Mean (SD) in cases	2133	0.499 (0.299)	0.210 (0.401)	0.867 (0.275)
OR (95% CI)*		1.28 (1.20-1.37)	1.37 (1.27-1.49)	1.23 (1.12-1.36)
P for testing interaction		*0.54*	*0.211*	*0.719*
Stratified by Percent of African Ancestry				
> 80% African Ancestry	3947			
Mean (SD) in controls	2116	0.395 (0.282)	0.034 (0.380)	0.796 (0.259)
Mean (SD) in cases	1831	0.473 (0.304)	0.213 (0.393)	0.847 (0.281)
OR (95% CI)*		1.30 (1.21-1.40)	1.47 (1.33-1.62)	1.13 (1.02-1.26)
< 80% African Ancestry	1874			
Mean (SD) in controls	1003	0.398 (0.308)	0.056 (0.414)	0.778 (0.274)
Mean (SD) in cases	871	0.489 (0.301)	0.188 (0.416)	0.882 (0.270)
OR (95% CI)*		1.29 (1.17-1.43)	1.32 (1.17-1.49)	1.44 (1.23-1.67)
P for testing interaction		*0.84*	*0.145*	*0.017*

* Odds ratio (95% confidence intervals) per 1 SD, adjusted for studies

1.371913
1.537085


(a)
Overall Breast Cancer
Absolute Risk Estimate
Age (years)
(b)
ER-positive Breast Cancer
Absolute Risk Estimate
Age (years)
(c)
ER-negative Breast Cancer
Absolute Risk Estimate
Age (years)
(d)
Overall Breast Cancer
10-year Absolute Risk Estimate
0.02
Age (years)
(e)
ER-positive Breast Cancer
10-year Absolute Risk Estimate
Age (years)
(f)
ER-negative Breast Cancer
10-year Absolute Risk Estimate
Age (years)
>99%
95-99%
90-95%
80-90%
60-80%
40-60%
20-40%
10-20%
5-10%
1-5%
<1%