Polygenic Risk Scores for Prediction of Breast Cancer in Korean women

Authors:

Yon Ho Jee, MS[1], Weang-Kee Ho, PhD[2,3], Sohee Park, PhD [4], Douglas F. Easton, PhD [5,6], Soo-Hwang Teo, PhD [3,7], Keum Ji Jung, PhD [8,9], Peter Kraft, PhD [1,10*]

Author affiliations:

[1]*Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America.*

[2]*School of Mathematical Sciences, Faculty of Science and Engineering, University of Nottingham Malaysia, Semenyih, Selangor, Malaysia.*

[3]*Cancer Research Malaysia, Subang Jaya, Selangor, Malaysia.*

[4]*Department of Biostatistics, Yonsei University Graduate School of Public Health, Seoul, Republic of Korea.*

[5]*Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom.*

[6]*Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, United Kingdom.*

[7]*Sime Darby Medical Centre, Subang Jaya, Selangor, Malaysia.*

[8]*Institute for Health Promotion, Graduate School of Public Health, Yonsei University, Seoul, Republic of Korea.*

[9]*Nuffield Department Population Health, University of Oxford, Oxford, United Kingdom.*

[10]*Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America.*

[*]Correspondence authors:

Peter Kraft, PhD

Department of Epidemiology, Department of Biostatistics,

Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America.

Email: pkraft@hsph.harvard.edu

ABSTRACT

Background: Polygenic risk scores (PRS) developed in large European GWAS have been shown to predict breast cancer risk in European-ancestry women and perform well in Asian women relative to PRS developed in smaller Asian studies. However, prospective validation of absolute risk predictions from models combining lifestyle and PRS in Asian women is limited.

Objectives: Here we evaluate the calibration of four PRS developed using Asian or European training samples; two PRS were restricted to genome-wide significant SNPs (GRS-11_{ASN} and GRS-136_{EUR}) and two included sub-genome-wide significant SNPs (PRS-42_{ASN} and PRS-209_{EUR}). We also assessed the improvement in risk prediction and risk stratification by incorporating the PRS into classical risk factor model.

Methods: For each PRS, we compared discrimination (area under the curve [AUC]) and calibration (expected-to-observed ratio [E/O]) of three absolute risk models in a cohort of 41,031 women from the Korean Cancer Prevention Study-II Biobank: (i) a model using incidence, mortality, and risk factor distributions among U.S. non-Hispanic white women and European-ancestry relative risks; (ii) a recalibrated model, using Korean incidence mortality and risk factor distributions but European-ancestry relative risks; and (iii) a fully Korean-based model using Korean incidence mortality and risk factor distributions and relative risk estimates from the KCPS.

Results: All Asian and European PRS improved risk prediction for breast cancer in Korean women under 50 (Qx: AUC=0.65, Qx+PRS-42_{ASN}: AUC=0.68, Qx+PRS-209_{EUR}: AUC=0.69

in Korean-based model for age<50). European PRS had larger effect sizes and greater discrimination than Asian PRS in a Korean population (PRS-42_{ASN}: HR per SD = 1.40, AUC = 0.60 vs. PRS-209_{EUR}: HR per SD = 1.54, AUC = 0.62). We found that the U.S.-based absolute risk models overestimated the risks for women age ≥50 years (PRS-42_{ASN}: E/O=1.93, 95% CI=1.69, 2.19, PRS-209_{EUR}: E/O = 1.92, 95% CI=1.69, 2.19). Our absolute risk projections suggest that risk-reducing lifestyle changes would lead to larger absolute risk reductions among women at higher PRS.

Conclusions: Incorporation of PRS previously developed in Asian and European-ancestry populations can improve discrimination in Korean women. Our finding suggests that PRS may be useful for prioritizing individuals for targeted intervention on their lifestyle such as alcohol intake and obesity. Larger Asian training samples should improve PRS discrimination among Korean women. Further studies are needed to evaluate the value of incorporating PRS into risk models in ancestrally diverse populations.

INTRODUCTION

Breast cancer is the leading cancer diagnosed among women in most countries in the world. While the incidence of breast cancer in Asian women is currently lower than that in Western countries, the age distribution of breast cancer incidence in Asian women is markedly different from that in the Western countries, with a peak at 45–49 years in the Asian countries vs. 60–70 years in the Western countries [1, 2]. We previously found this age difference led to overestimation of risk in Korean women when conventional breast cancer risk models developed in European-ancestry populations were used [3]. This underscores the need to validate Western-derived risk prediction models in Asian women and adapt them to improve their predictive ability.

In addition to lifestyle, clinical and environmental breast cancer risk factors, genetic susceptibility can play an important role in the development of breast carcinogenesis [4]. A large proportion of genetic variation in risk for breast cancer is polygenic due to multiple common single nucleotide polymorphisms (SNPs) with a small risk individually. These common breast cancer susceptibility SNPs have been discovered by genome-wide association studies (GWAS) [5-12], and their combined effects can be summarized as polygenic risk scores (PRS) [13-15]. The incorporation of 313-SNP PRS developed for European ancestry women (69,732 controls and 88,916 cases) into classical risk prediction model further improved discrimination and risk stratification in women of European descent [15]. A 46-SNP PRS was developed in Asian women (22,113 controls and 22,013 cases), which was shown to be less predictive than the European-specific PRS in Asian women [16]. This result held in Korean women (hazard ratio [HR] per unit standard deviation [SD]=1.57 for European PRS vs. HR per SD=1.40 for Asian PRS). However, prospective validation of

absolute risk predictions from models incorporating both lifestyle and PRS in Asian women is limited.

We previously used Individualized Coherent Absolute Risk Estimation (iCARE) [17] to validate three risk prediction models (the U.S.-based European-ancestry model, a recalibrated model, and a fully Korean-based model) based on classical breast cancer risk factors in a Korean population [3]. Here we evaluate the predictive capacity of four PRS developed using Asian or European training samples; two PRS were restricted to genome-wide significant SNPs (GRS-11_{ASN} and GRS-136_{EUR}) and two included sub-genome-wide significant SNPs (PRS-42_{ASN} PRS-209_{EUR}). We also assessed the improvement in risk prediction and risk stratification by incorporating the PRS into classical risk factor model.

METHODS

Study population for discrimination and calibration analyses.

The validation cohort includes 41,031 women age 20-80 years at enrollment from the Korean Cancer Prevention Study-II (KCPS-II) Biobank. The study participants undertook routine health assessments at nationwide health promotion centers between 2004 and 2013. The study design and recruitment have been described in detail previously [18]. Exclusion criteria included history of any cancer, no genetic consent or DNA source, and age at entry below 20 or above 80 years. Of 41,031 women, 705 breast cancer cases occurred over 10 years of follow-up (Supplementary Table 1). These 41,031 women were not a representative sample of the women with available DNA in the KCPS-II since cancer cases were oversampled for genotyping (Supplementary Figure 1). We account for this oversampling in calibration analyses using inverse-probability-of-sampling-weights as implemented in the iCARE

software. All participants gave written informed consent before participation. The Institutional Review Board of Yonsei University approved this study protocol (IRB approval number 4-2011-0277).

Data collection

Data on reproductive and lifestyle factors were collected at enrollment: age at menarche, age at menopause, parity, age at first birth, oral contraceptive (OC) use (never, ever), hormone replacement therapy (HRT) use, alcohol intake, history of benign breast disease (BBD), and family history of breast cancer. Height and weight were measured while participants wore light clothing. Body mass index (BMI) was calculated as the weight (kg) divided by the height squared (m^2).

Genotyping methods.

All samples in the validation cohort were genotyped using Illumina Global Screening Array v2.0. Based on a subset of autosomal variants with minor allele frequency (MAF) ≥ 0.01 and genotype call rate ≥ 98%, samples with call rate < 98% and extremes in heterozygosity were excluded. Data were imputed using SHAPEIT [19] for phasing and MINIMAC3 [20] for imputation, with 1000 Genomes Project (Phase 3) as the reference panel [21]. Post-imputation quality was evaluated using the R^2 as provided by MIMINAC3 [20]. We excluded SNPs with R^2 below 0.80 which were taken to be indicative of poor imputation quality and were not analyzed. Principal components analyses were used to identify genetic outliers and define ancestry-informative covariates.

Follow-up for breast cancer.

The outcome was incidence of breast cancer (ICD-10 codes C50). The follow-up was nearly 100% complete because all participants have a unique identification number assigned at birth, allowing linkage with the national cancer registry and hospital admission records. The data on cancer diagnoses were obtained based on histologic type, resulting in high accuracy.

Study population for relative risk estimation

The Korean Cancer Prevention Study (KCPS) was used to independently estimate the relative risks (RRs) for breast cancer risk factors (Supplementary Table 2). The KCPS is a 1.3-million-member prospective cohort study, designed to assess risk factors for mortality, incidence, and hospital admission from cancer, with a follow-up of 25 years [22]. The KCPS cohort includes the 443,627 women ages 20–80 years who received health insurance from the Korean Medical Insurance Corporation and who had biennial medical evaluations between 1992 and 1995. The collection of risk factors was similarly done to the KCPS-II. Because history of BBD was not asked for women in the KCPS, we defined history of BBD based on ICD-10 code D24. In the KCPS cohort, an incident breast cancer was coded based on a hospital admission for a cancer diagnosis.

Statistical analysis.

Polygenic risk scores. We compared the performances of four PRS using genome-wide significant SNPs and polygenic SNPs previously reported in Asian and European-ancestry studies: (i) Asian genome-wide significant SNPs found in the Biobank Japan [23], (ii) Asian sub-genome-wide significant SNPs included in Asian-based PRS [16], (iii) European genome-wide significant SNPs reported in the Breast Cancer Association Consortium (BCAC) [5], and (iv) European sub-genome-wide significant SNPs included in European-based PRS

[15]. We removed SNPs with low imputation quality ($R^2 < 0.8$) in the KCPS-II. There was one SNP overlapping between the Asian genome-wide significant SNPs and the SNPs in the Asian-based PRS and 40 SNPs overlapping between European genome-wide significant SNPs and the SNPs in the European-based PRS. For the PRS based on genome-wide significant SNPs, the genome-wide significance threshold was set at $p < 5 \times 10^{-8}$, and the sub-genome-wide-significant PRS were developed by clumping and thresholding (C+T) methods [15, 16]. The Asian genome-wide significant SNPs included significant SNPs at $p < 5 \times 10^{-8}$ after clumping (R^2=0.1) using the East Asian 1000 Genomes reference panel. The European genome-wide significant SNPs were drawn from collaborations of multiple studies including the Breast Cancer Association Consortium (BCAC).

We derived PRS for breast cancer using the formula PRS $\sum_{i=1}^{k} \beta_i x_i$ where x_i is the number of risk alleles (0,1,2) for SNP i and β_i is the corresponding weight. For the PRS based on genome-wide significant SNPs, we used publicly available weights from a GWAS of Japanese or European ancestry women [5, 23]. For the Asian and European PRS, we used the corresponding published weights [15, 16]. All PRS were standardized to have mean zero and standard deviation (SD) one in the total sample.

The questionnaire-based models included baseline data on reproductive, anthropometric, behavioral, and clinical risk factors: age at menarche, age at menopause, parity, age at first birth, oral contraceptive use, hormone replacement therapy use, BMI, height, alcohol intake, history of benign breast disease, and family history of breast cancer.

As a sensitivity analysis, we calculated a PRS using the SNPs from the European ancestry scores but taking the Asian per-allele odds ratios (ORs) [23] as weights. We compared the

performances of these PRS with European PRS weighted by European per-allele ORs. In addition, we performed a sensitivity analysis of including all SNPs without imputation R^2 threshold.

Association analyses. A Cox proportional hazards regression was used to estimate HRs per SD of each PRS (time scale = number of years since baseline, adjusted for the first 2 principal components). We calculated area under the receiver operating characteristic curve (AUC) using logistic regression.

Discrimination and calibration. To evaluate the performance of five-year absolute risk models based on questionnaire data only (Qx), PRS only, and both questionnaire and PRS data (Qx+PRS), we used the iCARE software to estimate discrimination, measured by Area Under Operating Curve (AUC), and calibration (overall expected-to-observed ratio [E/O] and expected versus observed incidence by expected risk deciles). iCARE calculates five-year absolute risks based on several external inputs: RR estimates for included risk factors; average age-specific absolute risk rates; the distribution of risk factors in the target population (estimated using a reference sample); and the age-specific competing mortality rates. iCARE then compares the predicted absolute risks to the observed incidence. We performed these discrimination and calibration analyses separately for the first and second 5-years of follow-up, excluding women diagnosed with breast cancer or lost to follow-up in the first 5 years from the second 5 years of follow up. Summary AUC and E/O estimates across these two follow-up periods were calculated using the fixed effects inverse variance weighting method.

For each RR model incorporating questionnaire data, we evaluated three absolute risk models [14] in the KCPS-II Biobank: (i) a U.S.-based European-ancestry model, using incidence, mortality, and risk factor distributions among U.S. non-Hispanic white women and European-ancestry RRs; (ii) a recalibrated model, using Korean incidence/ mortality and risk factor distributions but European-ancestry RRs; and (iii) a fully Korean-based model using Korean incidence/ mortality and risk factor distributions and RR estimates from the KCPS. For the RR associated with a 1-SD difference in PRS, we used external estimates in models (i) and (ii) (from European-ancestry studies for European-ancestry PRS and from Asian-ancestry studies for Asian-ancestry PRS) and estimates from the KCPS-II for model (iii). Supplementary Table 2 provides detailed descriptions of RR estimates included in the models and population distribution.

To obtain information on risk factor distributions, iCARE uses an additional individual-level reference dataset of risk factors representing each population. The reference datasets were 2010 National Health and Nutrition Examination Survey (NHANES) for the U.S.-based model and 2010-2012 Korean NHANES (KNHANES) for the recalibrated model and the Korean-based model. Details on classical risk factor distributions are in Supplementary Table 2. Since neither reference dataset has genetic information, we randomly assigned each individual in the reference data sets a normally distributed log PRS value with variance 1, mean 0 for individuals without a first-degree family history, and mean $\beta/2$ for individuals with a first-degree family history, where β is the log RR per-SD difference in PRS.

Absolute risk of breast cancer by PRS and modifiable risk factors.

We estimated cumulative and 10-year absolute risk trajectories across strata defined by genetic and modifiable risk profiles using the probability method in the Korean-based Qx + Asian-based PRS model. The absolute risk of developing breast cancer for a woman of age *a* over the time interval $a + s$ can be calculated as

$$R_{a,a+s} = \int_a^{a+s} \lambda_0(t) \exp(\mathbf{Z}'\boldsymbol{\beta}) \exp\left(-\int_a^t [\lambda_0(u) exp(\mathbf{Z}'\boldsymbol{\beta}) + m(u)]\, du\right) dt$$

(a)

Formula (a) holds under the assumptions that the risk factors **Z** act in a multiplicative fashion on the baseline hazard function $\lambda_0(t)$. Here, $\boldsymbol{\beta}$ represents a vector of log RRs associated with each of the risk factors in **Z**. The baseline hazard is calculated by dividing the population average age-specific incidence rates by the average RR, where the average is taken over the risk factor distribution; risk factor distributions were calculated separately for women younger and older than fifty using the KNHANES data. Formula (a) accounts for competing risks due to mortality from other causes through the age-specific mortality rate function *m(t)*.

Cumulative risk is evaluated as absolute risk between age 20 years and a specific age. The 10-year risk is evaluated as absolute risk over the next 10 years for a woman who has attained a specific age without developing breast cancer. To use this method, we estimated the multivariable RRs of each woman in the KCPS-II based on their risk factors *Z*, the log RRs ($\boldsymbol{\beta}$) estimated in the KCPS, the age-specific mortality rates of breast cancer in Korea, and the risk factor distribution in KNHANES.

We classified individuals in the top 20% of the PRS distribution as high PRS (corresponding RR≥1.33), those in the bottom 20% as low PRS (corresponding RR≤0.75), and those in the

middle category (>20th to <80th percentile) as middle PRS. Individuals above the median of modifiable risk score distribution were classified as an elevated modifiable risk and those below the median were considered as reduced modifiable risk group. Modifiable risk scores were generated using predicted log RRs based on BMI, oral contraceptive use, alcohol intake, and hormonal replacement therapy use. For each of these six PRSxmodifiable risk groups, we calculated absolute risks for hypothetical women who have RR equal to the average RR in that group.

We performed bootstrapping to generate distributions of estimated lifetime (by age 80 years) absolute risk accounting for sampling uncertainty in risk factor RRs. We created 1,000 random samples of log RR parameters for *k* variables ($\hat{\beta}_{1i}, \hat{\beta}_{2i}, \ldots, \hat{\beta}_{ki}$, i=1,…,1000) from multivariate normal distribution with mean equal to the empirical log RRs ($\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$) and variance equal to the squared standard errors of the estimates. For each of the 1,000 replicates, we calculated absolute risks for each of the six strata as above. The quantiles of this simulated distribution were used to estimate 95% confidence intervals.

Furthermore, we performed 5-year absolute risk projections by three birth cohorts (1920-1949, 1950-1969, 1970-1995 years) to examine birth cohort effects. For birth-cohort specific rates, age-specific rates were estimated using data on breast cancer mortality and incidence in Korea from 2000 to 2015. Since the age-specific data were partially available in some birth cohorts, the absolute risk projections were estimated using different age ranges across the birth cohorts (20-40, 35-60, and 55-75 years old).

All statistical tests were two-sided at a significance level of 0.05 and calculated using SAS version 9.4 software (SAS Institute, Cary, NC) for descriptive statistics and RRs. Absolute risks were evaluated with R version 4.0.3 software using the iCARE package 1.0.0.

RESULTS

Study sample.

Data on 705 breast cancer cases and 40,326 control women from the KCPS-II Biobank were included. The population distributions of breast cancer risk factors differ between the U.S. non-Hispanic white and Korean population (Supplementary Table 2). Korean women tend to have later age at menarche, earlier age at menopause, and later age at first birth than the U.S. non-Hispanic white women. The proportions of women who use OC or HRT were markedly lower among Korea women than U.S. non-Hispanic white women. Korean women, on average, had a lower BMI than U.S. non-Hispanic white women. The RRs between European-ancestry and Korean women for most risk factors did not differ appreciably, except for BBD, which had a larger effect on breast cancer among Korean women than European-ancestry women (RR=5.05 vs. 1.68). This may reflect the narrower definition of BBD in the KPCS-II, focusing on women with a history of benign neoplasm of unspecified breast. In addition, the inverse association between BMI and breast cancer risk seen among premenopausal European-ancestry women was not seen among Korean women (Supplementary Table 2).

All PRS had higher mean in cases than controls (Table 1). Among breast cancer cases, the mean PRS was higher for the European-derived PRS than for the Asian-derived PRS (GRS-11_{ASN}: 0.29 vs. GRS-136_{EUR}: 0.43, PRS-42_{ASN}: 0.33 vs. PRS-209_{EUR}: 0.43). Of published SNPs for each PRS, we included SNPs that passed imputation $R^2 > 0.8$ in the KCPS-II. Four PRS were constructed using previously reported SNPs: i) Asian genome-wide significant SNPs (GRS-11_{ASN}); ii) Asian polygenic SNPs (PRS-42_{ASN}); iii) European genome-wide significant SNPs (GRS-136_{EUR}); and iv) European polygenic SNPs (PRS-209_{EUR}) (Table 2).

Evaluation of Asian and European PRS in KPCS-II Biobank.

Table 2 shows the estimated HR per unit increase of PRS and AUC for breast cancer. Compared with the Asian PRS, the European PRS had larger effect sizes (PRS-209_{EUR}: HR per SD = 1.54 vs. PRS-42_{ASN}: HR per SD = 1.40) and a greater discrimination (PRS-209_{EUR}: AUC = 0.62 vs. PRS-42_{ASN}: AUC = 0.60) in the KCPS-II. For Asian PRS, PRS-42_{ASN} had a larger estimated HR and AUC than GRS-11_{ASN} (GRS-11_{ASN}: HR per SD = 1.35, 95% confidence interval (CI) = 1.25, 1.45, AUC = 0.58). On the other hand, there was little difference in HR and AUC between European genome-wide SNPs PRS and European polygenic SNPs PRS (GRS-136_{EUR}: HR per SD = 1.55, 95% CI = 1.44, 1.66, AUC = 0.62; PRS-209_{EUR}: HR per SD = 1.54, 95% CI = 1.43, 1.66, AUC = 0.62).

We then evaluated the predictive performance of models incorporating the PRS into absolute risk models along with conventional questionnaire-based risk factors. The incorporation of PRS improved discrimination overall (Figure 1, Supplementary Table 3). The improvement was slightly greater for the incorporation of European PRS compared to the incorporation of Asian PRS, especially among women of age 50+ (Qx: AUC=0.54, Qx + PRS-42_{ASN}: AUC=0.59, Qx + PRS-209_{EUR}: AUC=0.60 in Korean-based model). For Asian PRS, the combined model with PRS-42_{ASN} had a slightly larger AUC than GRS-11_{ASN} (Korean-based model: Qx + GRS-11_{ASN} =0.66 vs. Qx + PRS-42_{ASN} =0.68 in age<50; Qx + GRS-11_{ASN} =0.58 vs. Qx + PRS-42_{ASN} =0.59 in age 50+). Such difference was minimal between combined model with GRS-136_{EUR} and PRS-209_{EUR}.

Model calibration differed depending on the population incidence rates and reference population used to calculate absolute risks and whether Qx risk factor data were included (Figure 2, Supplementary Table 4). We found that the U.S.-based absolute risk models

overestimated the risks for women age ≥50 years, even after the incorporation of PRS (PRS-42_{ASN}: E/O=1.94, 95% CI=1.70, 2.21, PRS-209_{EUR}: E/O = 1.92, 95% CI=1.68, 2.18). The overestimation was notably large in all U.S.-based PRS-only models (up to E/O=2.70). Further recalibrations of the models using the Korean incidence and mortality rates, risk factor distributions from Korean population, and RR estimates from Korean population showed improved calibration, especially among older women (Figure 3).

In our sensitivity analysis, we found the PRS consisting of all SNPs without R^2 threshold did not improve the AUCs (Supplementary Table 5). Although all variants with $R^2 > 0$ were considered, still fewer number SNPs were analyzed than the published SNPs (168 for 172-SNP PRS, 304 for 313-SNP PRS) because variants with low MAF (<0.01) were excluded. Also, the performance of European PRS using the Asian per-allele ORs as weights did not change as compared to the European PRS weighted by the European per-allele ORs (Supplementary Table 6).

Absolute breast cancer risk predictions.

We explored cumulative and 10-year absolute risk trajectories across strata defined by PRS-42_{ASN} and modifiable risk profiles in Korean-based model (Figure 4). The risk of developing breast cancer by age 80 was ~2% for women in the low PRS group with reduced modifiable risk and 7% for those in the high PRS group with elevated modifiable risk. This model suggests that interventions on modifiable risk factors have the potential to reduce breast cancer risk, even among women at high risk due to their inherited genetics. For example, women in the high PRS group with reduced modifiable risk (dotted red line) can have a lower risk than those with middle PRS but elevated modifiable risk (dashed pink line). Moreover, the differences in absolute risk between women with elevated and reduced modifiable risks

were larger for women with higher PRS. The amount of risk difference between elevated and reduced modifiable risk was higher in higher PRS group (~3% in the high PRS, ~2% in the middle PRS, and ~1.5% in the low PRS group). Although we acknowledge the 95% CIs of these strata would mostly overlap due to low precision in KCPS, the bootstrapping distribution of lifetime (by age 80 years) absolute risk showed that only small proportion would overlap, especially between the lowest and the highest risk stratum (Supplementary Table 7).

We investigated absolute risk projections by birth cohort (Supplementary Figure 2). The most recent birth cohort (1970-1995) showed an increasing trend in 5-year absolute breast cancer risk by age (up to age 40, the oldest available age for this cohort). Incidence increased up to age 45 for the 1950-1969 birth cohort, then plateaued. Incidence was low and decreased slightly with age in the available range (55-75) for the oldest birth cohort (1920-1949). The 5-year absolute breast cancer risk at age 40 increased over time: it was 5.7% for birth cohort 1970-1995 and 3.7% for birth cohort 1950-1969.

DISCUSSION

In this study evaluating four PRS for breast cancer risk prediction in Korean women, we found that European-based PRS performed better than Asian-based PRS. We also observed that inclusion of polygenic variants in additional to classical risk factors improved the prediction in Korean women. Our absolute risk projections suggest that a larger absolute risk reduction would occur among women at higher PRS by shifting to a healthier lifestyle.

We found that current European PRS showed better discrimination than current Asian PRS in a Korean population. A previous study also reported lower effect sizes and poorer performance of Asian PRS (ORs per 1 SD: 1.10~1.41 and corresponding AUCs: 0.533~0.586) than those for the European-ancestry based 287-SNP PRS derived from 313-SNP PRS (OR per 1 SD=1.51, AUC=0.617) [24]. Similarly, discrimination of European PRS (PRS-209_{EUR}) in this study was slightly lower in Korean women (AUC=0.62, HR per SD=1.54) than that in European-ancestry women (AUC=0.63, HR per SD~1.62) [15]. These findings highlight the challenges in transferability of current PRS across different populations, which arise from the overwhelming abundance of European-descent studies and the dearth of well-powered studies in globally diverse populations. Factors such as allele frequencies, linkage disequilibrium patterns, and demographic history and natural selection different across populations could lead to such differential transferability [25, 26].

Both Asian and European PRS, either comprised of top GWAS variants or polygenic variants, improved discrimination for breast cancer in Korean women, as compared to classical risk factor models. However, the magnitude of the resulting improvement in prediction varied by PRS and age groups. A greater improvement was observed for the incorporation of European PRS vs. Asian PRS in Korean women. The magnitude of improvement by PRS was lower for women age less than 50 for which the AUC of questionnaire model was higher than those aged 50 or older (AUC=0.66 in age <50 vs. AUC=0.54 in age 50+). This smaller improvement in those younger women could reflect the inability of AUC to increase when the baseline model (i.e., questionnaire model) already has good discrimination with important risk factors [27]. Amongst the combined (Qx+PRS) models, the discrimination was higher among women of age less than 50 than the discrimination among women age 50 or greater (PRS-209 EUR: AUC=0.69 in age <50 vs.

AUC=0.60 in age 50+). In a recent pan-cancer analysis, Kachuri et al. examined a PRS based on genome-wide significant SNPs from the BCAC meta-analysis and found the difference in C-index to be 0.063 between risk factors model and risk factors plus PRS model in the UK biobank population [28]. This is similar to an average value of our findings for the corresponding PRS (GRS-136_{EUR}) among age <50 and age 50+ in the US-based model (ΔAUC = 0.04 in age<50; ΔAUC = 0.09 in age 50+).

The U.S.-based absolute risk models were well calibrated among Korean women <50 years but overestimated the risks for those age ≥50 years, even after the incorporation of PRS. Notably, the overestimation was more extreme in all U.S.-based PRS-only models. This happens because the E/O estimate in the U.S.-based PRS-only model is completely driven by the differences in average incidence between the U.S. and KCPS-II, whereas the E/O estimates for the U.S.-based Qx-only and Qx+PRS models are driven by both the difference in average incidence and the difference in risk factor distributions. It turns out the latter two differences cancel out, making the calibration for Qx-only and Qx+PRS models look better than PRS-only models ("two wrongs make a less wrong"). The overestimation was corrected by recalibrating the models using the Korean-specific inputs, which underscores the importance of tailoring absolute risk models to the target population.

Consistent with previous studies [13, 29], our analysis indicates that a larger absolute risk reduction would occur among women at higher PRS by improving their lifestyle. This suggests that focusing on high-risk individuals could yield higher benefits of preventing cancers for certain risk factor modification interventions that may not be applicable to the whole population due to cost and other considerations. Moreover, we found the 5-year absolute breast cancer risk increased with age in the most recent birth cohorts (1970-1995), with higher risks at age 40, compared to those women who were born before 1970. Given that

young Asian women today are experiencing a dramatic increase in breast cancer incidence [1], PRS could be used to adjust the optimal age for screening initiation and/or intensity to maximize the early detection of aggressive cancers, while minimizing the harms of screening in Asian women. For example, a prior study projected that, based on their PRS, 12% of Chinese women born between 1960 and 1969 had 10-year risk over 2.3% at age 40—this being a recommended risk threshold for screening initiation—whereas the proportion of women born after 1979 passing that threshold at age 40 would rise to 29% [16]. However, ultimately evidence from clinical trials will be needed to understand the true effect of an intervention for the underlying population. Three trials of personalized risk-based breast screening incorporating PRS are underway to evaluate the efficacy, safety, and acceptability of risk-based screening in the USA and Europe [30, 31]. Such clinical evaluation of PRS is urgently needed in Asian populations, where the burden of breast cancer is growing due to its dramatic increase in breast cancer incidence .

There are some limitations in our study. First, we could not consider subtypes of breast cancer since data were not available in the KCPS-II Biobank. Given the distributions of breast cancer subtypes are different between Asia and Western countries [32], the prediction capacity may differ by subtypes in different populations. Second, although we provide important insight on the predictive capacity of multiple PRS in a Korean population, our findings may not be generalizable to other Asian countries. Ho et al. reported that the distribution of this PRS was different across seven Asian countries, with the magnitude of differences consistent with genetic distance between these ethnic groups, confirming the importance of ethnic-specific calibration for valid estimation of breast cancer risk [24]. Third, we only used baseline information for first and second 5-years of follow-up and did not consider the changes in the risk factors during the follow-up period. Although this may have

caused measurement error, we anticipate the magnitude of the error would not be large for most of the variables such as age at menarche, age at first birth, and family history of breast cancer. Finally, although we had information of a large number of risk factors, we lacked data on several known modifiable and non-modifiable risk factors. For instance, breastfeeding has been found to be the strongest protective factor in Korean women, whereas in European-ancestry women, the protective effect is relatively small [33, 34]. In addition, Asian women have been reported to have denser breasts on mammography, which could increase their breast cancer risk [35, 36]. Further validation studies using more comprehensive and Asian-specific risk factor models, along with Asian-specific PRS are needed.

Our study provides a comprehensive description of the utility of genetic and modifiable risk factors for the breast cancer risk prediction in a Korean population-based cohort. We established absolute risk models to reflect the age-specific incidence rates, distribution of risk factors, and RRs in the U.S. and Korea. Moreover, we evaluated model calibration stratified by levels of risk, which can be useful for risk-based breast cancer prevention and screening by identifying individuals at the extremes of risk.

CONCLUSION

We have shown that incorporation of PRS previously developed in Asian and European-ancestry populations can improve discrimination in Korean women. Our findings suggest that PRS may be useful for motivating targeted prevention in high PRS group before they accumulate a high burden of modifiable risk factors. Further studies are needed to evaluate the value of incorporating additional information on factors into a model in ancestrally diverse populations.

FUNDING

This work was supported by grants U01 CA249866 and U01 CA261339.

REFERENCES:

1. Leong SP, Shen ZZ, Liu TJ, *et al.* Is breast cancer the same disease in Asian and Western countries? World J Surg 2010;34(10):2308-24.

2. Ko BS, Noh WC, Kang SS, *et al.* Changing patterns in the clinical characteristics of korean breast cancer from 1996-2010 using an online nationwide breast cancer database. J Breast Cancer 2012;15(4):393-400.

3. Jee YH, Gao C, Kim J, *et al.* Validating Breast Cancer Risk Prediction Models in the Korean Cancer Prevention Study-II Biobank. Cancer Epidemiol Biomarkers Prev 2020;29(6):1271-1277.

4. Shiovitz S, Korde LA. Genetics of breast cancer: a topic in evolution. Annals of oncology : official journal of the European Society for Medical Oncology 2015;26(7):1291-1299.

5. Michailidou K, Lindstrom S, Dennis J, *et al.* Association analysis identifies 65 new breast cancer risk loci. Nature 2017;551(7678):92-94.

6. Hunter DJ, Kraft P, Jacobs KB, *et al.* A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 2007;39(7):870-4.

7. Turnbull C, Ahmed S, Morrison J, *et al.* Genome-wide association study identifies five new breast cancer susceptibility loci. Nat Genet 2010;42(6):504-7.

8. Fletcher O, Johnson N, Orr N, *et al.* Novel breast cancer susceptibility locus at 9q31.2: results of a genome-wide association study. J Natl Cancer Inst 2011;103(5):425-35.

9. Zheng W, Long J, Gao YT, *et al.* Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. Nat Genet 2009;41(3):324-8.

10. 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. Nat Genet 2014;46(8):886-90.

11. Long J, Delahanty RJ, Li G, *et al.* A common deletion in the APOBEC3 genes and breast cancer risk. J Natl Cancer Inst 2013;105(8):573-9.

12. Shu X, Long J, Cai Q, *et al.* Identification of novel breast cancer susceptibility loci in meta-analyses conducted among Asian and European descendants. Nat Commun 2020;11(1):1217.

13. 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 Oncol 2016;2(10):1295-1302.

14. Garcia-Closas M, Gunsoy NB, Chatterjee N. Combined associations of genetic and environmental risk factors: implications for prevention of breast cancer. J Natl Cancer Inst 2014;106(11).

15. Mavaddat N, Michailidou K, Dennis J, *et al.* Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. Am J Hum Genet 2019;104(1):21-34.

16. Ho WK, Tai MC, Dennis J, *et al.* Polygenic risk scores for prediction of breast cancer risk in Asian populations. Genet Med (in press).

17. Choudhury PP, Maas P, Wilcox A, *et al.* iCARE: R package to build, validate and apply absolute risk models. bioRxiv 2018; 10.1101/079954:079954.

18. Jee YH, Emberson J, Jung KJ, *et al.* Cohort Profile: The Korean Cancer Prevention Study-II (KCPS-II) Biobank. Int J Epidemiol 2018;47(2):385-386f.

19. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. Nature Methods 2012;9(2):179-181.

20. Das S, Forer L, Schönherr S, *et al.* Next-generation genotype imputation service and methods. Nature Genetics 2016;48(10):1284-1287.

21. Auton A, Abecasis GR, Altshuler DM, *et al.* A global reference for human genetic variation. Nature 2015;526(7571):68-74.

22. Jee SH, Sull JW, Park J, *et al.* Body-Mass Index and Mortality in Korean Men and Women. New England Journal of Medicine 2006;355(8):779-787.

23. Ishigaki K, Akiyama M, Kanai M, *et al.* Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases. Nat Genet 2020;52(7):669-679.

24. Ho WK, Tan MM, Mavaddat N, *et al.* European polygenic risk score for prediction of breast cancer shows similar performance in Asian women. Nat Commun 2020;11(1):3833.

25. Martin AR, Gignoux CR, Walters RK, *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am J Hum Genet 2017;100(4):635-649.

26. Martin AR, Kanai M, Kamatani Y, *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet 2019;51(4):584-591.

27. Pencina MJ, D'Agostino RB, Massaro JM. Understanding increments in model performance metrics. Lifetime Data Anal 2013;19(2):202-18.

28. Kachuri L, Graff RE, Smith-Byrne K, *et al.* Pan-cancer analysis demonstrates that integrating polygenic risk scores with modifiable risk factors improves risk prediction. Nature Communications 2020;11(1):6084.

29. Parkin DM, Boyd L, Walker LC. 16. The fraction of cancer attributable to lifestyle and environmental factors in the UK in 2010. Br J Cancer 2011;105 Suppl 2(Suppl 2):S77-81.

30. Shieh Y, Eklund M, Madlensky L, *et al.* Breast Cancer Screening in the Precision Medicine Era: Risk-Based Screening in a Population-Based Trial. JNCI: Journal of the National Cancer Institute 2017;109(5).

31. Allweis TM, Hermann N, Berenstein-Molho R, *et al.* Personalized Screening for Breast Cancer: Rationale, Present Practices, and Future Directions. Annals of Surgical Oncology 2021; 10.1245/s10434-020-09426-1.

32. Lin CH, Yap YS, Lee KH, *et al.* Contrasting Epidemiology and Clinicopathology of Female Breast Cancer in Asians vs the US Population. J Natl Cancer Inst 2019;111(12):1298-1306.

33. Kim Y, Choi JY, Lee KM, *et al.* Dose-dependent protective effect of breast-feeding against breast cancer among ever-lactated women in Korea. Eur J Cancer Prev 2007;16(2):124-9.

34. Key TJ, Verkasalo PK, Banks E. Epidemiology of breast cancer. Lancet Oncol 2001;2(3):133-40.

35. Rajaram N, Mariapun S, Eriksson M, *et al.* Differences in mammographic density between Asian and Caucasian populations: a comparative analysis. Breast Cancer Research and Treatment 2017;161(2):353-362.

36. Bae J-M, Kim EH. Breast Density and Risk of Breast Cancer in Asian Women: A Meta-analysis of Observational Studies. Journal of preventive medicine and public health = Yebang Uihakhoe chi 2016;49(6):367-375.

Table 1. Mean and standard deviation of standardized polygenic risk scores with breast cancer risk

	Breast cancer events	No breast cancer
n (%)	705 (1.7)	40,326 (98.3)
Mean (SD) of age at recruitment, in years	44.09 (9.78)	41.12 (11.41)
Mean (SD) of age of diagnosis, in years	50.49 (10.03)	-
Mean (SD) of GRS-11_{ASN}	0.29 (1.02)	-0.01 (1)
Mean (SD) of PRS-42_{ASN}	0.33 (0.99)	-0.01 (1)
Mean (SD) of GRS-136_{EUR}	0.43 (1.02)	-0.01 (1)
Mean (SD) of PRS-209_{EUR}	0.43 (1.03)	-0.01 (1)

Table 2. Association of polygenic risk scores and the occurrence of breast cancer.

PRS discovery population	SNP selection	PRS	Number of SNPs published	Number of SNPs used in analyses	HR (95% CI)*	AUC (95% CI)
Asian	Genome-wide significance[1] [23]	$GRS\text{-}11_{ASN}$	11	11	1.35 (1.25 to 1.45)	0.58 (0.56 to 0.61)
Asian	Sub-genome-wide significance[2] [16]	$PRS\text{-}42_{ASN}$	46	42	1.40 (1.30 to 1.51)	0.60 (0.58 to 0.62)
European	Genome-wide significance[1] [5]	$GRS\text{-}136_{EUR}$	172	136	1.55 (1.44 to 1.66)	0.62 (0.60 to 0.64)
European	Sub-genome-wide significance[2] [15]	$PRS\text{-}209_{EUR}$	313	209	1.54 (1.43 to 1.66)	0.62 (0.60 to 0.64)

*adjusted for principal component

[1]P-value $< 5x10^{-8}$

[2]Clumping and threshold method

PRS = polygenic risk score; SNP = single-nucleotide polymorphism; HR = hazard ratio; AUC = area under the curve; CI = confidence interval.

Figure 1. Discrimination for the breast cancer risk prediction models validated.

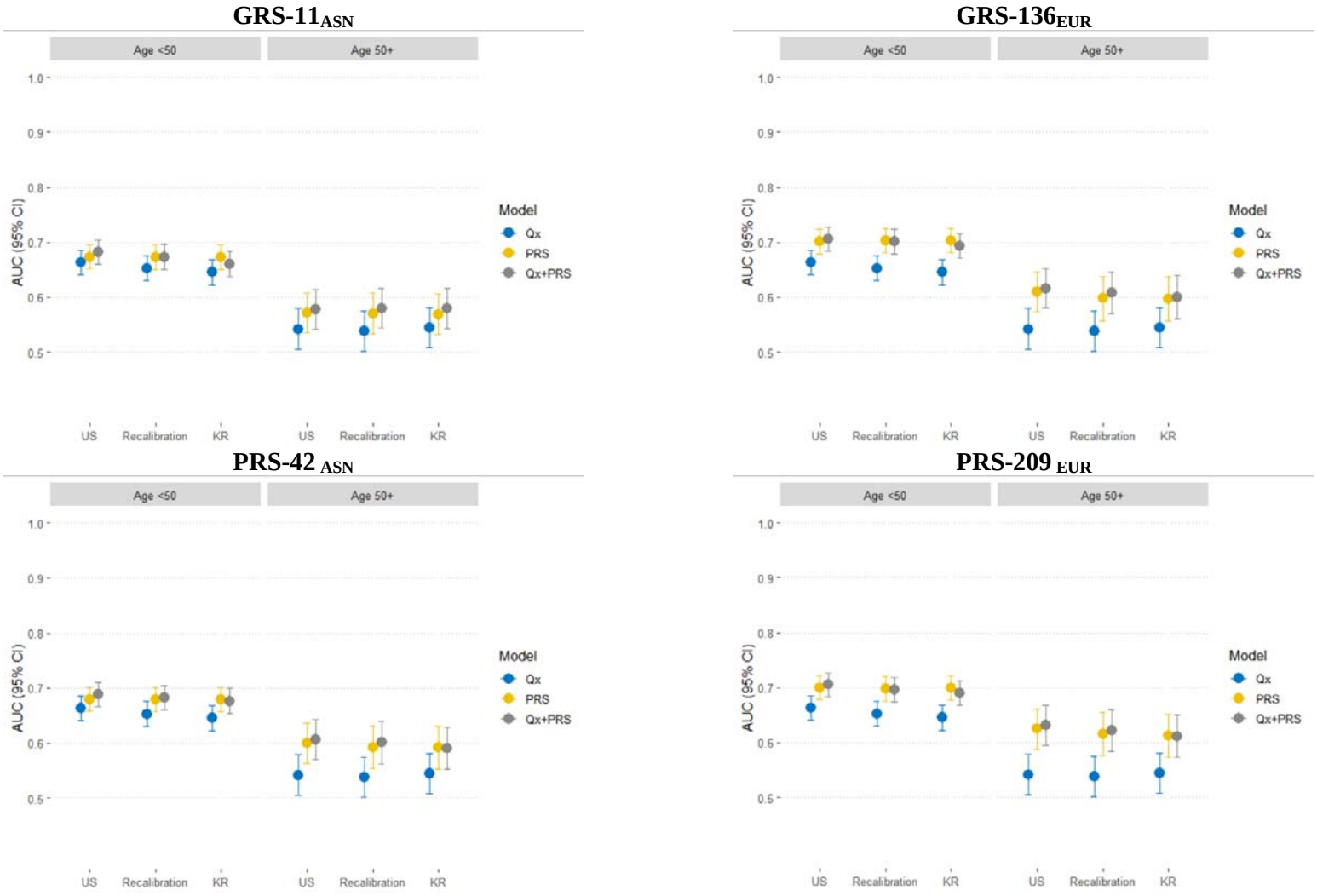


Figure 2. Calibration for the breast cancer risk prediction models validated.

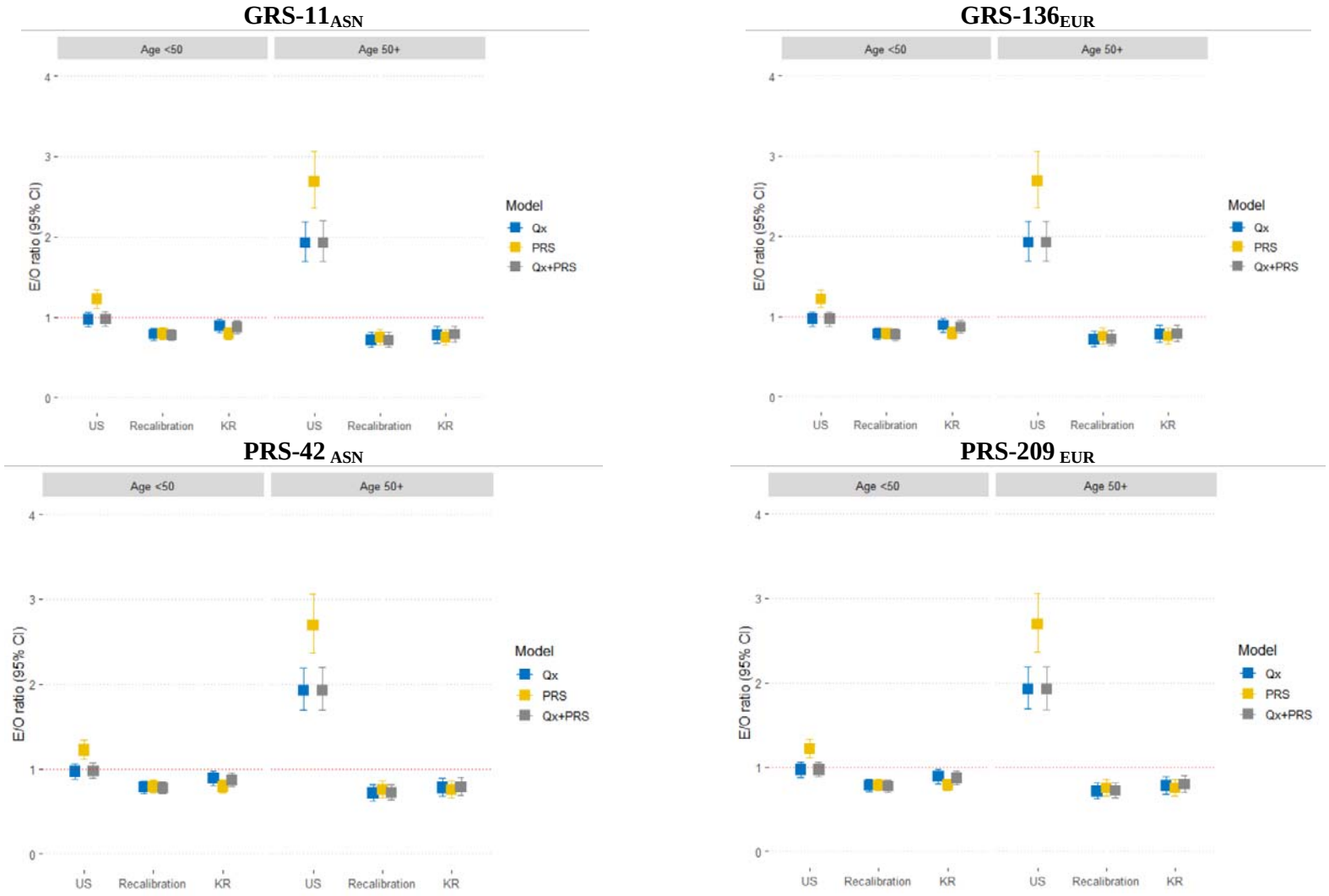
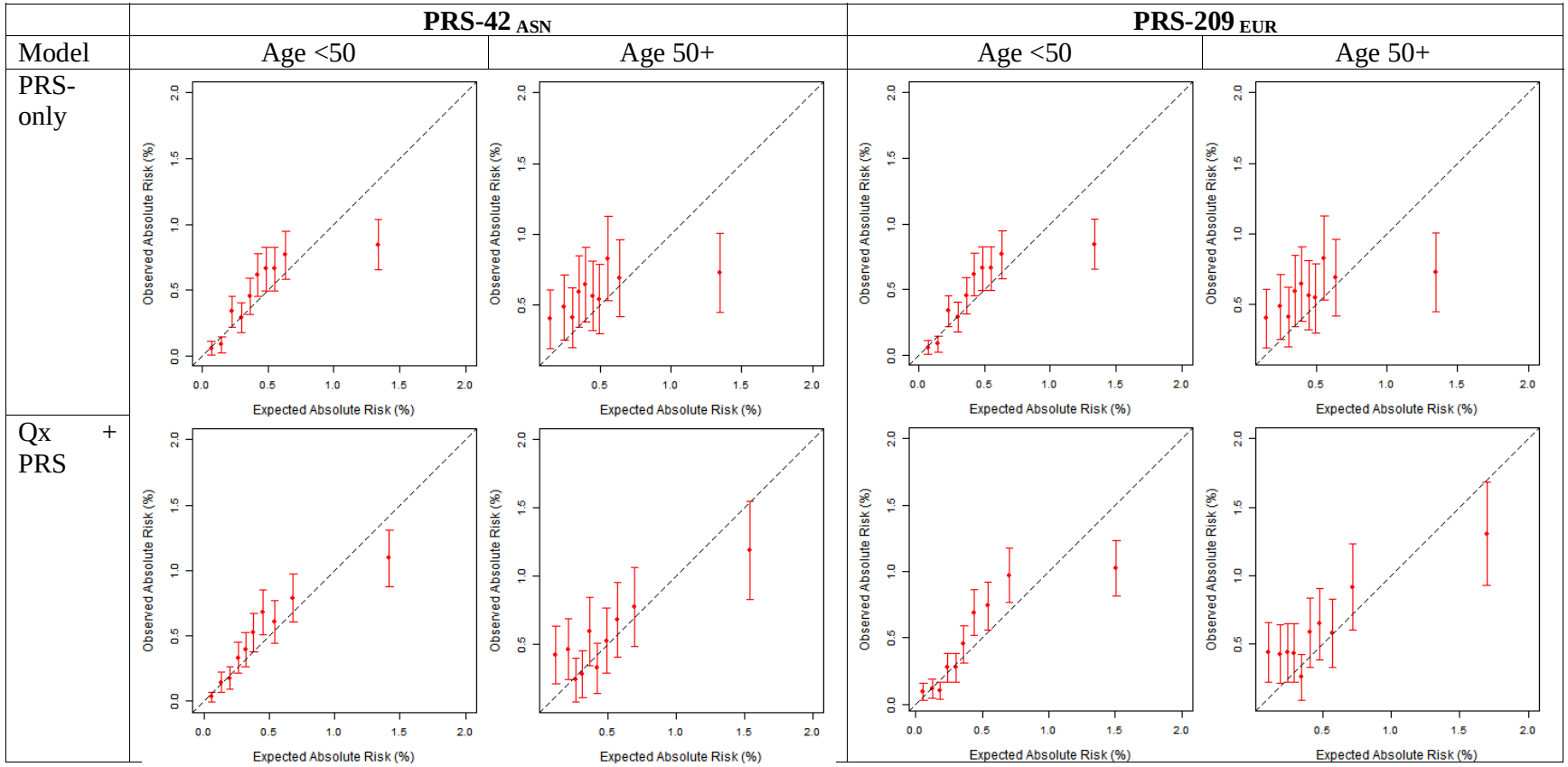
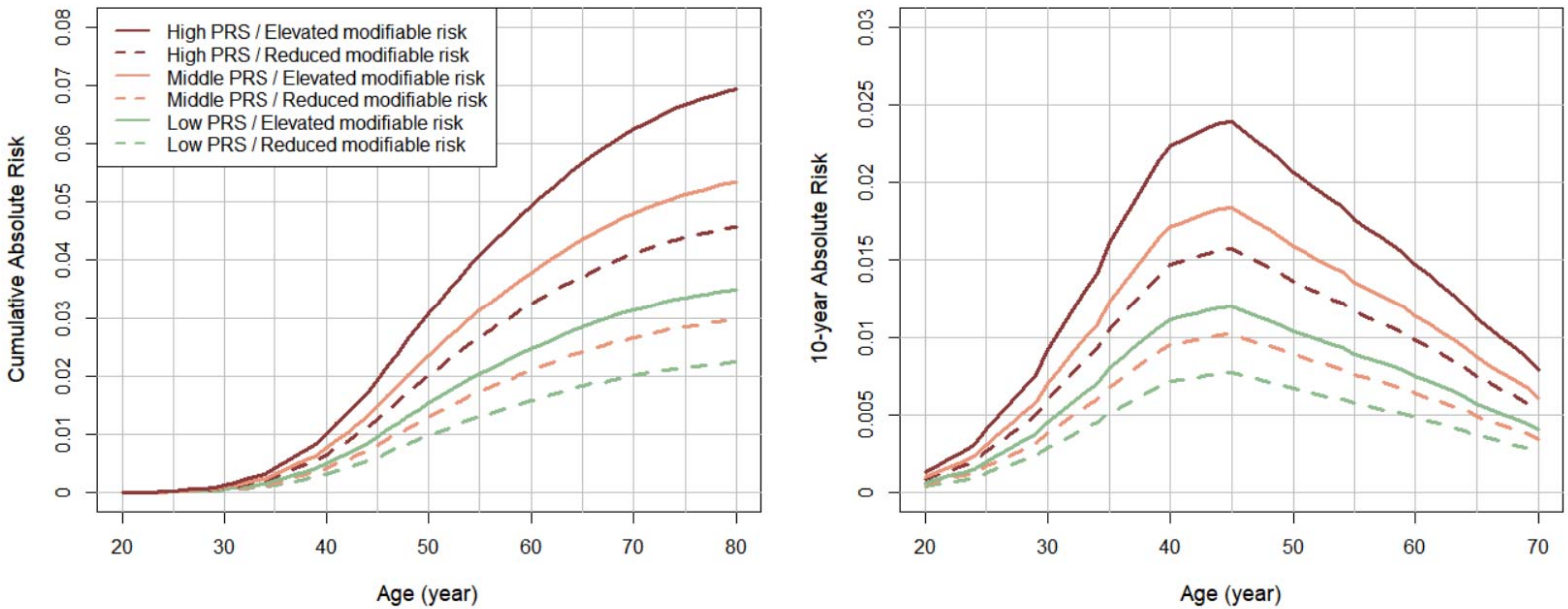


Figure 3. Absolute risk calibration for breast cancer risk prediction models in the Korean Cancer Prevention Study-II Biobank (Korean-based model*)



*Korean-based model using Korean incidence mortality and risk-factor distributions and relative risk estimates from the Korean Cancer Prevention Study. For each decile, observed absolute risks and expected absolute risks were meta-analyzed between the first and second 5-years of follow-up using inverse variance weighting method.

Figure 4. Lifetime and 10-year risk trajectories across strata defined by modifiable risk factors and percentiles of the polygenic risk score (PRS) distribution.



Modifiable risk factors include body-mass index, oral contraceptive use, alcohol intake, and hormonal replacement therapy use. (High PRS: ≥80th; Middle PRS: 20th-80th; Low PRS: ≤20th; Elevated modifiable risk: ≥median; Reduced modifiable risk: <median)