medRxiv preprint doi: https://doi.org/10.1101/2021.09.23.21263888; this version posted September 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Incorporating Polygenic Risk Scores and Nongenetic Risk Factors for Breast Cancer Risk Prediction among Asian Women, Results from Asia Breast Cancer Consortium

Yaohua Yang, PhD[1]; Ran Tao, PhD[2]; Xiang Shu, PhD[3]; Qiuyin Cai, PhD[1]; Wanqing Wen, PhD[1]; Kai Gu, MD[4]; Yu-Tang Gao, PhD[5]; Ying Zheng, PhD[6]; Sun-Seog Kweon, PhD[7,8]; Min-Ho Shin, PhD[7]; Ji-Yeob Choi, PhD[9,10,11]; Eun-Sook Lee, PhD[12,13,14]; Sun-Young Kong, PhD[12,13,14]; Boyoung Park, PhD[14,15]; Min Ho Park, PhD[16]; Guochong Jia, MPH[1]; Bingshan Li, PhD[17]; Daehee Kang, MD, PhD[10,11,18,19]; Xiao-Ou Shu, MD, PhD [1]; Jirong Long, PhD[1]; Wei Zheng, MD, PhD[1*]

[1] Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, TN, USA

[2] Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, USA

[3] Department of Epidemiology & Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA

[4] Shanghai Municipal Center for Disease Control and Prevention, Shanghai Institutes of Preventive Medicine, Shanghai, China.

[5] State Key Laboratory of Oncogene and Related Genes & Department of Epidemiology, Shanghai Cancer Institute, Renji Hospital, Shanghai Jiaotong University School of Medicine, Shanghai, China

[6] Department of Cancer Prevention, Fudan University Shanghai Cancer Center, Shanghai, China

[7] Department of Preventive Medicine, Chonnam National University Medical School, Hwasun, South Korea

[8] Jeonnam Regional Cancer Center, Chonnam National University Hwasun Hospital, Hwasun, South Korea

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.09.23.21263888; this version posted September 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

[9] Department of Biomedical Sciences, Seoul National University College of Medicine, Seoul, South Korea

[10] Department of Preventive Medicine, Seoul National University College of Medicine, Seoul, South Korea

[11] Cancer Research Institute, Seoul National University College of Medicine, Seoul, South Korea

[12] National Cancer Center Graduate School of Cancer Science and Policy, Goyang, South Korea

[13] Hospital, National Cancer Center, Goyang, South Korea

[14] Research Institute, National Cancer Center, Goyang, South Korea

[15] Department of Preventive Medicine, Hanyang University College of Medicine, Seoul, South Korea

[16] Department of Surgery, Chonnam National University Medical School & Hospital, Hwasun, South Korea

[17] Department of Molecular Physiology & Biophysics, Vanderbilt Genetics Institute, Vanderbilt University, Nashville, TN, USA

[18] Department of Biomedical Sciences, Seoul National University Graduate School, Seoul, South Korea

[19] Institute of Environmental Medicine, Seoul National University Medical Research Center, Seoul, South Korea

*** Correspondence to**: Dr. Wei Zheng, M.D., Ph.D., Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center, 2525 West End Ave, Suite 800, Nashville, TN, 37203. Tel: (615) 936-0682. Email: wei.zheng@vanderbilt.edu.

Word count: 2991

medRxiv preprint doi: https://doi.org/10.1101/2021.09.23.21263888; this version posted September 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Key Points

Question: What is the performance of breast cancer risk prediction models for Asian women incorporating polygenic risk scores (PRSs) and nongenetic risk factors?

Findings: A 111-genetic-variant PRS developed using data of 125 790 Asian women was significantly associated with breast cancer risk in an independent case-control study nested within a prospective cohort, with an odd ratio (OR) per standard deviation increase of 1.67 (95% confidence interval [CI]=1.46-1.92) and an area under the receiver operating characteristic curve (AUC) of 0.639 (95% CI=0.604-0.674). The prediction model including this PRS and six nongenetic risk factors improved the AUC to 0.650 (95% CI=0.616-0.685).

Meaning: Our study provides strong supports for the utility of prediction models in identifying Asian women at high risk of breast cancer.

medRxiv preprint doi: https://doi.org/10.1101/2021.09.23.21263888; this version posted September 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Abstract

Importance Polygenic risk scores (PRSs) have shown promises in breast cancer risk prediction; however, limited studies have been conducted among Asian women.

Objective To develop breast cancer risk prediction models for Asian women incorporating PRSs and nongenetic risk factors.

Design PRSs were developed using data from genome-wide association studies (GWAS) of breast cancer conducted among 123 041 Asian-ancestry women (including 18 650 cases) using three approaches (1) reported PRS for European-ancestry women; (2) breast cancer-associated single-nucleotide polymorphisms (SNPs) identified by fine-mapping of GWAS-identified risk loci; (3) genome-wide risk prediction algorithms. A nongenetic risk score (NgRS) was built including six well-established nongenetic risk factors using data from 1974 Asian women. Integrated risk scores (IRSs) were constructed using PRSs and the NgRS. PRSs were initially validated in an independent dataset including 1426 cases and 1323 controls and further evaluated, along with the NgRS and IRSs, in the second dataset including 368 cases and 736 controls nested withing a prospective cohort study.

Setting Case-control and prospective cohort studies.

Participants 20 444 breast cancer cases and 106 450 controls from the Asia Breast Cancer Consortium.

Main Outcomes and Measures Logistic regression was used to examine associations of risk scores with breast cancer risk to estimate odds ratios (ORs) with 95% confidence intervals (CIs) and area under the receiver operating characteristic curve (AUC).

Results In the prospective cohort, $PRS_{111,}$ a PRS with 111 SNPs, developed using the fine-mapping approach showed a prediction performance comparable to a genome-wide PRS

medRxiv preprint doi: https://doi.org/10.1101/2021.09.23.21263888; this version posted September 26, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

including over 855,000 SNPs. The OR per standard deviation increase of PRS_{111} was 1.67 (95% CI=1.46-1.92) with an AUC of 0.639 (95% CI=0.604-0.674). The NgRS had a limited predictive ability (AUC=0.565; 95% CI=0.529-0.601); while IRS_{111}, the combination of PRS_{111} and NgRS, achieved the highest prediction accuracy (AUC=0.650; 95% CI=0.616-0.685). Compared with the average risk group (40th-60th percentile), women in the top 5% of PRS_{111} and IRS_{111} were at a 3.84-folded (95% CI=2.30-6.46) and 4.25- folded (95% CI=2.57-7.11) elevated risk of breast cancer, respectively.

Conclusions and Relevance PRSs derived using breast cancer-associated risk SNPs have similar prediction performance in Asian and European descendants. Including nongenetic risk factors in models further improved prediction accuracy. Our findings support the utility of these models in developing personalized screening and prevention strategies.

medRxiv preprint doi: https://doi.org/10.1101/2021.09.23.21263888; this version posted September 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Introduction

Breast cancer is the most commonly diagnosed malignancy among women worldwide.[1] The incident rate of breast cancer has been increasing substantially in many Asian countries, although the overall rate is still significantly lower than those seen in the U.S. and many European countries.[2] Currently, many Asian countries do not have a population-based breast cancer screening program, leading to delayed cancer diagnoses and poor survival rates.[3] Because of the differences in breast cancer risk, screening programs currently implemented in the U.S. and European countries may not be appropriate for Asian countries. Hence, a cost-efficient, population-specific breast cancer screening strategy for Asian women is imminently needed.

In 2006, we established the Asia Breast Cancer Consortium (ABCC) to identify single nucleotide polymorphisms (SNPs) associated with breast cancer risk through genome-wide association studies (GWAS). To date, approximately 50 risk loci were identified in our studies using Asian data alone or meta-analyses of data from both Asian and European descendants.[4-12] However, most of breast cancer risk loci were identified in GWAS conducted in European descendants.[13] Multiple studies have attempted to aggregate effects of SNPs identified by GWAS as polygenic risk scores (PRSs) to stratify women into different breast cancer risk groups.[14-18] The vast majority of PRSs for breast cancer were established specifically in women of European ancestry. Among them, a 313-SNP PRS showed the highest predictive ability, with an area under the receiver operating characteristic (ROC) curve (AUC) of 0.630 to 0.642.[14] Few studies of breast cancer PRSs for Asian women were conducted and limited prediction accuracy was observed.[19-23] A recent study showed that the 313-SNP PRS performed better than PRSs derived from Asian

medRxiv preprint doi: https://doi.org/10.1101/2021.09.23.21263888; this version posted September 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

data.[23] However, in that study, the sample size was relatively small and the Asian-specific PRSs included limited number of SNPs.

In addition to genetic variations, nongenetic factors are also associated with breast cancer risk.[24] Several studies have explored the potential of incorporating PRSs and nongenetic risk factors in improving the prediction accuracy.[18,24-27] Among them, a recent study among European women revealed that the 313-SNP PRS was more predictive than a model including 16 nongenetic risk factors, and the best risk stratification performance was achieved when PRS and nongenetic factors were combined.[24] However, similar studies have rarely been carried out among Asian women. In the present study, we aimed to develop and validate PRSs for Asian women using data from the largest GWAS of breast cancer ever conducted among Asian women and further evaluate the performance of risk prediction models including both PRSs and known nongenetic risk factors.

Methods

Study Participants

As shown in **Table 1**, the PRS development datasets included GWAS data of 20 076 breast cancer cases and 105 714 controls of Asian ancestry from the ABCC. Detailed information on the ABCC is described elsewhere.[11] We divided these datasets to a training set (18 650 cases and 104 391 controls) for PRS derivation and a validation set (1426 cases and 1323 controls) for prediction performance evaluation (**eMethods** in the **Supplement**). For each PRS development approach, the most predictive PRS in our validation set were further evaluated in an independent case-control study nested within a prospective cohort study, comprising 368 cases and 736

medRxiv preprint doi: https://doi.org/10.1101/2021.09.23.21263888; this version posted September 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

individually matched controls by age (<5 years). Included in the nested case-control study were participants from the Shanghai Women's Health Study (SWHS) and none of them had a diagnosis of any cancers at the time of enrollment (**eMethods** in the **Supplement**).[11,28] All studies involved in the current analyses have been approved by their respective Institutional Review Boards.

Genetic Data

Detailed descriptions of genetic data are provided in our recent publication and **eMethods**.[11] Genotyping was conducted using several platforms and genotyping data imputation was performed separately by study (**eTable 1**). After quality controls and imputation, 5 947 015 SNPs were included in our analyses. GWAS was conducted within each study/sub-study and association results were combined via fixed-effects meta-analyses.

PRS Development

We applied three approaches to develop PRSs as described below briefly and in detail in **eMethods**. PRSs were calculated using the formula: $PRS = \sum_{k=0}^{n} \beta_k SNP_k$, in which SNP_k and β_k represent the allelic dosage and corresponding weight of $SNP\ k$, and n is the number of SNPs used.

Reported PRS for European-ancestry Women

The best breast cancer PRS to date was the 313-SNP PRS among European-ancestry women.[14] Most recently, this PRS was updated by adding 17 novel breast cancer susceptibility SNPs.[13] Of these 330 SNPs, 263 could be found in our validation set and prospective test set, and three PRSs

medRxiv preprint doi: https://doi.org/10.1101/2021.09.23.21263888; this version posted September 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

($PRS_{263\text{-}ASN}$, $PRS_{263\text{-}EUR}$ and $PRS_{263\text{-}META}$) were derived using weights of these SNPs from data from European-ancestry women included in the Breast Cancer Association Consortium (BCAC-European),[13] data from Asian-ancestry women in our training set (**Table 1**), and meta-analyses of these two datasets, respectively.

PRSs based on SNPs selected from fine-mapping of GWAS-identified risk loci

The overall workflow of this approach is presented in **eFigure 1**. For each of the 238 independent susceptibility loci for breast cancer, [4-12,29,30] fine-mapping analyses were performed using summary statistics of our training set to identify SNPs that were independently associated with breast cancer risk using GCTA-COJO.[31,32] Within each locus, a COJO-*P* threshold of 10^{-5} was used to identify independently associated risk SNPs and re-estimate weights of them on breast cancer for PRS construction. Some loci were ineligible for fine-mapping because no SNPs within them had an association with breast cancer risk at $P<10^{-5}$ in our training set. Based on fine-mapping results, three PRSs were derived using (1) all SNPs selected from fine-mapping; (2) SNPs selected by fine-mapping and showing consistent association directions with $P<.05$ in the BCAC-European data; (3) SNPs in (2), plus lead SNPs from loci that were ineligible for fine-mapping but showed $P<.05$ in our training set (**eFigure 1**). We repeated the fine-mapping analyses using COJO-*P* thresholds of 10^{-3} and 10^{-4} to identify independent risk SNPs and used them to construct three sets of PRSs for each threshold following the same steps described above.

PRSs based on genome-wide risk prediction algorithms

LDpred, LDpred2, and PRS-CSx were used to derive genome-wide PRSs using data from the training set. The detailed description of these three algorithms can be obtained elsewhere.[33-35] Of

medRxiv preprint doi: https://doi.org/10.1101/2021.09.23.21263888; this version posted September 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

the 5 947 015 SNPs, indels and ambiguous SNPs were excluded by LDpred, and weights of the remaining 4 487 284 SNPs with breast cancer risk were re-evaluated. Both LDpred2 and PRS-CSx recommends using SNPs included in HapMap 3; thus, of the 5 948 258 SNPs, weights of only 855 680 HapMap 3 SNPs were re-estimated using each of these two algorithms (**eMethods** in **Supplements**).

Models Incorporating PRSs and Nongenetic Risk Factors

Established nongenetic breast cancer risk factors included body mass index (BMI), waist-to-hip ratio (WHR), a prior diagnosis of benign breast disease, age at menarche, age at first live birth, and family history of breast cancer. An interaction term between BMI and menopause status was included in the model as BMI shows a different association with breast cancer risk by menopausal status.[19] Data of 1974 women from the SWHS but independent from those in the prospective test set were used to estimate the weights of these six nongenetic factors and the interaction term on breast cancer risk (**eTable 2**). A logistic regression model was fitted with case/control status of breast cancer as the outcome and these eight factors as predictors. Weights estimated from this model were then used to construct a nongenetic risk score (NgRS) for each subject using the following formula: $NgRS = \sum_{k=0}^{7} w_k F_k \ + \ w_i BMI * Menopause$, where F_k and w_k are the value and corresponding weight of factor k, and w_i is the weight of the interaction term between BMI and menopause status.

Prediction Performance Evaluation

PRSs derived from the training set were first evaluated for their associations with breast cancer risk and prediction performance in our case-control validation set. Then, the most predictive

medRxiv preprint doi: https://doi.org/10.1101/2021.09.23.21263888; this version posted September 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

PRSs from each PRS development approach were further appraised in our prospective test set. Finally, for the PRS showing the highest prediction accuracy in our prospective test set, and the reported European PRS,[14] an integrated risk score (IRS) was built through incorporating each PRS with the NgRS using this formula: $IRS = PRS + NgRS$. Logistic regression was used to evaluate ORs and 95% confidence intervals (CIs) per standard deviation (SD) increase in these risk scores. Prediction performance was measured by AUCs and 95% CIs using the R function *pROC:roc*.[36] We also investigated the utility of these scores in classifying participants with two- and three-fold increased risk compared to the average risk group (40th-60th percentiles), through logistic regression analyses.

Absolute Risk of Developing Breast Cancer According to PRS/IRS Percentiles

We estimated the 10-year absolute risk of developing breast cancer using the most predictive PRS in our prospective test set and the reported European PRS,[14] and their corresponding IRSs. Considering that the prospective test set has a limited sample size (*N*=1104), which would lead to unstable OR estimates, 10 207 Chinese women (5087 cases and 5120 controls) from the whole ABCC datasets with both genetic and nongenetic data available were also included in this analysis. Logistic regression was used to estimate breast cancer ORs of different PRS/IRS percentile groups compared to the middle quintile (40%-60%) group. Then 10-year absolute risks were calculated utilizing these ORs and the incidence and mortality rates of breast cancer in Shanghai in 2017 following the strategy described previously.[19]

Results

Prediction Performance of PRSs

medRxiv preprint doi: https://doi.org/10.1101/2021.09.23.21263888; this version posted September 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

The three PRSs derived based on the reported European PRS[14] had similar prediction performance in our case-control validation set (**Table 2; eTable 3; eTable 4**). However, in our prospective test set, $PRS_{263\text{-}META}$ had the best prediction accuracy (AUC=0.626, 95% CI=0.592-0.661) (**Table 2**). The OR of breast cancer per SD increase of $PRS_{263\text{-}META}$ was 1.63 (95% CI=1.43-1.87).

To identify SNPs more specifically associated with breast cancer risk in Asian women and derive Asian-specific PRS, we performed fine-mapping analyses. At each fine-mapping threshold, three PRSs were developed (**eTable 3**) and among them, PRS_{111} showed the strongest association with breast cancer risk as well as highest prediction performance in both validation and prospective test sets (**Table 2**). This PRS was developed using 57 SNPs selected by fine-mapping and showing consistent association directions with $P<.05$ in the BCAC-European data,[13] plus 54 lead SNPs in GWAS loci with $P<.05$ in our training set (**eFigure 1; eTable 5**). The OR for breast cancer per SD increase in PRS_{111} was 1.45 (1.34-1.57) and 1.67 (1.46-1.92), with AUCs of 0.603 (95% CI=0.582-0.624) and 0.639 (95% CI=0.604-0.674), in our case-control validation set and prospective test set, respectively (**Table 2**). Compared to the average risk group (40th-60th percentile), women in the top 5% of PRS_{111} were at 3.84-folded (95% CI=2.30-6.46) increased risk of breast cancer. As shown in **Figures 1**. For both PRS_{111} and $PRS_{263\text{-}META}$, distribution curves for cases were shifted to the right compared to those for controls, and the overlap was less for PRS_{111} than $PRS_{263\text{-}META}$ (**Figures 1A** and **Figure 1C**). The difference in median percentile between cases and controls (64 vs 43) was higher for PRS_{111}, compared to $PRS_{263\text{-}META}$ (60 vs 44) (**Figures 1B** and **Figure 1D**).

medRxiv preprint doi: https://doi.org/10.1101/2021.09.23.21263888; this version posted September 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

The most predictive PRS derived by each genome-wide risk prediction algorithm in our case-control validation set are shown in **Table 2**. Among them, $PRS_{PRS\text{-}CSx}$ was the best-performing PRS not only in our case-control validation set, but also in our prospective test set with AUC (95% CI) of 0.613 (0.592-0.634) and 0.642 (0.608-0.676), respectively (**Table 2; eTable 3**). The OR (95% CI) for breast cancer per SD increase of this PRS in these three datasets was 1.51 (1.39-1.63) and 1.70 (1.49-1.95) respectively. Although in our prospective test set, $PRS_{PRS\text{-}CSx}$ performed slightly better than PRS_{111} (AUC: 0.642 vs 0.639), we chose PRS_{111} as the best PRS of the present study and used it in downstream analyses because compared to $PRS_{PRS\text{-}CSx}$, PRS_{111} used much fewer SNPs (111 vs 855 680) but had almost equal predictive ability.

Prediction Performance of NgRS and IRSs

In our prospective test set, the NgRS was associated with breast cancer risk with an OR per SD increase of 1.29 (95% CI=1.14-1.46) with an AUC of 0.565 (95% CI=0.529-0.601) (**Table 3**). Incorporating this NgRS with PRS_{111} or $PRS_{263\text{-}META}$, we created IRS_{111} and $IRS_{263\text{-}META}$, respectively. In our prospective test set, IRS_{111} showed a better prediction accuracy (AUC=0.650; 95% CI=0.616-0.685; OR=1.77; 95%CI=1.55-2.04) than $IRS_{263\text{-}META}$ (AUC=0.639; 95% CI=0.605-0.673; OR=1.72; 95% CI=1.51-1.98) **(Table 3)**. Compared to the average risk group, women in the top 5% of IRS_{111} and $IRS_{263\text{-}META}$ were at a 4.25-folded (95% CI=2.57-7.11) and 2.79-folded (95% CI=1.70-4.63) elevated risk of breast cancer, respectively. Among all risk scores developed in the present study, IRS_{111} had the best risk stratification capability in our prospective test set. Approximately 14.0% and 38.7% of participants could be identified by IRS_{111} as having a three- and two-folded increased breast cancer risk, respectively, compared to the average risk group (**eTable 6**).

medRxiv preprint doi: https://doi.org/10.1101/2021.09.23.21263888; this version posted September 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Absolute Risk of Developing Breast Cancer According to PRS/IRS Percentiles

Among the 10 207 Chinese women from the whole ABCC datasets, a dose-response association of breast cancer risk with percentiles of PRS_{111} or IRS_{111} was observed (**Figure 2A-B**). Compared to the average risk group, women in the top 5% of PRS_{111} and IRS_{111} were at a 3.39-folded (95% CI=2.80-4.10) and 5.22-folded (95% CI=4.37-6.25) increased risk of breast cancer, respectively; while those at the bottom 5% were at 0.30-folded (95% CI=0.23-0.39) and 0.27-folded (95% CI=0.21-0.35) decreased risk of breast cancer, respectively (**eTable 7**). The 10-year absolute risks were estimated by PRS_{111}/IRS_{111} percentiles and age groups. As shown in **Figure 2C-D**, in the same percentile group, risks estimated by IRS_{111} were higher than those by PRS_{111} across age groups. For women aged 60 years, the ranges of 10-year absolute risks estimated by PRS_{111} and IRS_{111} were 0.35%-7.68% and 0.38%-14.9%, respectively. Similar results were obtained from analyses using $PRS_{263\text{-}META}$ and $IRS_{263\text{-}META}$ (**eFigure 2**).

Discussion

In the present study, leveraging large GWAS datasets of Asian-ancestry women, we developed PRSs for breast cancer risk using three approaches and validated their prediction performance in an independent prospective test cohort. PRS_{111}, derived using the fine-mapping approach, was the best-performing PRS in this study (AUC=0.639). The prediction model incorporating PRS_{111} and six nongenetic risk factors achieved a further improved prediction accuracy (AUC=0.650).

A recent study compared the predictive ability of five Asian-specific PRSs with that of the 313-SNP European PRS in a retrospective dataset of Asian women.[23] The 313-SNP PRS was

medRxiv preprint doi: https://doi.org/10.1101/2021.09.23.21263888; this version posted September 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

significantly more predictive (AUC=0.617) than any of the five Asian-specific PRSs (best AUC=0.586).[23] However, because most of breast cancer risk variants were identified in GWAS conducted among European-ancestry women, the Asian-specific PRSs were derived using limited number of SNPs (n=5 to 51) in that study. In the present study, the most predictive PRSs based on these 313 SNP, $PRS_{263\text{-}META}$, under-performed the PRS_{111}, which was derived entirely using Asian data. The prediction ability of PRS_{111} in Asian women (AUC=0.639) is almost equivalent with that of the 313-SNP PRS in European-ancestry women (AUC=0.642).[14]

Most studies of prediction models incorporating PRS and nongenetic risk factors were carried out among women of European ancestry.[18,26,37-40] In general, including nongenetic risk factors could lead to improved prediction accuracy although the magnitude of improvement is relatively small. In a recent analysis using data from a prospective cohort of Dutch women, the 313-SNP European PRS was found to have an AUC of 0.636.[39] Incorporating this PRS with nine nongenetic risk factors improved the AUC to 0.653,[39] similar to the level achieved in our study. In 2010, we built an Asian-specific prediction model incorporating a 12-SNP PRS, and five nongenetic risk factors, which showed an AUC of 0.629 among Chinese women.[19] In the present study, IRS_{111}, the combination of PRS_{111} and the NgRS, the aggregation of six nongenetic risk factors, outperformed both PRS_{111} and the NgRS in predicting breast cancer risk.

The strengths of this study include the use of large GWAS datasets as the training set to improve the accuracy of estimating weights of breast cancer-associated SNPs for PRS construction. Instead of using only the lead SNPs identified in original GWAS of breast cancer, we performed fine-mapping analyses to identify additional breast cancer risk SNPs specifically for Asian

medRxiv preprint doi: https://doi.org/10.1101/2021.09.23.21263888; this version posted September 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

women. Because most of the breast-cancer associated SNPs were identified in European-ancestry populations and there are differences in genetic architectures between Asian and European-ancestry populations, we believe that this approach is necessary to conduct a PRS that is more appropriate for Asian women. In addition, state-of-the-art algorithms deriving PRSs using genome-wide SNPs were also employed in the present study and different combinations of parameters were tested for each algorithm. We demonstrated the ability of PRS-CSx in developing more predictive PRSs compared to other algorithms. Finally, the availability of both genetic and nongenetic risk factors data made it possible to establish and validate prediction models incorporating PRSs and nongenetic risk factors.

Limitations

This study also has several limitations. First, all PRSs had better prediction performance in our prospective test set than in our case-control validation set, which may be attributed to the design of our case-control validation set, in which case and control subjects were from two different studies, which could reduce the comparability between the case and control groups. Second, the sample size of our prospective test set is relatively small, which led to relatively wide ranges of 95% CIs for ORs and AUCs. Third, we included participants from both testing and training sets to increase the sample size in the analysis of 10-year absolute risks. Although the PRSs and IRSs used for relative risk estimation were externally validated, there might still be some potential for overfitting in risk estimation. Finally, our prospective test set only includes Chinese women, hence the prediction performance of PRSs in other Asian populations could not be investigated.

medRxiv preprint doi: https://doi.org/10.1101/2021.09.23.21263888; this version posted September 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

In summary, using data from the largest GWAS conducted in Asian women, we demonstrated that PRSs derived using breast cancer-associated risk SNPs show similar performance in predicting breast cancer risk in Asian and European descendants. Including known nongenetic risk factors in the models could further improve the accuracy of risk prediction. Our study provides strong supports for the utility of risk prediction models in developing personalized screening and prevention strategies.

Acknowledgements

The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agents. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. This research was primarily supported in part by the US National Institutes of Health grants R01CA235553, R01CA124558, R01CA158473, and R01CA148667. Sample preparation and genotyping assays at Vanderbilt were conducted at the Survey and Biospecimen Shared Resources and Vanderbilt Microarray Shared Resource, which are supported in part by the Vanderbilt-Ingram Cancer Center (P30CA068485). Data analyses were conducted using the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University. The SeBCS was supported by the BRL (Basic Research Laboratory) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (2011-0001564). KOHBRA/KOGES was supported by a grant from the National R&D Program for Cancer Control, Ministry for Health, Welfare and Family Affairs, Republic of Korea (#1020350). The KPOP was supported by Grant-in-Aid for Cancer Research and Control from the National Cancer Center of Korea (Grant Numbers 1410690 and 1710170). Studies conducted among

medRxiv preprint doi: https://doi.org/10.1101/2021.09.23.21263888; this version posted September 26, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Asian women include (Principal Investigator, grant support): the Shanghai Breast Cancer Study (W.Z. and X.-O.S., R01CA064277), the Shanghai Women's Health Study (W.Z., R37CA070867 and UM1CA182910), the Shanghai Breast Cancer Survival Study (X.-O. S., R01CA118229), the Shanghai Endometrial Cancer Study (X.-O.S., R01CA092585, controls only), the Seoul Breast Cancer Study [D.K., BRL (Basic Research Laboratory) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (2012-0000347)], the BioBank Japan Project (S.-K.L., the Ministry of Education, Culture, Sports, Sciences and Technology from the Japanese Government); the Hwasun Cancer Epidemiology Study-Breast (S.-S.K., the Biobank of Chonnam National University Hwasun Hospital, a member of the Korea Biobank Network, # 07SA2014020). The BCAC is funded by Cancer Research UK [C1287/A16563], the European Community's Seventh Framework Programme under grant agreement 223175 [HEALTH-F2-2009-223175] (COGS).

medRxiv preprint doi: https://doi.org/10.1101/2021.09.23.21263888; this version posted September 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

References

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Ca-Cancer J Clin*. May 2021;71(3):209-249. doi:10.3322/caac.21660
2. DeSantis CE, Ma J, Gaudet MM, et al. Breast cancer statistics, 2019. *CA: a cancer journal for clinicians*. 2019;69(6):438-451.
3. Sankaranarayanan R, Swaminathan R, Brenner H, et al. Cancer survival in Africa, Asia, and Central America: a population-based study. *The lancet oncology*. 2010;11(2):165-173.
4. Zheng W, Long J, Gao Y-T, et al. Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25. 1. *Nature genetics*. 2009;41(3):324-328.
5. Long J, Cai Q, Shu X-O, et al. Identification of a functional genetic variant at 16q12. 1 for breast cancer risk: results from the Asia Breast Cancer Consortium. *PLoS Genet*. 2010;6(6):e1001002.
6. Cai Q, Long J, Lu W, et al. Genome-wide association study identifies breast cancer risk variant at 10q21. 2: results from the Asia Breast Cancer Consortium. *Human molecular genetics*. 2011;20(24):4991-4999.
7. Long J, Cai Q, Sung H, et al. Genome-wide association study in east Asians identifies novel susceptibility loci for breast cancer. *PLoS Genet*. 2012;8(2):e1002532.
8. Cai Q, Zhang B, Sung H, et al. Genome-wide association analysis in East Asians identifies breast cancer susceptibility loci at 1q32. 1, 5q14. 3 and 15q26. 1. *Nature genetics*. 2014;46(8):886-890.
9. Long J, Delahanty RJ, Li G, et al. A common deletion in the APOBEC3 genes and breast cancer risk. *J Natl Cancer I*. 2013;105(8):573-579.
10. Han M-R, Long J, Choi J-Y, et al. Genome-wide association study in East Asians identifies two novel breast cancer susceptibility loci. *Human molecular genetics*. 2016;25(15):3361-3371.
11. Shu X, Long J, Cai Q, et al. Identification of novel breast cancer susceptibility loci in meta-analyses conducted among Asian and European descendants. *Nature Communications*. 2020;11(1):1-9.
12. Jia G, Shu X, Yang Y, et al. Meta-analyses of genome-wide association study data from Asian and European-ancestry women identify 17 novel breast cancer susceptibility loci. *Nature communications*. (Submitted)
13. Zhang H, Ahearn TU, Lecarpentier J, et al. Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nature genetics*. 2020:1-10.
14. Mavaddat N, Michailidou K, Dennis J, et al. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *The American Journal of Human Genetics*. 2019;104(1):21-34.
15. Kramer I, Hooning MJ, Mavaddat N, et al. Breast cancer polygenic risk score and contralateral breast cancer risk. *The American Journal of Human Genetics*. 2020;107(5):837-848.
16. Jia G, Lu Y, Wen W, et al. Evaluating the utility of polygenic risk scores in identifying high-risk individuals for eight common cancers. *JNCI cancer spectrum*. 2020;4(3):pkaa021.

medRxiv preprint doi: https://doi.org/10.1101/2021.09.23.21263888; this version posted September 26, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

17. Hughes E, Tshiaba P, Gallagher S, et al. Development and validation of a clinical polygenic risk score to predict breast cancer risk. *JCO precision oncology*. 2020;4:585-592.
18. Maas P, Barrdahl M, Joshi AD, et al. Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the United States. *JAMA oncology*. 2016;2(10):1295-1302.
19. Zheng W, Wen W, Gao Y-T, et al. Genetic and clinical predictors for breast cancer risk assessment and stratification among Chinese women. *J Natl Cancer I*. 2010;102(13):972-981.
20. Wen W, Shu X-o, Guo X, et al. Prediction of breast cancer risk based on common genetic variants in women of East Asian ancestry. *Breast Cancer Res*. 2016;18(1):1-8.
21. Lee CPL, Irwanto A, Salim A, et al. Breast cancer risk assessment using genetic variants and risk factors in a Singapore Chinese population. *Breast Cancer Res*. 2014;16(3):1-13.
22. Hsieh Y-C, Tu S-H, Su C-T, et al. A polygenic risk score for breast cancer risk in a Taiwanese population. *Breast cancer research and treatment*. 2017;163(1):131-138.
23. Ho W-K, Tan M-M, Mavaddat N, et al. European polygenic risk score for prediction of breast cancer shows similar performance in Asian women. *Nature communications*. 2020;11(1):1-11.
24. Lee A, Mavaddat N, Wilcox AN, et al. BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genetics in Medicine*. 2019;21(8):1708-1718.
25. Rudolph A, Song M, Brook MN, et al. Joint associations of a polygenic risk score and environmental risk factors for breast cancer in the Breast Cancer Association Consortium. *International journal of epidemiology*. 2018;47(2):526-536.
26. Vachon CM, Scott CG, Tamimi RM, et al. Joint association of mammographic density adjusted for age and body mass index and polygenic risk score with breast cancer risk. *Breast Cancer Res*. 2019;21(1):1-10.
27. Kapoor PM, Mavaddat N, Choudhury PP, et al. Combined associations of a polygenic risk score and classical risk factors with breast cancer risk. *JNCI: Journal of the National Cancer Institute*. 2020;
28. Zheng W, Chow W-H, Yang G, et al. The Shanghai Women's Health Study: rationale, study design, and baseline characteristics. *Am J Epidemiol*. 2005;162(11):1123-1131.
29. Michailidou K, Lindström S, Dennis J, et al. Association analysis identifies 65 new breast cancer risk loci. *Nature*. 2017;551(7678):92-94.
30. Zhang H, Ahearn TU, Lecarpentier J, et al. Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nature genetics*. 2020;52(6):572-581.
31. Yang J, Ferreira T, Morris AP, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature genetics*. 2012;44(4):369-375.
32. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*. 2011;88(1):76-82.
33. Vilhjálmsson BJ, Yang J, Finucane HK, et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The american journal of human genetics*. 2015;97(4):576-592.

medRxiv preprint doi: https://doi.org/10.1101/2021.09.23.21263888; this version posted September 26, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

34. Privé F, Arbel J, Vilhjálmsson BJ. LDpred2: better, faster, stronger. *Bioinformatics*. 2020;36(22-23):5424-5431.
35. Ge T, Chen C-Y, Ni Y, Feng Y-CA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature communications*. 2019;10(1):1-10.
36. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *Bmc Bioinformatics*. 2011;12(1):1-8.
37. van Veen EM, Brentnall AR, Byers H, et al. Use of single-nucleotide polymorphisms and mammographic density plus classic risk factors for breast cancer risk prediction. *JAMA oncology*. 2018;4(4):476-482.
38. Shieh Y, Hu D, Ma L, et al. Breast cancer risk prediction using a clinical risk model and polygenic risk score. *Breast cancer research and treatment*. 2016;159(3):513-525.
39. Lakeman IM, Rodríguez-Girondo M, Lee A, et al. Validation of the BOADICEA model and a 313-variant polygenic risk score for breast cancer risk prediction in a Dutch prospective cohort. *Genetics in Medicine*. 2020;22(11):1803-1811.
40. Zhang X, Rice M, Tworoger SS, et al. Addition of a polygenic risk score, mammographic density, and endogenous hormones to existing breast cancer risk prediction models: a nested case–control study. *PLoS medicine*. 2018;15(9):e1002644.

medRxiv preprint doi: https://doi.org/10.1101/2021.09.23.21263888; this version posted September 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Figure Legends

Figure 1. Distributions of PRS_{111} and $PRS_{263\text{-}META}$ between breast cancer cases and controls in our prospective test set. Density plot between cases were controls for PRS_{111} (A) and $PRS_{263\text{-}META}$ (C). Violin and boxplots between cases and controls for PRS_{111} (B) and $PRS_{263\text{-}META}$ (D).

Figure 2. Ten-year absolute risk of developing breast cancer estimated using data from 10 207 Chinese women. ORs of breast cancer for percentiles of PRS_{111} (A) and IRS_{111} (B) compared to the average risk group (40%-60% percentiles). Ten-year absolute risk of breast cancer by percentiles of PRS_{111} (C) and IRS_{111} (D) for women in different age categories.

medRxiv preprint doi: https://doi.org/10.1101/2021.09.23.21263888; this version posted September 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Tables

Table 1. Summary of participating studies included in the current project

Study	No. of cases	No. of controls	Age at enrollment [a]	
			Cases	Controls
PRS training and testing				
Training set				
SBCGS	5384	6347	52.8 ± 9.3	52.1 ± 9.2
HCES-Br	274	273	49.1 ± 10.8	54.0 ± 7.4
KPOP	963	921	-	-
BBJ2	5552	89 731	-	-
SeBCS	2246	2052	-	-
BCAC-Asians	4231	5067	54.4 ± 10.4	53.8 ± 10.0
Validation set				
SBCGS	1426	1323	50.1 ± 11.3	50.6 ± 9.5
Sub-Total	20 076	105 714		
Prospective study				
SWHS	368	736	52.1 ± 8.7	51.6 ± 9.5

Abbreviations: SBCGS, Shanghai Breast Cancer Genetic Study; HCES-Br, Hwasun Cancer Epidemiology Study-Breast; KPOP, Korea Precision Oncology Program; BBJ2, The Biobank Japan Project 2; SeBCS, Seoul Breast Cancer Study; BCAC, Breast Cancer Association Consortium; SWHS, Shanghai Women's Health Study.

[a] Mean ± standard deviation (SD) is presented. Individual level data was not available for KPOP, BBJ2 and SeBCS.

medRxiv preprint doi: https://doi.org/10.1101/2021.09.23.21263888; this version posted September 26, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

ons of PRSs with breast cancer risk in the validation set and prospective test set, the Asia Breast Cancer Consortiu

methods	**Validation set (1426 cases vs. 1323 controls)**			**Prospective test set (368 cases**	
	OR (95% CI) [a]	**AUC (95% CI)**	***P* [a]**	**OR (95% CI) [a]**	**AUC (95%**
an PRS [b]					
	1.42 (1.31-1.53)	0.597 (0.575-0.618)	2.47×10^{-18}	1.62 (1.42-1.85)	0.625 (0.590
	1.44 (1.33-1.56)	0.601 (0.580-0.622)	5.47×10^{-20}	1.58 (1.38-1.80)	0.616 (0.582
	1.44 (1.33-1.55)	0.600 (0.579-0.621)	1.54×10^{-19}	1.63 (1.43-1.87)	0.626 (0.592
$<10^{-5}$)	1.45 (1.34-1.57)	0.603 (0.582-0.624)	2.72×10^{-20}	1.67 (1.46-1.92)	0.639 (0.604
$<10^{-4}$)	1.42 (1.31-1.53)	0.597 (0.575-0.618)	1.38×10^{-18}	1.63 (1.42-1.87)	0.632 (0.597
$<10^{-3}$)	1.38 (1.28-1.49)	0.592 (0.571-0.613)	3.30×10^{-16}	1.54 (1.35-1.76)	0.619 (0.584
prediction algorithms [d]					
284 SNPs)	1.44 (1.34-1.56)	0.600 (0.579-0.621)	4.96×10^{-20}	1.52 (1.34-1.74)	0.616 (0.581
80 SNPs)	1.40 (1.29-1.51)	0.591 (0.570-0.612)	4.77×10^{-17}	1.51 (1.33-1.72)	0.612 (0.577
80 SNPs)	1.51 (1.39-1.63)	0.613 (0.592-0.634)	3.03×10^{-24}	1.70 (1.49-1.95)	0.642 (0.608

k score; OR, odds ratio; CI, confidence interval; AUC, aera under the receiver operating characteristic curve.

er SD increase in PRS and *P* values were estimated using logistic regression.

included in the European-ancestry PRS reported by Zhang et al. *Nat Genet*. 2020, data on 263 SNPs were available in ou s and thus included in this analysis. These PRSs were developed using weights from BCAC-European data only, Asian d vo datasets, respectively.

SNPs selected from fine-mapping of Asian data and showing consistent association directions in BCAC-European data w ed using Asian data.

n, only the most predictive PRS in the validation set is presented. Weights of SNPs from our training set were estimated

medRxiv preprint doi: https://doi.org/10.1101/2021.09.23.21263888; this version posted September 26, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

medRxiv preprint doi: https://doi.org/10.1101/2021.09.23.21263888; this version posted September 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Table 3. Performance of risk scores in the prospective test set

Risk score	OR (95% CI) [a]	AUC (95% CI)	*P* [a]
NgRS [b]	1.29 (1.14-1.46)	0.565 (0.529-0.601)	6.36×10^{-5}
PRS_{111} [c]	1.67 (1.46-1.92)	0.639 (0.604-0.674)	1.28×10^{-13}
IRS_{111} [c]	1.77 (1.55-2.04)	0.650 (0.616-0.685)	4.24×10^{-16}
$PRS_{263\text{-}META}$ [d]	1.63 (1.43-1.87)	0.626 (0.592-0.661)	1.25×10^{-12}
$IRS_{263\text{-}META}$ [d]	1.72 (1.51-1.98)	0.639 (0.605-0.673)	6.76×10^{-15}

PRS, polygenic risk score; NgRS, nongenetic risk score; IRS, integrated risk score; OR, odds ratio; CI, confidence interval; AUC, aera under the receiver operating characteristic curve.

[a] OR and 95% CI per SD increase and *P* values was estimated using logistic regression.

[b] Based on weighted six nongenetic risk factors and an interaction item.

[c] PRS_{111}: the best PRS derived in the present study. IRS_{111}: the combination of PRS_{111} and the NgRS.

[d] $PRS_{263\text{-}META}$ was derived based on meta-analysis results of Asian and BCAC-European data for 330 SNPs initially reported in European-ancestry populations (Zhang et al. *Nat Genet. 2020)*. $IRS_{263\text{-}META}$ was the combination of $PRS_{263\text{-}META}$ and the NgRS.