의학박사 학위논문

Development of a breast cancer risk prediction model incorporating polygenic risk scores and non-genetic risk factors for Korean women

다유전자성 위험도와 비유전적 위험요인을 결합한 한국형 유방암 예측모델의 개발

2023년 08월

서울대학교 대학원

의학과 외과학 전공

최 지 혜

다유전자성 위험도와 비유전적 위험요인을 결합한 한국형 유방암 예측모델의 개발

지도 교수 한 원 식

이 논문을 의학박사 학위논문으로 제출함
2023년 04월

서울대학교 대학원
의학과 외과학
최 지 혜

최지혜의 의학박사 학위논문을 인준함
2023년 07월

위 원 장 박수경 (인)

부위원장 한원식 (인)

위 원 이규언 (인)

위 원 문형곤 (인)

위 원 이종원 (인)

Abstract

Background: To develop a breast cancer prediction model and stratify breast cancer risks for Korean women using published polygenic risk scores (PRSs) combined with non-genetic risk factors (NGRFs).

Methods: Thirteen PRS models generated from single or multiple combinations of the Asian and European PRSs were evaluated among 20,434 Korean women. The area under the curve (AUC) and increase in odds ratio (OR) per standard deviation (SD) were compared for each PRS. The PRSs with the highest predictive performance were combined with NGRFs; then, an integrated prediction model was established using the iCARE tool. The progress in prediction power was assessed in terms of AUC and expected to observed (E/O) ratio. The absolute breast cancer risk was stratified for 18,142 women whose follow-up data were available.

Results: PRS_{38_ASN}+PRS_{190_EB}, a combination of Asian and European PRSs, had the highest AUC (0.621) among PRSs, with an OR per SD increase of 1.45 (95% CI: 1.31-1.61). Compared with the average risk group (35-65%), women in the top 5% had a 2.5-fold higher risk of breast cancer. Incorporating NGRFs yielded a modest

increase in the AUC of women aged >50 years (0.570 to 0.607 for age≥50). For PRS_{38_ASN}+PRS_{190_EB}+NGRF, the average absolute risk was 5.06%. The lifetime absolute risk at age 80 years for women in the top 5% was 9.93%, whereas that of women in the lowest 5% was 2.22%. Women at higher risks were more sensitive to NGRF incorporation.

Conclusion: Combined Asian and European PRSs were predictive of breast cancer in Korean women. Incorporation of NGRF further enhanced predictive performance in women aged >50. These findings support the use of these models for personalized screening and prevention of breast cancer.

Significance: This study provides insights into genetic susceptibility and NGRFs for predicting breast cancer in Korean women.

……

Keywords: Breast cancer; prediction model; absolute risk; polygenic risk score; non-genetic risk factor; Koreans.

Student Number: 2018-31469

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1 Background

1.1.1 Breast cancer incidence in Korea

Breast cancer is the most common cancer worldwide, and it's incidence is substantially increasing in Asian countries [1, 2]. In Korea, breast cancer incidence has increased over the last 25 years [3]. It has been on the rise since 1999 and breast cancer has become the most common cancer among women after 2016. According to the Korea National Cancer incidence database, the crude breast cancer incidence rate was 92.9 per 1,000,000 in 2019 [4]. Furthermore, it is expected to continue to grow within the next decades with the continuation of western lifestyles and changes in reproductive patterns [5]. Along with interest in these changes, there has been paramount interest in the development of a risk model to estimate and stratify an individual's susceptibility to breast cancer [6].

1.1.2 Breast cancer prediction models

Breast cancer has a multifactorial etiology resulting from a complex interaction of genetic and non-genetic risk factors

(NGRFs), such as environmental, reproductive, and lifestyle factors. A comprehensive model encompassing both genetic factors and NGRF would yield the best predictive ability. Therefore, it is essential to develop a prediction model that accurately captures risk factors.

Earlier breast cancer prediction models were based exclusively on NGRFs such as clinical factors, lifestyle factors, reproductive factors and family history. For instance, the BCRAT, a widely accepted prediction model known as the modified Gail model, originally included five risk factors (age, number of first-degree relatives with breast cancer, age at birth of first child, age at menarche and number of previous biopsies) when first developed in 1989[7]. Although some modifications were made thereafter, the discriminatory accuracy of BCRAT among different ethnicities was moderate, resulting in an AUC of 0.55 when validated in Korean women [8].

A recent analysis on the 10-year performance of breast cancer risk models including BCRAT, BRCAPRO, BOADICEA and Tyrer-Cuzick model, showed that misclassification of risks still exists. They suggested that the prediction power should be further improved, by constructing a hybrid model incorporating PRS [9].

With the leveraging advances in Genome wide association studies, increasing efforts have been made to incorporate genetic factors, such as polygenic risk scores (PRSs), into breast cancer prediction models in order to improve the predictive value of risk prediction models [10-13]. Emerging evidence suggests that PRSs, which provide a joint effect of numerous common genetic susceptibility variants, may explain a significant portion of genetic susceptibility to breast cancer [14, 15]. As expected, addition of a PRS to the existing prediction models improved AUC. It was shown that the AUC of the Gail and Tyrer-Cuzick model improved up to 0.06, when PRS was combined, depending on the model and study population [16-18]. Moreover, recent studies support the idea that a better prediction model could be achieved when NGRFs were integrated to a PRS. According to Zhang et al, the AUC for the BCRAT model improved from 0.56 to 0.65 when PRS and risk factors such as mammographic density and endogenous hormone use were incorporated [19].

1.1.3 Transferability of European models

While these findings support the potential feasibility of incorporating PRS in to breast cancer risk models, the SNPs that

constitute the PRS should be applicable to the population being examined. However, the majority of PRSs developed to date are based on European ancestry, and PRSs for Asian women have been under-evaluated, due to insufficient sample size [20-22]. One of the first attempts to solve this issue, adopted European-based PRS to Asian women. It was proved that European PRS could be applied in Asian women with better prediction power than PRS originated from Asian-specific SNPs [23]. Consequently, more studies have succeeded in integrating Asian-specific SNPs and Asian-specific weights and combining different ethnic PRSs using diverse statistical adjustments to examine the transferability of European PRSs to Asian women [12, 24]

Beyond genetic characteristics, the incidence of breast cancer, reproductive risk factor characteristics may differ significantly among ethnic groups. For instance, menopause (median 45-49 years), have continuously attributed to the highest breast cancer incidence of Korean women at the age of 40-49 years while that from the western countries peak at 60-69 years [28]. These distinct differences have hindered the transferability of western-based breast cancer prediction models to Korea as well.

In 2013, the Korean Breast Cancer Risk Assessment Tool was

developed based on the Gail model and Korean risk factor distribution [25]. This model stratified Korean-base non-genetic risk factors such as a family history of breast cancer in first-degree relatives, age at menarche, menopausal status, age at first full-term pregnancy, duration of breast feeding oral contraceptive usage and exercise, parity, BMI at the cut-off age of 50 years. The discriminatory accuracy of KoBCRAT was 0.63 for women aged under 50, and 0.65 for those over 50. However, risk prediction models incorporating genetic factors for Korean women are lacking [26, 27] and a Korean-specific risk prediction model is highly demanded.

1.2 Aims of the study

There were two aims of this study. Firstly, to develop a breast cancer PRS for Korean women using previously published PRSs for those of Asian and European ancestry. Secondly, to investigate whether the integration of NGRFs based on Korean data could improve its predictive performance.

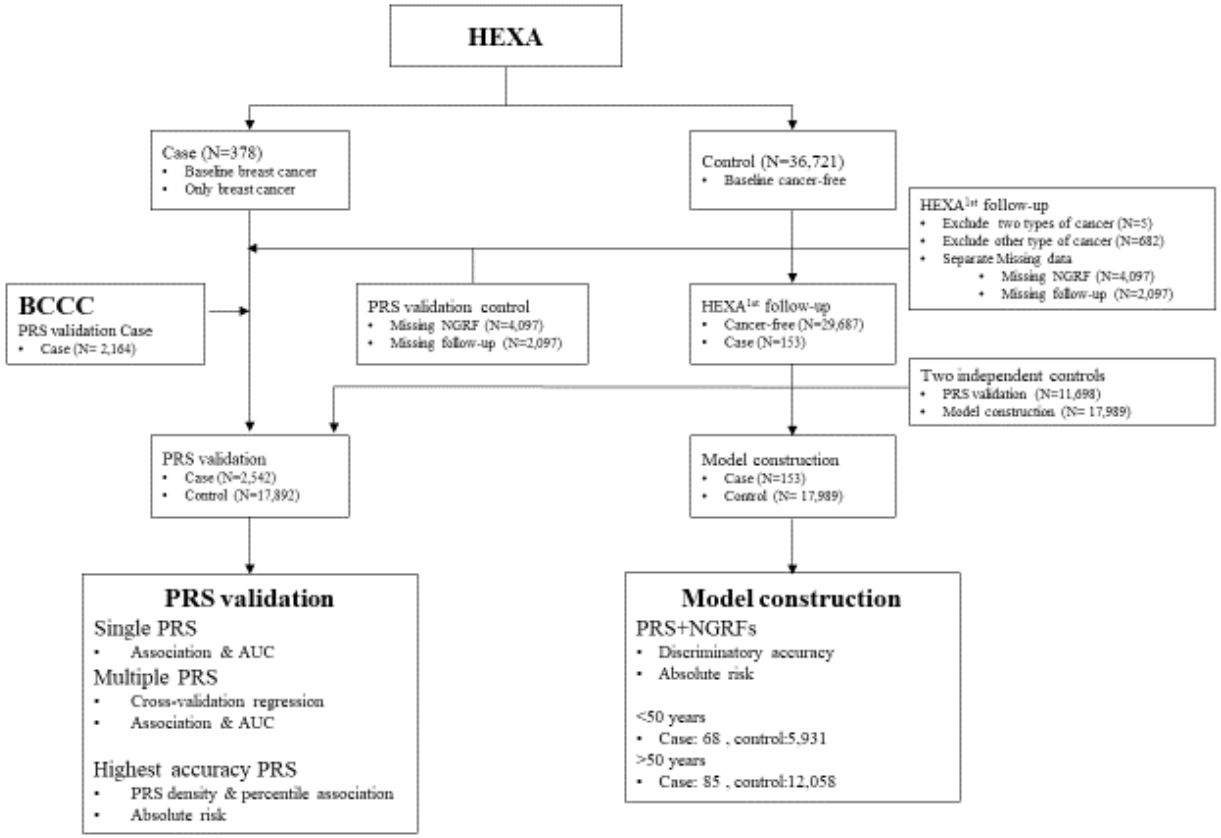
Chapter 2. Materials and Methods

2.1 Study design overview

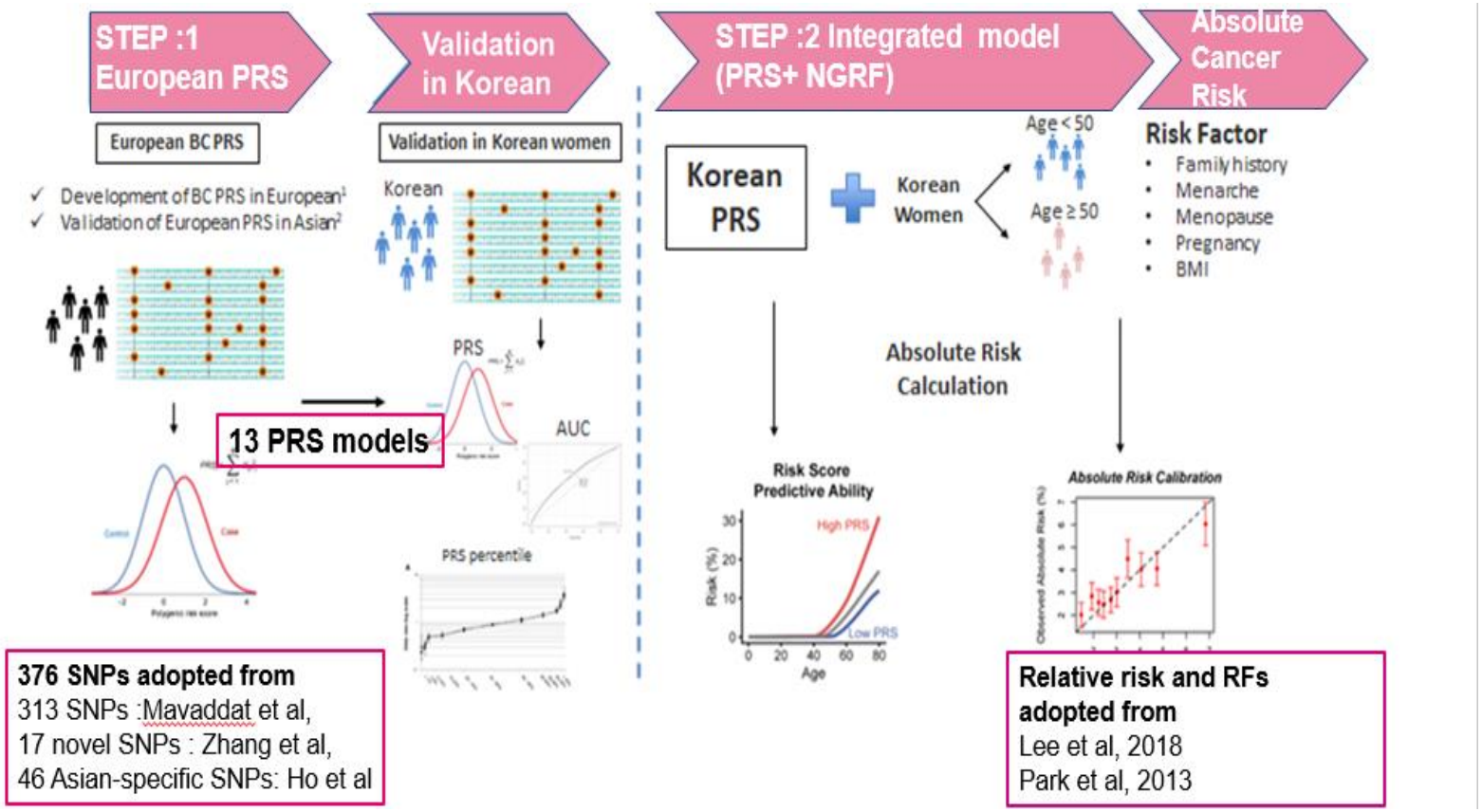
This study was conducted using two steps according to the aims of the study. First, previously reported breast cancer SNPs were validated among Korean women using various published PRSs, and absolute breast cancer risks were evaluated using the PRS with the highest accuracy (Figure 1A, left). Second, the performance of the prediction models incorporating PRSs and NGRFs were evaluated, and the absolute breast cancer risks were estimated (Figure 1A, right). The schematic diagram of the study process is shown in Figure 1B.

Figure 1. (A) Flow chart showing the study process. Previously reported breast cancer SNPs were validated among Korean women using various published PRSs, and absolute breast cancer risks were evaluated using the PRS with the highest accuracy (Figure 1, left). Prediction models incorporating PRSs and NGRFs were constructed using a different cohort (Figure 1, right). After

evaluating the predictive performance of the models, the absolute breast cancer risk was estimated.



B. Schematic diagram of study process.



2.2.1 Study populations

All participants in this study belonged to either the Health Examinee (HEXA) and Korean Association Resource (KARE) cohort of the Korean Genome and Epidemiology Study (KoGES) or the Breast Cancer Case-Cohort (BCCC). The detailed design of the KoGES study, a large cohort study with publicly available data, has been described elsewhere [29]. HEXA, initiated in 2004, recruited 173,357 participants aged >40 years from 38 health examination centers and training hospitals located in eight regions of Korea. Among them, 58,697 participants who had genotype data and met the sample quality control criteria were selected for the analyses. Samples with a low genotype call rate (<97%), cryptic relatedness, or gender discrepancy were excluded. Women who had not been diagnosed with cancer were selected for further analysis (Figure 1). Cases were defined as those who were diagnosed with breast cancer but not with other types of cancers. Controls were defined as those who were cancer-free at baseline and at the time of the follow-up surveys as well. The participants of HEXA had been followed up using active and passive methods [29]. The first follow up cohort of HEXA at a median of 4.6 ± 1.5 years, was labeled as HEXA1st. KARE, initiated in 2001, recruited 10,038 participants

aged 40–69 years from two cities, Ansan and Ansung, in Korea. Among them, 5,493 participants who had genotype data and met the sample quality criteria used for HEXA were selected. The KARE cohort was used as a reference dataset for risk factor distributions during an absolute risk estimation model construction. The BCCC was initiated in 2008, and it only recruited patients with breast cancer (N=2,165) from Seoul National University Hospital. One BCCC participant whose age at onset was >80 years was excluded from further analysis. Participants in the BCCC cohort had genotype data but lacked information about NGRFs. Therefore, genetic information about the BCCC cohort was used for PRS validation only. To perform PRS validation, we used 378 cases detected during the baseline HEXA survey. To construct the PRS+NGRF model, we used 153 cases detected at the time of the follow-up survey. Individuals who were not included in the follow-up survey (N=4,097) or had missing NGRF data (N=2,097) were excluded from the prediction model construction process. Instead, they were used as controls for PRS validation. To ensure a balanced sample distribution, two independent subsets from the HEXA controls at a 1:1 ratio to analyze 2,542 cases and 17,892 controls for final PRS validation (Figure 1) were generated. The PRS+NGRF model was

constructed using 153 cases and 17,989 cancer-free individuals based on the conditions aforementioned (Figure 1).

2.2.2 Ethics Approval and Consent to Participate

The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of National Biobank of Korea (P01-2023108-31-001) and Seoul National University Hospital (1507-132-689). Informed consent was obtained from all subjects

2.3 Genetic data acquisition

The HEXA and KARE participants were genotyped using the Korean Chip (K-CHIP), which was designed by the Center for Genome Science, Korea National Institute of Health (KNIH), based on the UK Biobank Axiom Array, and manufactured by Affymetrix [30]. BCCC participants were genotyped using the Affymetrix Genome-wide Human SNP array 6.0. We used the Michigan imputation server for phasing (via Eagle v2.4) and imputation (via minimac4) using 1000 Genomes Phase 3 data as the reference panel [31]. After imputation, we excluded SNPs with low imputation

quality scores (INFO <0.3), minor allele frequency (MAF) ≤0.01, genotype call rates (<95%), and Hardy–Weinberg equilibrium (HWE) P ≤ 1E–06.

2.4 Construction of Breast Cancer Prediction Models

2.4.1 Construction of breast Cancer PRS model

A total of 376 breast cancer–associated SNPs, including 313 SNPs selected from Mavaddat et al., 17 novel SNPs from Zhang et al., and 46 Asian–specific SNPs from Ho et al., were investigated in this study [15,23,24]. We examined seven single PRS models according to previously published beta weights of SNPs (Table 1). Single PRSs were constructed using the following equation (1):

$$\text{Single PRS}_k = \beta_1 x_1 + \beta_2 x_2 + \ldots + \ldots \beta_k x_k \quad (1),$$

where β is a coefficient representing the association between each SNP and breast cancer and k is the number of SNPs used. Of the 376 SNPs, 239 were available in the imputed genotype data of the three Korean cohorts. Accordingly, seven single PRS models were constructed. Depending on the numbers of incorporated SNPs and

types of β weights used, PRSs were denoted as PRS_{38_ASN}, PRS_{196_EUR}, PRS_{196_ASN}, PRS_{196_EB}, PRS_{201_EUR}, PRS_{201_ASN}, and PRS_{201_META}. PRS_{ASN} or PRS_{EUR} indicates that β weights of each PRS were inferred from Asian or European weights, respectively [10,15,23]. PRS_{EB} applied β weights based on a combination of Asian and European weights using the Empirical-Bayes approach [24]. PRS_{META} utilized β weights generated by a meta-analysis of European and Asian weights reported by a previous study [16] (see Table 1 for details).

Six multiple PRSs were constructed using a linear combination of the Asian and European PRSs. Multiple PRSs were constructed using equation (2):

$$\text{Multiple PRS} = \alpha_0 + \alpha_1 \cdot PRS_{38\_ASN} + \alpha_2 \cdot PRS_{\_EUR} \quad (2)$$

α_0, α_1, and α_2 were obtained by fitting a logistic regression model with breast cancer incidence as the outcome. PRSs were standardized to the respective standardized deviations of the HEXA controls. Ten-fold cross-validation by regression was conducted for multiple PRS models. The relative contributions of each PRS to multiple PRS models are shown in Table 2.

Table 1. Number of breast cancer associated SNPs and beta weights used to construct single PRS models

Single PRS	Original nSNPs[b]	Selected nSNPs[a]	Beta weight
PRS_{38_ASN}[c]	46	38	Asian SNPs and Asian weights reported in Ho et al [23].
PRS_{190_EUR}	313	190	European SNPs and European weights reported in Mavaddat et al [31].
PRS_{190_ASN}	313	190	European SNPs and Asian weights reported in Ho et al.
PRS_{190_EB}	313	190	European SNPs and EB (Empirical Bayes) weights reported in Ho et al [23].
PRS_{201_EUR}	330	201	European SNPs and European weights reported in Mavaddat et al. and Zhang et al [32].
PRS_{201_ASN}	330	201	European SNPs and Asian weights reported in Yang et al [24].
PRS_{201_META}	330	201	European SNPs and meta-analysis between European and Asian SNPs reported in Yang et al [37].

[a] Selected SNPs with imputation score (INFO) ≥ 0.7, minor allele frequency (MAF) ≥ 0.01, (HWE) ≥ 1E-06 and call rate ≥ 0.95.[b] SNP: single nucleotide polymorphism

[c] Depending on the numbers of incorporated SNPs and types of β weights used, PRS was denoted as PRS_{38_ASN}, PRS_{196_EUR}, PRS_{196_ASN}, PRS_{196_EB}, PRS_{201_EUR}, PRS_{201_ASN}, and PRS_{201_META}

Table 2. Weights used for constructing multiple PRS models

PRS combination	α_0 [a]	α_1 [a]	α_2 [a]	W [b]
$\alpha_0 + \alpha_1 PRS_{38\_ASN} + \alpha_2 PRS_{190\_EUR}$	−2.045	0.181	0.317	0.363
$\alpha_0 + \alpha_1 PRS_{38\_ASN} + \alpha_2 PRS_{190\_ASN}$	−2.041	0.118	0.335	0.260
$\alpha_0 + \alpha_1 PRS_{38\_ASN} + \alpha_2 PRS_{190\_EB}$	−2.050	0.148	0.346	0.300
$\alpha_0 + \alpha_1 PRS_{38\_ASN} + \alpha_2 PRS_{201\_EUR}$	−2.047	0.180	0.321	0.359
$\alpha_0 + \alpha_1 PRS_{38\_ASN} + \alpha_2 PRS_{201\_ASN}$	−2.038	0.121	0.327	0.270
$\alpha_0 + \alpha_1 PRS_{38\_ASN} + \alpha_2 PRS_{201\_META}$	−2.044	0.137	0.336	0.290

[a] Multiple PRSs have been constructed using formula $\alpha_0 + \alpha_1$ Asian PRS + α_2 European PRS. α_0, α_1, and α_2 were obtained by fitting a logistic regression model with breast cancer incidence as outcome. [b] Contribution of Asian PRS to the linear combination. $W = \alpha_1 / (\alpha_1 + \alpha_2)$ and $(1 - W)$ represents the contribution of European PRS to the linear combination. PRSs were standardized to respective standardized deviation of the controls in the HEXA.

2.4.2 Models incorporating non-genetic risk factors

For the PRS models developed during step one, NGRFs were incorporated to establish an integrated risk prediction model. Depending on the menopausal status, the incidence of breast cancer differs in Korea, and risk factors (RFs) linked to the development of breast cancer exert varying effects [25, 27]. Thus, the PRS+NGRF models were constructed separately using the cut-off age of 50 years by applying different relative risks (RRs) and RFs. Information about estrogen-dependent NGRFs in the HEXA and KARE were taken from the survey data. The BMI measured at the time of enrollment (average age, 53±8.37 years) was used. Breast cancer-associated NGRFs and respective RRs were obtained from external studies [25, 27]. For women aged <50 years, age at menarche, familial history of breast cancer, menopausal status, age at first full-term pregnancy, height, and BMI were included. For women aged ≥50 years, age at menopause and pregnancy experience (nullipara or para; Table 3) were additionally included, whereas age at first full-term pregnancy and menopausal status were excluded. Table 3 provides a description of the RR estimates used in this study. In all prediction models incorporating NGRF

scores, equation (3) was used, where F_k and w_k are the value and corresponding weight of factor k, respectively:

$$NGRF\ Score = \sum_{k=0}^{6} w_k F_k \quad (3)$$

2.5 Statistical analysis

2.5.1 PRS association analysis and evaluation of predictive performance

For PRS association analyses, we used logistic regression adjusted for covariates. We examined the odds ratio (OR) per standard deviation (SD) of the PRS for seven percentile groups (0-5%, 5-15%, 15-35%, 35-65%, 65-85%, 85-95%, 95-100%), with 35-65% being the average risk group.

The prediction performance of the PRS was measured by the area under the receiver-operating characteristic curve (AUC) using logistic regression. To compare the predictive function of either NGRF-based or PRS-based models and integrated models (PRS+NGRF), the AUC and expected to observed (E/O) ratios were evaluated

Table 3. Relative risks of NGRFs and participant distributions for breast cancer

Breast cancer risk factors	HEXA[1st c] Age<50 (N=5,999)	HEXA[1st c] Age≥50 (N=12,143)	Relative risk (range) Age < 50	Relative risk (range) Age ≥ 50
Age at baseline, years				
Mean (±SD)	44.67 ± 3.13	57.34 ± 5.32	–	–
On set age of breast cancer				
Mean (±SD)	47.49 ± 3.08	60.44 ± 5.9	–	–
Age at menarche, years [a]				
≤10	7 (0.001)	9 (0.001)	1.27 (1.23 −1.31)	1.35 (1.25−1.46)
11	73 (0.012)	52 (0.004)	1.13 (1.09 −1.16)	1.16 (1.08−1.25)
12	418 (0.07)	320 (0.026)	1.13 (1.09 −1.16)	1.16 (1.08−1.25)
13	1071 (0.179)	1098 (0.09)	Ref.	Ref.
14	1721 (0.287)	2160 (0.178)	1.00 (0.97 −1.03)	1.00 (0.93−1.08)
15	1438 (0.24)	2702 (0.223)	0.89 (0.86 −0.92)	0.86 (0.80−0.93)
≥16	1271 (0.212)	5802 (0.478)	0.79 (0.77 −0.81)	0.74 (0.69−0.80)
Breast cancer family history in first degree [b]				
No	5888 (0.981)	11880 (0.978)	Ref.	Ref.
Yes	111 (0.019)	263 (0.022)	1.12 (0.81 −1.56)	2.01 (1.28−3.31)
Menopause [b]				
Premenopausal	5609 (0.935)	–	1.74 (1.442−2.14)	–
Postmenopausal	390 (0.065)	–	Ref.	–
Age at menopause (year) [b]				
Premenopausal	–	3309 (0.273)		2.50 (1.78−3.51)
<44	–	286 (0.024)	–	Ref.
45−49	–	2404 (0.198)	–	1.34 (0.99−1.83)
50−54	–	5149 (0.424)	–	1.36 (1.01−1.82)
≥50	–	995 (0.082)	–	1.62 (1.09−2.39)
Age at first full−term pregnancy (year)[b]				
Nullipara	93 (0.016)	–	1.08 (0.80−1.45)	–
<24	828 (0.138)	–	Ref.	–
24−30	4434 (0.739)	–	1.16 (0.97−1.39)	–
≥30	644 (0.107)	–	1.25 (0.93−1.69)	–
Pregnancy [b]				
Nullipara	–	120 (0.01)	–	1.88 (1.24−2.84)
Para	–	12023 (0.99)	–	Ref.
Height, m [a]				
Mean (±SD)			1.20 (1.15 −1.27)	1.24 (1.16−1.33)
BMI, (kg/m2) [a]				
<18.5	164 (0.027)	8284 (0.682)	0.98 (0.93 −1.03)	Ref.
18.5 − <25	4538 (0.756)		Ref.	Ref.
25 − <30	1179 (0.197)	3482 (0.287)	1.02 (0.97 −1.07)	1.35 (1.22−1.49)
≥30	118 (0.02)	377 (0.031)	1.04 (0.99 −1.10)	1.82 (1.65−2.01)

[a] Relative risk of NGRFs used in KCPS cohort [19], [b] Relative risk of risk factors used in KoBCRAT model [17]

[c] HEXA[1st]: 1st follow−up cohort of the Health Examinee cohort, SD: standard deviation, Ref: reference

2.5.2 Estimation of the absolute risk of breast cancer according to PRS percentiles

For the PRS showing the highest prediction accuracy, the lifetime absolute risks of breast cancer were estimated. Furthermore, using an integrated PRS+NGRF model, the lifetime and 5-year absolute breast cancer risks were recalculated. The absolute risk of breast cancer for women of age α over the time interval $\alpha + \tau$ was defined according to equation (4):

$$R_{\alpha,\alpha+\tau} = \int_{\alpha}^{\alpha+\tau} \lambda_0(t) \exp(\beta^T Z) \exp(-\int_{\alpha}^{\tau} [\lambda_0(u) \exp(Z\beta) + m(u)] du) dt \quad (4)$$

Equation (4) assumes that risk factor (RF) Z acts in a multiplicative fashion on the baseline hazard function $\lambda_0(t)$. It accounts for competing risks originating from mortality due to other causes through m(t), the age-specific mortality rate function. The lifetime absolute risk was evaluated as the risk between the age of 20 years and a specific age with a maximum of 80 years. The 5-year absolute risk was defined as the risk within the next 5 years for a woman who has reached a specific age. The iCARE tool requires the RRs of RFs (Z), log-relative risks (β), age-specific incidence rate of all-cause mortality excluding breast cancer mortality, incidence rates of breast cancer, and RF distributions within a population. For this study, RRs were obtained from external

studies [17, 19], and RF distributions were derived from KARE, which was used as a reference cohort. The age-specific breast cancer incidence and mortality rates of Korean women in 2010 were obtained from the Korean Statistical Information Service [26]. Absolute risks were evaluated with R 4.2.1 using the Individualized Coherent Absolute Risk Estimation (iCARE) R package (version 1.18.0) [27]. $P<0.05$ was considered significant.

To investigate the associated effect of PRS and NGRF according to the magnitude of risks, absolute lifetime risks using multiple PRS and NGRF risk strata were analyzed. PRSs were classified into three risk groups (0-20%: low; 20-80%: mid; 80-100%: high), and NGRF scores were classified into two groups divided using a median distribution (0-50%: low; 50-100%: high).

Chapter 3. Results

3.1 Study population

PRS validation was performed among 20,434 Korean women (Figure 1). The PRS+NRGF model was evaluated among 18,142 cancer-free individuals. In this subset, 153 cancer cases were detected during the follow-up period. Among 153 newly developed cases, 68 occurred in women aged <50 years and 85 occurred in those aged >50 years.

3.2 Performance of PRS in the Korean population

Thirteen PRSs were constructed using previously reported Asian and European SNPs (Table 4). In general, the multiple PRS models performed better than the single PRS models for Korean women. Among the PRS models, the most predictive was PRS_{38_ASN}+PRS_{190_EB} (AUC: 0.621), although the overall AUC differences between PRS models were marginal. The contribution of PRS_{38_ASN} to PRS_{38_ASN}+PRS_{190_EB} was approximately 30% (Table 2). We did not observe a significant interaction between PRS_{38_ASN}+PRS_{190_EB} and age (Table 5). The density plot of

Table 4. Mean, SD, and the associations of the PRS with the breast cancer risks for Korean women

PRS	HEXA[a] + BCCC[b] Case (N = 2,542) Mean ± (SD)	Control (N = 17,892) Mean ± (SD)	OR per SD[c] (95% CI)	AUC[d] (95% CI)
Single PRS				
PRS_{38_ASN}	−0.10 ± 0.41	−0.24 ± 0.41	1.37 (1.24−1.52)	0.592 (0.581−0.604)
PRS_{190_EUR}	0.68 ± 0.51	0.47 ± 0.50	1.38 (1.24−1.53)	0.611 (0.599−0.623)
PRS_{190_ASN}	−0.06 ± 0.56	−0.29 ± 0.54	1.41 (1.27−1.56)	0.612 (0.600−0.624)
PRS_{190_EB}	0.24 ± 0.46	0.05 ± 0.45	1.41 (1.27−1.56)	0.616 (0.604−0.627)
PRS_{201_EUR}	0.46 ± 0.52	0.25 ± 0.50	1.37 (1.23−1.51)	0.612 (0.600−0.624)
PRS_{201_ASN}	0.03 ± 0.55	−0.19 ± 0.54	1.41 (1.28−1.56)	0.612 (0.600−0.624)
PRS_{201_META}	0.61 ± 0.59	0.37 ± 0.57	1.41 (1.28−1.57)	0.614 (0.603−0.626)
Multiple PRS				
PRS_{38_ASN}+PRS_{190_EUR}	−1.85 ± 0.21	−1.94 ± 0.20	1.44 (1.3−1.59)	0.619 (0.607−0.631)
PRS_{38_ASN}+PRS_{190_ASN}	−2.07 ± 0.22	−2.17 ± 0.21	1.44 (1.3−1.59)	0.615 (0.604−0.627)
PRS_{38_ASN}+PRS_{190_EB}	−1.98 ± 0.20	−2.07 ± 0.19	1.45 (1.31−1.61)	0.621 (0.609−0.633)
PRS_{38_ASN}+PRS_{201_EUR}	−1.92 ± 0.21	−2.01 ± 0.21	1.43 (1.29−1.58)	0.620 (0.608−0.631)
PRS_{38_ASN}+PRS_{201_ASN}	−2.04 ± 0.21	−2.13 ± 0.21	1.44 (1.30−1.59)	0.615 (0.603−0.627)
PRS_{38_ASN}+PRS_{201_META}	−1.85 ± 0.23	−1.95 ± 0.23	1.45 (1.31−1.60)	0.618 (0.607−0.630)

[a] HEXA: Health Examinee, [b] BCCC: Breast cancer case−cohort, [c] OR: odds ratios were estimated using a logistic regression model adjusted for age and study, SD: standard deviation, [d] AUC: Area under the curve.

Table 5. Association between PRS_{38_ASN}+PRS_{190_EB} and breast cancer risks in different age groups

Age category	OR (95% CI) [a]	*P*−value	N (Case / Control)
40−50	1.39 (1.16−1.67)	2.95E−04	1,128/6,255
50−60	1.54 (1.29−1.85)	2.42E−06	684/7,702
60−70	1.44 (1−20.080)	5.06E−02	259/3,746

[a] OR, 95% CI and P values were estimated using a logistic regression model adjusted for age and study

PRS_{38_ASN}+PRS_{190_EB} is shown in Figure 2. The distribution curve for cancer participants shifted to the right compared with that of the controls. The percentile association of PRS_{38_ASN}+PRS_{190_EB} stratified into seven percentile groups is presented in Table 6. Women in the top 5% had a 2.5−fold higher risk and women in the lowest 5% had a 0.61−fold lower risk of breast cancer than the average risk group (35−65%). The risk distributions were well−distinguished between the risk percentile groups, although the associations were not statistically significant. Lifetime and 5−year absolute risks of PRS_{38_ASN}+PRS_{190_EB} are shown in Figure 3. The lifetime absolute risk at age 80 years for women in the highest 5%

was 9.91%, and that of women in the lowest 5% was 2.18%

(average lifetime absolute risk: 4.89%).

Table 6. Percentile association of PRS_{38_ASN}+PRS_{190_EB} and distributions of participants

	Sample size				
PRS (percentiles)	BCCC [a] –Case	HEXA [b] –Case	HEXA–Control	OR (95% CI) [c]	*P*–value
00–05(%)	47	11	964	0.61 (0.31–1.09)	1.20E–01
05–15(%)	125	20	1,898	0.57 (0.34–0.90)	2.11E–02
15–35(%)	290	63	3,734	0.91 (0.66–1.24)	5.39E–01
35–65(%)	610	101	5,419	Reference	–
65–85(%)	534	94	3,459	1.46 (1.10–1.94)	9.35E–03
85–95(%)	320	54	1,669	1.74 (1.24–2.42)	1.17E–03
95–100(%)	238	35	749	2.50 (1.67–3.67)	4.52E–06

[a] BCCC: Breast cancer case–cohort, [b] HEXA: Health Examinee cohort, [c] Odds ratios were estimated using a logistic regression model adjusted for age and study, CI: confidence interval.

Figure 2. Density plot of PRS_{38_ASN}+PRS_{190_EB} among participants with (case) and without breast cancer (control) showing PRS distribution.

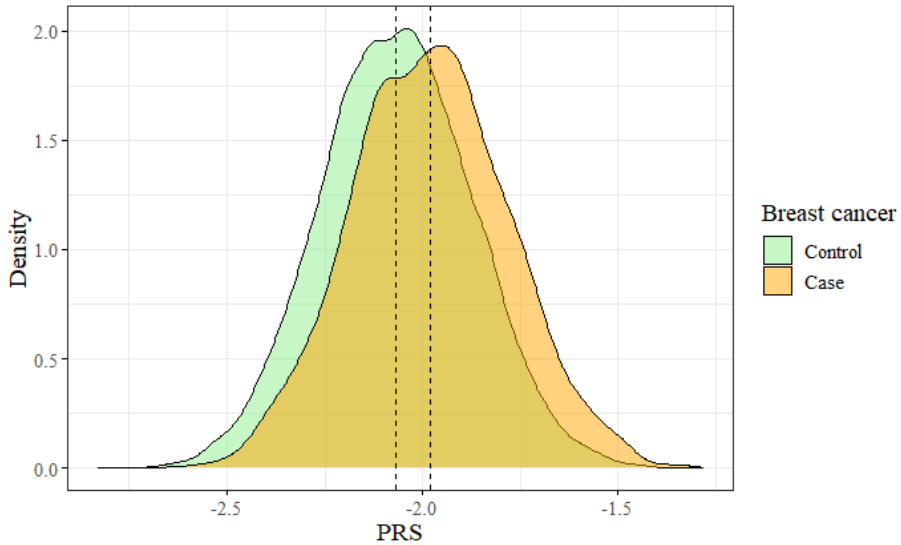
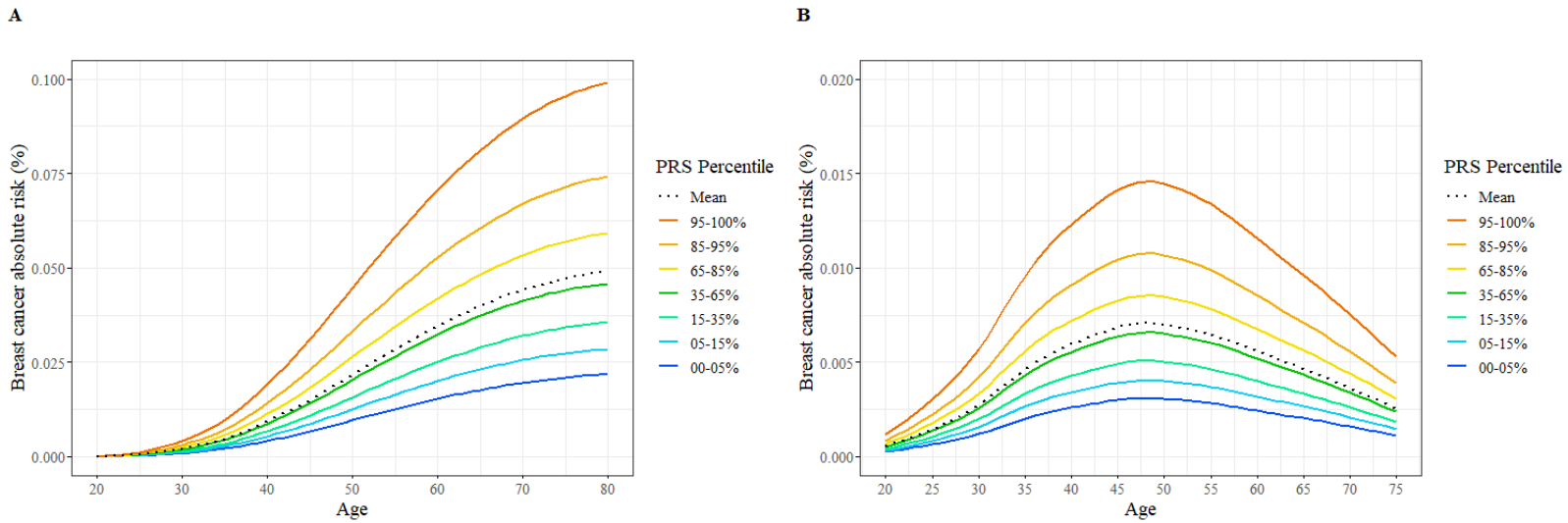


Figure 3. Absolute risk of developing breast cancer estimated using data from 20,434 Korean women. Lifetime absolute risk (A) and 5-year absolute risk (B) of breast cancer estimated by PRS_{38_ASN}+PRS_{190_EB} for women at different age categories and risk percentiles. Dotted lines represent the average risks.

A. Lifetime absolute risk

B. 5-year absolute risk



3.3 Performance of the prediction model incorporating non-genetic risk factors

The prediction model established using NGRFs had limited predictive power (Figure 4, Table 7). NGRF models were more predictive for women aged ≥50 years than for those aged <50 years (AUC 0.564 vs. 0.503). The results showed that there were noticeable differences in the AUC changes when NGRF models were added to PRS models, depending on the age group. For women aged ≥50 years, the addition of NGRFs led to an increase in the AUC, although the initial PRS had a lower AUC. However, for participants aged <50 years, the incorporation of NGRFs had only a small effect on the AUC, whereas PRS alone had better predictive performance. This implies that women aged ≥50 are more dependent on NGRFs, whereas women aged <50 are more genetically predisposed. PRS_{38_ASN}+PRS_{201_META}+NGRF had the highest predictive power for women aged <50 years, whereas PRS_{38_ASN}+PRS_{190_EB}+NGRF was the most predictive model for women aged ≥50 years. Nevertheless, because there was not much difference in the overall AUC (0.012) between the best and worst multiple PRSs for all age groups, we decided to use PRS_{38_ASN}+PRS_{190_EB}+NGRF (hereafter referred to as the integrated

Figure 4. Area under the curve (AUC) for various PRS models and NGRFs predicting the breast cancer risk. The AUC was compared among NGRFs, PRSs, and PRS+NGRF (integrated) models for women aged <50 years (A) and ≥50 years (B) resepectively.

A. Age <50, B. Age ≥50

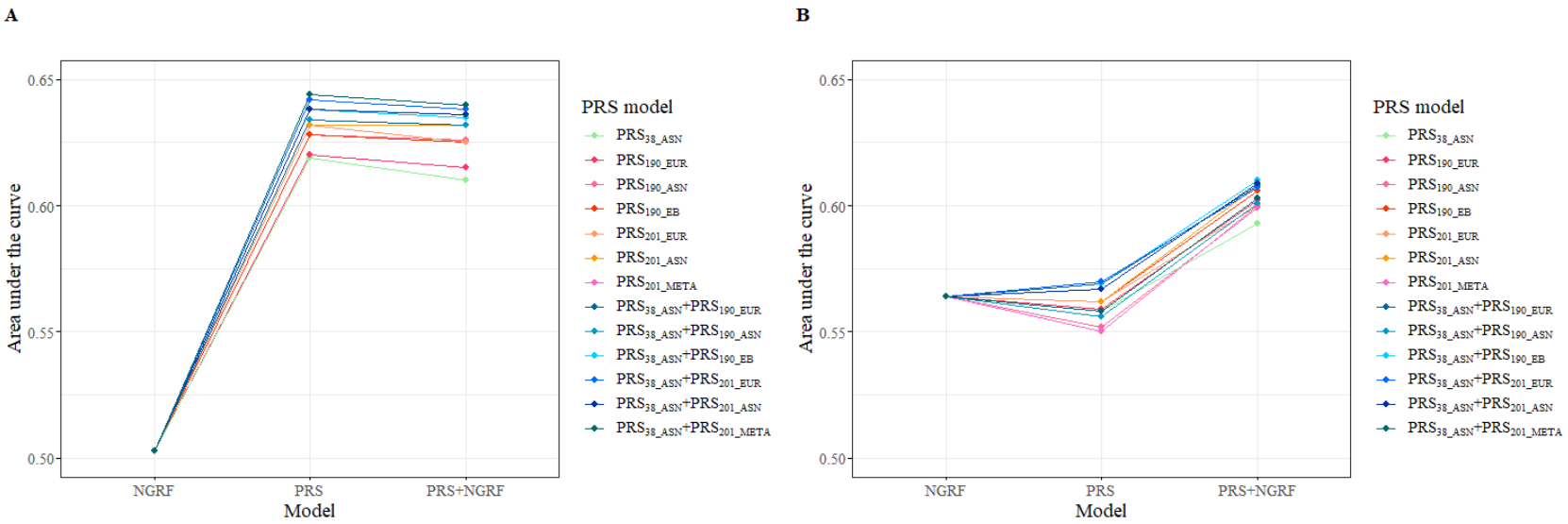


Table 7. Comparison of AUC and E/O ratio between NGRF, PRS and PRS+NGRF models

	AUC [a] (95% CI)				E/O[b] (95% CI)			
	Age < 50		Age ≥ 50		Age < 50		Age ≥ 50	
NGRF	0.503 (0.403-0.572)		0.564 (0.498-0.630)		0.478 (0.376-0.607)		0.740 (0.594-0.922)	
	PRS	PRS + NGRF	PRS	PRS + NGRF	PRS	PRS + NGRF	PRS	PRS + NGRF
Single PRS								
PRS_{38_ASN}	0.619 (0.545-0.694)	0.610 (0.541-0.678)	0.558 (0.487-0.629)	0.593 (0.531-0.654)	0.504 (0.396-0.64)	0.535 (0.421-0.68)	0.776 (0.623-0.967)	0.833 (0.668-1.037)
PRS_{190_EUR}	0.62 (0.553-0.687)	0.615 (0.547-0.683)	0.559 (0.493-0.624)	0.602 (0.544-0.66)	0.455 (0.358-0.578)	0.484 (0.381-0.615)	0.694 (0.557-0.864)	0.739 (0.593-0.92)
PRS_{190_ASN}	0.628 (0.555-0.702)	0.626 (0.553-0.699)	0.552 (0.482-0.622)	0.599 (0.536-0.661)	0.45 (0.354-0.572)	0.478 (0.376-0.608)	0.692 (0.556-0.862)	0.737 (0.592-0.918)
PRS_{190_EB}	0.628 (0.559-0.697)	0.625 (0.556-0.695)	0.562 (0.494-0.63)	0.606 (0.548-0.665)	0.453 (0.356-0.576)	0.482 (0.379-0.613)	0.692 (0.556-0.862)	0.736 (0.591-0.917)
PRS_{201_EUR}	0.632 (0.565-0.699)	0.625 (0.557-0.693)	0.562 (0.496-0.627)	0.600 (0.540-0.660)	0.455 (0.358-0.578)	0.484 (0.381-0.615)	0.694 (0.557-0.865)	0.740 (0.594-0.922)
PRS_{201_ASN}	0.632 (0.560-0.704)	0.632 (0.560-0.703)	0.562 (0.492-0.633)	0.608 (0.546-0.670)	0.453 (0.357-0.576)	0.481 (0.379-0.612)	0.694 (0.557-0.864)	0.740 (0.594-0.922)
PRS_{201_META}	0.638 (0.569-0.708)	0.636 (0.566-0.706)	0.550 (0.481-0.618)	0.600 (0.539-0.661)	0.453 (0.356-0.576)	0.482 (0.379-0.612)	0.693 (0.556-0.863)	0.739 (0.593-0.921)
Multiple PRS								
$PRS_{38\_ASN}+PRS_{190\_EUR}$	0.634 (0.563-0.704)	0.632 (0.563-0.700)	0.569 (0.499-0.638)	0.608 (0.549-0.666)	0.495 (0.389-0.629)	0.526 (0.414-0.669)	0.756 (0.607-0.942)	0.807 (0.648-1.006)
$PRS_{38\_ASN}+PRS_{190\_ASN}$	0.634 (0.561-0.707)	0.632 (0.560-0.704)	0.556 (0.485-0.627)	0.601 (0.539-0.663)	0.479 (0.377-0.609)	0.509 (0.400-0.647)	0.737 (0.592-0.918)	0.786 (0.631-0.979)
$PRS_{38\_ASN}+PRS_{190\_EB}$	0.638 (0.567-0.708)	0.635 (0.566-0.704)	0.569 (0.498-0.639)	0.61 (0.551-0.668)	0.491 (0.386-0.624)	0.522 (0.411-0.664)	0.751 (0.603-0.936)	0.801 (0.643-0.998)
$PRS_{38\_ASN}+PRS_{201\_EUR}$	0.642 (0.572-0.712)	0.638 (0.570-0.707)	0.570 (0.501-0.639)	0.607 (0.547-0.666)	0.493 (0.388-0.627)	0.524 (0.413-0.667)	0.754 (0.605-0.94)	0.805 (0.646-1.003)
$PRS_{38\_ASN}+PRS_{201\_ASN}$	0.638 (0.566-0.710)	0.636 (0.565-0.707)	0.567 (0.495-0.638)	0.609 (0.548-0.671)	0.483 (0.38-0.614)	0.513 (0.403-0.652)	0.74 (0.594-0.922)	0.791 (0.635-0.985)
$PRS_{38\_ASN}+PRS_{201\_META}$	0.644 (0.573-0.715)	0.640 (0.570-0.710)	0.558 (0.488-0.628)	0.603 (0.542-0.664)	0.484 (0.381-0.615)	0.514 (0.405-0.654)	0.741 (0.595-0.924)	0.792 (0.635-0.986)

[a] AUC: Area Under the Curve, [b] Expected/Observed (E/O) ratio. E: expected 5-year absolute risk, O: observed 5-year incidence.

model), which is the model containing PRS_{38_ASN}+PRS_{190_EB} (hereafter referred to as the multiple PRS model), which had the highest accuracy during step one, to further estimate the absolute breast cancer risk.

3.4 Absolute risk of breast cancer according to PRS percentiles

Figure 5 depicts the lifetime and 5-year absolute risks of the integrated model. The lifetime absolute risk of breast cancer ranged from 2% to 10%, with an average of 5.06% (Figure 5A). The absolute risk at age 80 years for women in the highest 5% was 9.93%, whereas that for women in the lowest 5% was at 2.22%. The zenith of the 5-year absolute risk for women in the top 5% at age 48 years was 1.47%, and it declined thereafter (Figure 5B). The 5-year absolute risk of the average risk group at 40 years, which is the age when the first breast cancer screening program is recommended in Korea, was 0.6%. However, women in the top 5% risk reached this level of risk much earlier at age 33 years. This outcome may support the need for individualized screening strategies for high-risk women, particularly for those aged <40 years.

Figure 5. Estimation of the absolute breast cancer risk by seven percentiles. Lifetime absolute risk (A) and 5-year absolute risk (B) of developing breast cancer predicted using PRS_{38_ASN}+PRS_{190_EB}+NGRF (integrated model.) Dotted lines represent the average risks.

A. Lifetime absolute risk

B. 5-year absolute risk

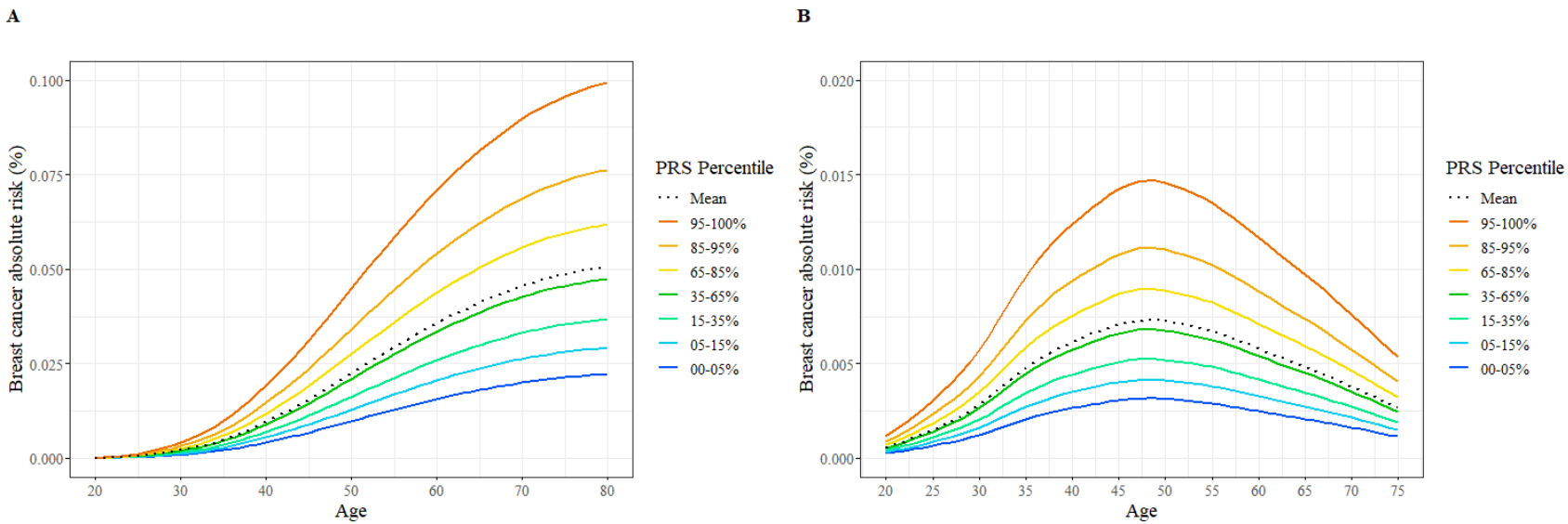
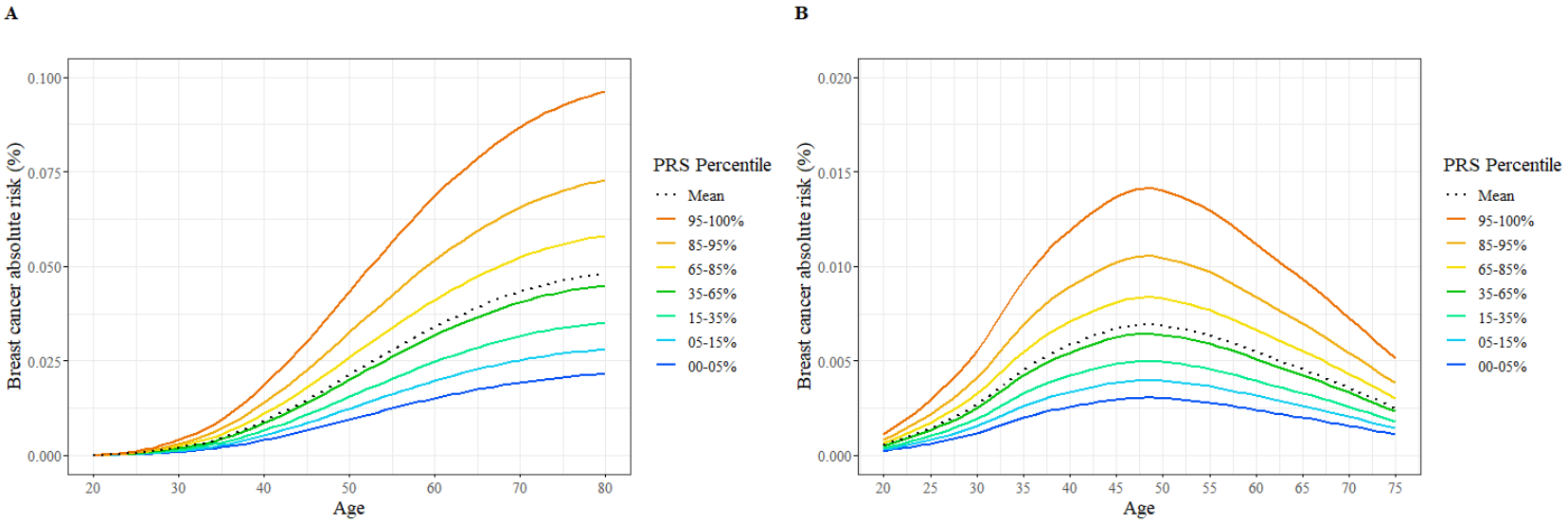


Figure 6. Absolute risk of breast cancer estimated at different ages and risk percentiles using data from HEXA [1st]. Lifetime absolute risk (A) and 5-year absolute risk (B) of developing breast cancer predicted by PRS_{38_ASN}+PRS_{190_EB} (multiple PRS model). Dotted lines represent the average risks.

A. Lifetime absolute risk B. 5-year absolute risk



To compare the changes induced by incorporating NGRFs, the lifetime and 5-year absolute risks of the multiple PRS model were also estimated using HEXA 1st (Figure 6). The average absolute risks of the integrated model (5.06%) and multiple PRS model (4.81%) were quite similar. However, the distributions of density in both models were different. Figure 7 depicts the density plot of the multiple PRS and integrated models at age 80 years. Adding NGRFs to multiple PRS models increased the SD of the integrated model. A greater increase in the mean was observed for the higher-risk group, especially those in the top 5% (Table 8). This implies that women at higher risk are more sensitive to risk-reducing interventions.

To investigate the associated effect of PRS and NGRF, we analyzed the absolute lifetime risks using different PRS and NGRF risk levels (Figure 8). In the model, the curves for mid PRS+high NGRF versus high PRS+low NGRF as well as the curves for mid PRS+low NGRF versus low PRS+high NGRF nearly overlapped. This supports the idea that risk modifications might reduce the breast cancer risk of some individuals despite their inherited genetic risks. The difference in the absolute risk with high NGRF and low NGRF levels was greater for women at higher risk,

indicating that there is greater potential for risk reduction among this group of women.

Table 8. Mean, SD of PRS_{38_ASN}+PRS_{190_EB}(multiple PRS model) and of PRS_{38_ASN}+PRS_{190_EB}+NGRF (integrated model) at age 80 at seven percentiles

PRS percentiles	Multiple PRS model Mean[a] (SD[b])	Integrated model Mean (SD)	Sample size (N)
00−05 (%)	0.0215 (0.0025)	0.0222 (0.0079)	907
05−15 (%)	0.0280 (0.0018)	0.0292 (0.0099)	1,814
15−35 (%)	0.0350 (0.0024)	0.0368 (0.0140)	3,629
35−65 (%)	0.0450 (0.0037)	0.0474 (0.0174)	5,442
65−85 (%)	0.0581 (0.0040)	0.0618 (0.0236)	3,629
85−95 (%)	0.0727 (0.0045)	0.0762 (0.0269)	1,814
95−100 (%)	0.0962 (0.0152)	0.0993 (0.0369)	907

[a] Mean: Average absolute risk, [b] SD: Standard deviation.

Figure 7. Density plot showing absolute breast cancer risk for the multiple PRS model (A) and integrated model (B) at age 80 years, stratified by seven PRS percentiles.

A. Multiple PRS model

B. Integrated model

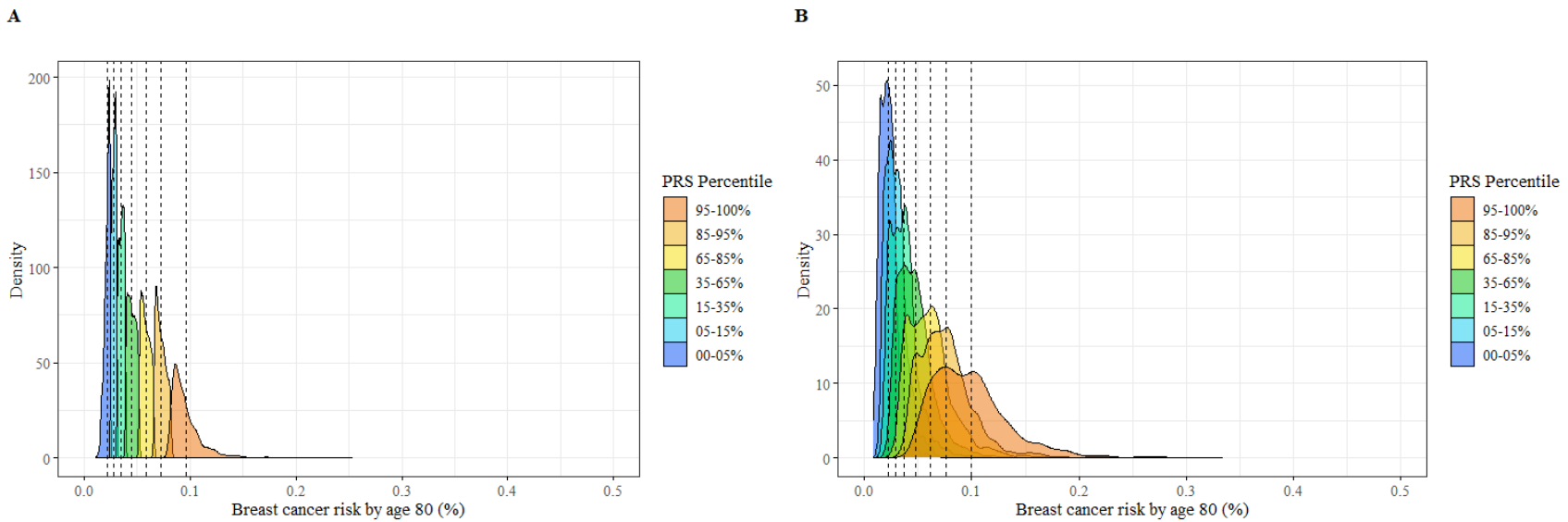
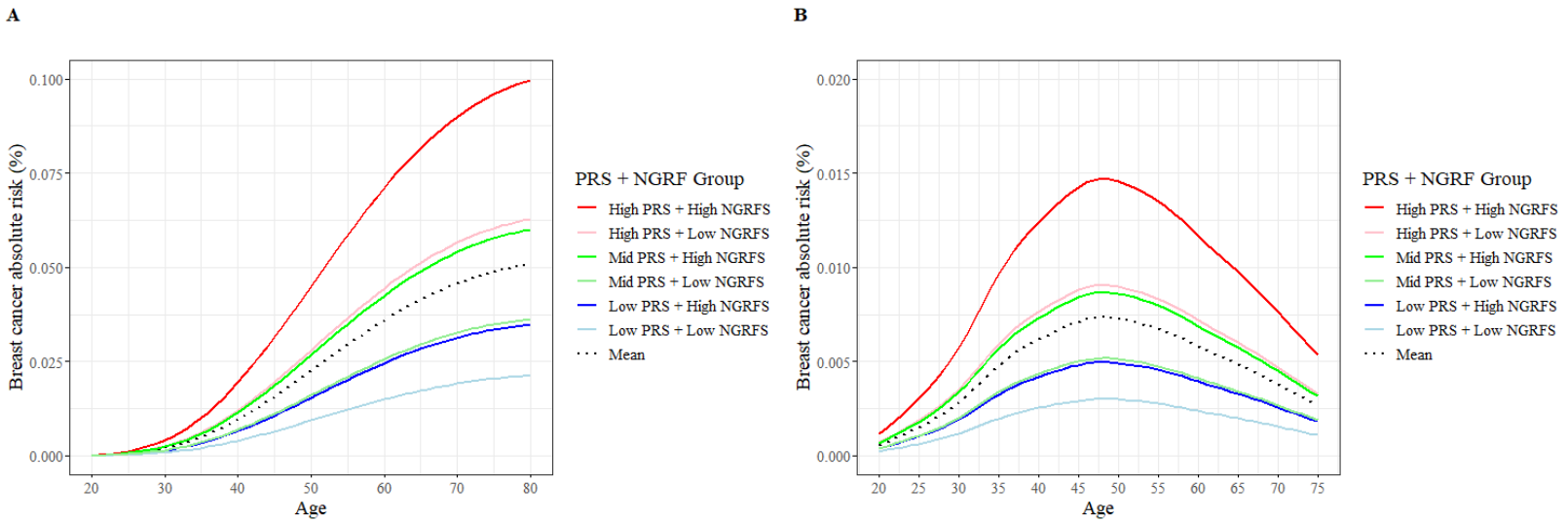


Figure 8. Lifetime absolute risk (A) and 5-year absolute risk (B) estimated using different combinations of PRS and risk levels of NGRF. PRS was classified into three risk groups according to percentile distributions (0-20%: low, 20-80%: mid, 80-100%: high). NGRF was classified into two levels divided at median (0-50%: low, 50-100%: high). Dotted lines represent the average risks.

A. Lifetime absolute risk

B. 5-year absolute risk



Chapter 4. Discussion

In this study, we developed a breast cancer prediction model for Korean women based on 13 PRS and NGRFs. We demonstrated that i) the combined Asian and European PRS was predictive of breast cancer among Korean women and ii) the incorporation of NGRFs improved breast cancer risk stratification for women aged ≥50 years. The findings of this study provide essential insights into genetic susceptibility and NGRFs for predicting breast cancer among Korean women.

Previous studies have reported a lesser AUC for European-ancestry based PRS in Asian women, raising the issue of transferability among Asian population. The predictive value of European-ancestry base PRS (PRS_{201_EUR}) in this study was 0.612, which is slightly lower than that of European-ancestry (0.63) [32]. However, combined Asian and European PRS performed better in Korean women (AUC 0.615-0.621). These results are consistent to the findings of Ho et al, which have proved that combined Asian and European PRS (PRS_{333}) could enhance breast cancer risk stratification for Asian women (AUC 0.621) [24].

To my knowledge, very few studies of the Korean PRS have

been conducted, and most of them were conducted as a part of large-scale Asian studies [22-24]. In a previous analysis, a PRS constructed with 44 SNPs among East Asian women in the Breast Cancer Association Consortium was examined among Korean women. In that study, although the Korean-specific AUC was not provided, the overall AUC was 0.606 [22], which is consistent with the AUC of PRS_{38_ASN} in this study (AUC: 0.592). In addition, one recent study explored a combined Asian and European PRS (PRS_{46}+PRS_{287_EB}) for Asian women, and an AUC of 0.630 (OR per SD: 1.59) was reported for a subset of Korean women [24]. The current study is one of the few studies to establish a breast cancer risk prediction model incorporating PRSs for Korean women only. In this study, the AUC of the PRS with the highest predictive power was 0.621. This is consistent with those of published studies that examined PRSs for Asian women [21, 24, 36, 37]. In addition, the increase in AUC resulting from combining NGRFs with multiple PRSs was comparable to the findings of previous studies (Table 7). Most studies report a modest increase in AUC (0.01 to 0.10) depending on the numbers, weights of SNPs and combinations of various risk factors [11-13, 38].

In this study, two separate models incorporating NGRFs

were established according to age using different RFs and RRs. In Korea, menopause is an important RF contributing to a distinctive breast cancer incidence curve that peaks at age 50 years and declines thereafter, as shown in this study (Figure 3) [6, 27]. For this reason, Korean prediction models often have been used to analyze age groups separately at the cut-off age of 50 [25-27]. In this study, it was found that the contributions of NGRFs and PRSs to the prediction of breast cancer were distinctively different among age groups (Figure 2, Table 7). For those aged <50 years, the AUC of the PRS alone was initially higher compared to that of women aged >50 years. Adding NGRFs did not contribute to an increase in the AUC of this group. These findings suggest that young women have higher genetic susceptibility to breast cancer than their older counterparts and indicate that the contribution of NGRF in this age group is relatively small. An analysis that evaluated the interactions between the PRS and NGRFs showed a stronger association between the PRS and premenopausal women (OR: 2.46), thus supporting the findings of this study [39]. In contrast to the findings of this study, the magnitude of AUC improvement by incorporating PRS was lower for women age less than 50 (AUC 0.66) than for those aged 50 or over (AUC 0.54) in one study based on the

questionnaire model. It was interpreted that the baseline questionnaire model contained important risk factors in advance. However, the predictive performances of risk prediction models still were higher among women of age less than 50, than their counterparts [40].

In this study, the performance of the prediction models among patients aged <50 years was greater than it was among women aged ≥50 years. Perhaps, risk prediction for patients aged ≥50 years is further complicated by menopause and confounding factors such as BMI. In a previous study, a prediction model constructed using iCARE based on the Korean incidence of mortality and Korean-based risk distributions showed an AUC of 0.584 for women aged ≥50 years and an AUC of 0.697 for women aged <50 years [27]. The study suggested that that adopting Korean risk distributions is especially important for age≥50. Korean prediction model utilizing European risk distributions have overestimated breast cancer risk of women age≥50 (E/O=2.472) while E/O was markedly improved after re-estimating with Korean based risk distributions (E/O=1.018). The difference resulted from the tendency to have later menarche, earlier menopause, later age at first birth, and lower BMI in Korean women. On the other hand,

the Korean Breast Cancer Risk Assessment Tool incorporated RFs such as age at menopause, pregnancy experience, BMI, oral contraceptive usage, and exercise, unlike that for premenopausal women, yielding an AUC of 0.65 for women aged ≥50 [25]. In this study, reproductive RFs were selected as the main NGRFs. The effects of risks related to estrogen-dependent reproductive factors, including age at menarche and number of pregnancies, on breast cancer development have been clearly established [5], whereas the effects of lifestyle factors (alcohol intake, hormone use, and exercise) are complicated and still controversial [41-46]. In addition, owing to the characteristics of questionnaire-based surveys, self-reported lifestyle RFs that change over time are associated with a high risk of confounding the study by creating recall bias or misclassification [47]. For instance, distinction between current hormone use or past history of hormonal use and duration is not clear. "Missing", or "No", "Unknown" values could not be differentiated. Therefore, it was managed to build a consistent model based mainly on reproductive risk factors.

The results of this study support the usage of personalized screening and prevention strategies based on PRS risk stratification. To date, most guidelines have been updated to

implement personalized screening for those at high risk, notably those with a family history or identified pathogenic variants [48, 49]. However, the current Korean national breast cancer screening program follows a one-size-fits-all approach, encouraging women aged 40 to 69 years in the general population to undergo a biennial mammography based on the evidence that screening reduces breast cancer mortality in this age group [50]. However, the optimal age for initiating screening may be adjusted based on the results of this study. This study suggests that earlier screening should be implemented for women in the high-risk group, particularly for those aged <40 years who are currently ineligible for screening programs. In addition, preventive interventions and lifestyle modifications for women with higher risk scores should be focused on to yield maximum risk reductions [51]. For instance, lifestyle changes affecting BMI may modify the breast cancer risk.

However, the best screening tool for breast cancer in young women and its effect on reducing real-world breast cancer mortality should be carefully evaluated before implementation [45]. In western countries, clinical trials evaluating the efficacy and feasibility of personalized screening programs incorporating PRS

are already ongoing [52–54]. Evidence from such trials could help comprehend the genuine effect of PRS based screening programs on reality.

This study had several limitations. First, the small sample size limited the optimization of the prediction model. A sufficient sample size may have provided significant p–values for all risk estimates in the multiple PRS model (Table 4). Also, the inclusion of few participants aged >70 years (1.57%) may have under–represented the risk estimates of women in this age group. However, this distribution was similar to data from the hospital–based nationwide cancer registry of Korea, in which the proportion of breast cancer diagnosed over 70 years of age is 1.3%. The average life expectancy of Koreans is increasing and reached 84 years in 2021 [53]. Further accumulation of follow–up data may enable us to develop a more accurate model that covers all age strata. In addition, the overestimated E/O ratio of PRS_{196_EB} + PRS_{38}+NGRF in age≥50 may have approximated to a better outcome. The E/O of all models among age≥50 were over 1 while in women aged <50, E/O ranged between 0.7–0.8. Presumably, this result is due to the discrepancy of breast cancer incidence over number of controls in this age group. The ratio of controls in age

<50 versus ≥50 was nearly 1:2 while only few more cancer diagnosis was announced in age≥50. When this discrepancy was adjusted, the E/O of age≥50 were similar to those in age <50. Second, because the BCCC and HEXA cohorts of case-control studies were recruited from either a teaching hospital or a health examination center in an urban area, their characteristics may not be representative of the entire Korean population, thus inducing selection bias. Further external validation among the general population is warranted. Third, this model does not include a modifiable RF. Further adjustments are needed to include information about modifiable RFs, and other RFs associated with breast cancer risk, such as mammographic density, might improve the predictive power [54, 55]. Finally, although common variants were included in the current model, pathogenic variants that are known to confer a higher risk of breast cancer susceptibility may be included in future prediction models.

Chapter 5. Conclusion

In conclusion, this study showed that the combined PRS is predictive of breast cancer in Korean women. The incorporation of NGRFs further enhanced the predictive power for women aged >50 years. These models can be helpful to developing optimal screening strategies for and effective preventive measures against breast cancer

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2018;68(6):394-424.

2. Ghoncheh M, Mahdavifar N, Darvishi E, Salehiniya H. Epidemiology, incidence and mortality of breast cancer in Asia. Asian Pacific journal of cancer prevention. 2016;17(sup3):47-52.

3. Hong S, Won Y-J, Park YR, Jung K-W, Kong H-J, Lee ES. Cancer statistics in Korea: incidence, mortality, survival, and prevalence in 2017. Cancer research and treatment: official journal of Korean Cancer Association. 2020;52(2):335-50.

4. Jung KW, Won YJ, Kong HJ, Lee ES. Prediction of cancer incidence and mortality in Korea, 2019. Cancer Res Treat 2019;51 :431-7.

5. Lee JE, Lee SA, Kim TH, Park S, Choy YS, Ju YJ, et al. Projection of breast cancer burden due to reproductive/lifestyle changes in Korean women (2013-2030) using an age-period-cohort model. Cancer research and treatment: official journal of

Korean Cancer Association. 2018;50(4):1388-95.

6. Park HL. Breast cancer risk prediction in Korean women: review and perspectives on personalized breast cancer screening. Journal of breast cancer. 2020;23(4):331.

7. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 1989;81:1879-86.

8. Min JW, Chang MC, Lee HK, Hur MH, Noh DY, Yoon JH, et al. Validation of risk assessment models for predicting the incidence of breast cancer in Korean women. J Breast Cancer 2014;17:226-35

9. Terry MB, Liao Y, Whittemore AS, Leoce N, Buchsbaum R, Zeinomar N, et al. 10-year performance of four models of breast cancer risk: a validation study. Lancet Oncol 2019;20:504-17

10. Rudolph A, Song M, Brook MN, Milne RL, Mavaddat N, Michailidou K, 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):526-36.

11. van Veen EM, Brentnall AR, Byers H, Harkness EF, Astley

SM, Sampson S, et al. Use of single-nucleotide polymorphisms and mammographic density plus classic risk factors for breast cancer risk prediction. JAMA oncology. 2018;4(4):476-82.

12. Lakeman IM, Rodríguez-Girondo M, Lee A, Ruiter R, Stricker BH, Wijnant SR, 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):1803-11.

13. Shieh Y, Hu D, Ma L, Huntsman S, Gard CC, Leung JW, et al. Breast cancer risk prediction using a clinical risk model and polygenic risk score. Breast cancer research and treatment. 2016;159(3):513-25.

14. Mavaddat N, Antoniou AC, Easton DF, Garcia-Closas M. Genetic susceptibility to breast cancer. Molecular oncology. 2010;4(3):174-91.

15. Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nature genetics. 2013;45(4):353-61.

16. Dite GS, MacInnis RJ, Bickerstaffe A, Dowty JG, Allman R, Apicella C, et al. Breast cancer risk prediction using clinical models

and 77 independent risk-associated SNPs for women aged under 50 years: Australian Breast Cancer Family Registry. Cancer Epidemiol Biomarkers Prev 2016;25:359-65. PUBMED | CROSSREF

17. Allman R, Dite GS, Hopper JL, Gordon O, Starlard-Davenport A, Chlebowski R, et al. SNPs and breast cancer risk prediction for African American and Hispanic women. Breast Cancer Res Treat 2015;154:583-9. PUBMED | CROSSREF

18. Starlard-Davenport A, Allman R, Dite GS, Hopper JL, Spaeth Tuff E, Macleod S, et al. Validation of a genetic risk score for Arkansas women of color. PLoS One 2018;13:e0204834.

19. Zhang X, Rice M, Tworoger SS, Rosner BA, Eliassen AH, Tamimi RM, 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 Med 2018;15:e1002644.

20. Chan CHT, Munusamy P, Loke SY, Koh GL, Yang AZY, Law HY, et al. Evaluation of three polygenic risk score models for the prediction of breast cancer risk in Singapore Chinese. Oncotarget. 2018;9(16):12796.

21. Lee CPL, Irwanto A, Salim A, Yuan J-m, Liu J, Koh WP, et

al. Breast cancer risk assessment using genetic variants and risk factors in a Singapore Chinese population. Breast Cancer Research. 2014;16(3):1-13.

22. Wen W, Shu X-o, Guo X, Cai Q, Long J, Bolla MK, et al. Prediction of breast cancer risk based on common genetic variants in women of East Asian ancestry. Breast Cancer Research. 2016;18(1):1-8.

23. Ho W-K, Tai M-C, Dennis J, Shu X, Li J, Ho PJ, et al. Polygenic risk scores for prediction of breast cancer risk in Asian populations. Genetics in Medicine. 2022;24(3):586-600.

24. Yang Y, Tao R, Shu X, Cai Q, Wen W, Gu K, et al. Incorporating polygenic risk scores and nongenetic risk factors for breast cancer risk prediction among Asian women. JAMA network open. 2022;5(3):e2149030-e.

25. Park B, Ma SH, Shin A, Chang M-C, Choi J-Y, Kim S, et al. Korean risk assessment model for breast cancer risk prediction. PLoS One. 2013;8(10):e76736.

26. Lee C, Lee JC, Park B, Bae J, Lim MH, Kang D, et al. Computational discrimination of breast cancer for Korean women based on epidemiologic data only. Journal of Korean medical science. 2015;30(8):1025-34.

27. Jee YH, Gao C, Kim J, Park S, Jee SH, Kraft P. Validating breast cancer risk prediction models in the Korean Cancer Prevention Study-II Biobank. Cancer Epidemiology, Biomarkers & Prevention. 2020;29(6):1271-7.

28. Kim Y, Han B-G, Group K. Cohort profile: the Korean genome and epidemiology study (KoGES) consortium. International journal of epidemiology. 2017;46(2):e20-e.

29. Moon S, Kim YJ, Han S, Hwang MY, Shin DM, Park MY, et al. The Korea Biobank Array: design and identification of coding variants associated with blood biochemical traits. Scientific reports. 2019;9(1):1-11.

30. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nature genetics. 2016;48(10):1284-7.

31. Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, et al. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. The American Journal of Human Genetics. 2019;104(1):21-34.

32. Zhang H, Ahearn TU, Lecarpentier J, Barnes D, Beesley J, Qi G, et al. Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific

analyses. Nature genetics. 2020;52(6):572-81.

33. https://kosis.kr/index/index.do

34. Pal Choudhury P, Maas P, Wilcox A, Wheeler W, Brook M, Check D, et al. iCARE: An R package to build, validate and apply absolute risk models. PloS one. 2020;15(2):e0228198.

35. Hsieh Y-C, Tu S-H, Su C-T, Cho E-C, Wu C-H, Hsieh M-C, et al. A polygenic risk score for breast cancer risk in a Taiwanese population. Breast cancer research and treatment. 2017;163(1):131-8.

36. Zheng W, Wen W, Gao Y-T, Shyr Y, Zheng Y, Long J, et al. Genetic and clinical predictors for breast cancer risk assessment and stratification among Chinese women. Journal of the National Cancer Institute. 2010;102(13):972-81.

37. Anastasiadi Z, Lianos GD, Ignatiadou E, Harissis HV, Mitsis M. Breast cancer in young women: an overview. Updates in surgery. 2017;69(3):313-7.

38. Gómez-Flores-Ramos L, Álvarez-Gómez RM, Villarreal-Garza C, Wegman-Ostrosky T, Mohar A. Breast cancer genetics in young women: What do we know? Mutation Research/Reviews in Mutation Research. 2017;774:33-45.

39. Shi M, O' Brien K, Weinberg C. interactions between a

polygenic Risk Score and non-genetic Risk factors in Young-onset Breast cancer. Scientific reports. 2020;10(1):1-7.

40. Jee YH, Ho WK, Park S, Easton DF, Teo SH, Jung KJ, Kraft P. Polygenic risk scores for prediction of breast cancer in Korean women. Int J Epidemiol. 2022 Nov 7:dyac206. doi: 10.1093/ije/dyac206. Epub ahead of print. PMID: 36343017.

41. Hsieh CC, Trichopoulos D, Katsouyanni K, Yuasa S. Age at menarche, age at menopause, height and obesity as risk factors for breast cancer: associations and interactions in an international case-control study. International journal of cancer. 1990;46(5):796-800.

42. Liu R, Kitamura Y, Kitamura T, Sobue T, Sado J, Sugawara Y, et al. Reproductive and lifestyle factors related to breast cancer among Japanese women: an observational cohort study. Medicine. 2019;98(51).

43. Park SY, Kolonel LN, Lim U, White KK, Henderson BE, Wilkens LR. Alcohol consumption and breast cancer risk among women from five ethnic groups with light to moderate intakes: the Multiethnic Cohort Study. International journal of cancer. 2014;134(6):1504-10.

44. Gao Y, Huang Y-B, Liu X-O, Chen C, Dai H-J, Song F-J, et

al. Tea consumption, alcohol drinking and physical activity associations with breast cancer risk among Chinese females: a systematic review and meta-analysis. Asian Pacific Journal of Cancer Prevention. 2013;14(12):7543-50.

45. Irwin ML, Varma K, Alvarez-Reeves M, Cadmus L, Wiley A, Chung GG, et al. Randomized controlled trial of aerobic exercise on insulin and insulin-like growth factors in breast cancer survivors: the Yale Exercise and Survivorship study. Cancer Epidemiology Biomarkers & Prevention. 2009;18(1):306-13.

46. Vinogradova Y, Coupland C, Hippisley-Cox J. Use of hormone replacement therapy and risk of breast cancer: nested case-control studies using the QResearch and CPRD databases. bmj. 2020;371.

47. Chen WY. Postmenopausal hormone therapy and breast cancer risk: current status and unanswered questions. Endocrinology and Metabolism Clinics. 2011;40(3):509-18.

48. Evans DG, Graham J, O' Connell S, Arnold S, Fitzsimmons D. Familial breast cancer: summary of updated NICE guidance. Bmj. 2013;346.

49. Monticciolo DL, Newell MS, Moy L, Niell B, Monsees B, Sickles EA. Breast cancer screening in women at higher-than-

average risk: recommendations from the ACR. Journal of the American College of Radiology. 2018;15(3):408-14.

50. Lee EH, Park BY, Kim NS, et al. The Korean guideline for breast cancer screening. Journal of the Korean Medical Association. 2015;58(5):408-19.

51. Arthur RS, Wang T, Xue X, Kamensky V, Rohan TE. Genetic factors, adherence to healthy lifestyle behavior, and risk of invasive breast cancer among women in the UK Biobank. JNCI: Journal of the National Cancer Institute. 2020;112(9):893-901.

52. Roberts MC. Implementation challenges for risk-stratified screening in the era of precision medicine. JAMA oncology. 2018;4(11):1484-5.

53. Cooperation OfE, Development. OECD Health Statistics 2021. 2021.

54. Brentnall AR, Harkness EF, Astley SM, Donnelly LS, Stavrinos P, Sampson S, et al. Mammographic density adds accuracy to both the Tyrer-Cuzick and Gail breast cancer risk models in a prospective UK screening cohort. Breast Cancer Research. 2015;17(1):1-10.

55. Kim EY, Chang Y, Ahn J, Yun JS, Park YL, Park CH, et al. Mammographic breast density, its changes, and breast cancer risk

in premenopausal and postmenopausal women. Cancer. 2020;126(21):468

국문초록

연구배경: 서구화된 생활습관의 보편화로 최근 한국에서도 유방암 발생률이 급격하게 증가하고 있어, 한국인의 특성에 맞는 한국형 유방암 예측 모델의 수립이 절실히 필요한 실정이다. 유방암은 다인성(multi-factorial) 질환으로, 유전적 요인 (genetic risk factor)과 비유전적(non-genetic)요인의 영향을 받기 때문에 두 요소를 모두 포함하는 예측 모델이 유리한데, 한국에는 유전적 요인을 이용한 유방암 예측 모델이 아직까지 부재하다. 최근 유방암의 유전형질(genetic susceptibility)을 다수의 단일염기다형성(single nucleotide polymorphism, SNP)의 합(joint effect)인 다유전자성 위험도(polygenic risk score)를 통해 설명하려는 연구가 증가하고 있어 본 연구에서는 다유전자성 위험도와 비유전적 위험요인을 결합한 한국형 유방암 예측 모델을 수립하려고 한다.

방법: 출간된 (published) 아시아인과 유럽인의 다유전자성 위험도 모델 일곱 개를 이용하여, 결합된 형태의 다유전자성 위험도 모델을 형성하여, 20,434 명의 한국인에서 유방암 위험도 예측력을 검증하고 최적의 한국형 유방암 예측 모델을 수립하였다. 예측력은 곡선아래면적(AUC)과 표준편차당 승산비(OR per SD)로 측정하였다. 이후에 다유전자성 예측 모델에 한국인의 비유전적 위험요인을 결합하여 통합 예측 모델을 개발하여 위험도 예측력 향상을 검증하였다. 위험도 예측 모델은 유방암의 위

험인자인 폐경을 고려해 50세 전후로 구분하여 각각 수립하였다. 완성된 최종 통합 모델로 한국인의 유방암 평생 절대 유방암 위험도(absolute risk)를 예측하였다. 절대 유방암 위험도를 일곱 개 그룹(0-5%, 5-15%, 15-35%, 35-65%, 65-85%, 85-95%, 95-100%)으로 나누어 유방암 위험군을 분류하고, 위험군 및 연령 대에 따른 유방암 발생의 특성을 살펴보았다. 또한 위험도와 비유전적 위험요인 사이의 연관성을 분석하여, 비유전적 위험 요인(수정 가능한 요인)의 영향력이 가장 큰 위험군을 살펴보았다.

결과: 13개의 다유전자성 모델 중 PRS_{38_ASN}+PRS_{190_EB} 이 가장 높은 AUC (0.521)와 OR per SD 1.45 (95%CI: 1.31-1.61)을 보였다. 평균수준의 위험도 그룹과 (35-65%) 비교하였을 때, 최상위 5% 위험도 그룹의 유방암 위험도는 2.5배 높았다. 비유전적 위험요소를 다유전자성 모델에 결합시켰을 때 50세 이상의 여성에서 AUC가 상승하였다. PRS_{38_ASN}+PRS_{190_EB}+NGRF(비유전 위험요소) 모델을 이용해 예측한 한국인의 평균 절대 유방암 위험도는 5.06%였다. 80세의 평생 절대 유방암 위험도는 최상위 5%그룹에서 9.93%였던 반면, 최하위 5% 그룹에서는 2.22% 였다. 유방암의 고위험군 일수록 비유전적 위험요소의 결합에 따라 분산이 커져서, 생활 습관 교정에 민감한 반응을 보일 수 있을 가능성이 있음을 확인하였다.

결론: 아시아인과 유럽인의 다유전자성 위험도를 결합시킨 모델이 한국 여성의 유방암을 잘 예측하였다. 또한, 다유전자성 위험도에 한국인의

비 유전적 위험요인을 결합시킨 통합 유방암 예측 모델 수립하여 한국 여성의 유방암의 절대 위험도를 측량하였다. 이 연구의 결과를 한국여성에게 유방암 위험도에 따른 개별화된 최적의 검진 및 예방 전략을 수립하는 데 필요한 기초 자료로 활용할 수 있을 것이다.

중요성: 한국 여성의 유방암을 예측하는데 있어서 유전학적 감수성과 비 유전적 위험요소의 관계에 대한 통찰을 준다.

……

주요어: 유방암; 예측 모델; 다유전자성 위험도; 비유전적 위험요인; 절대 위험도; 한국인 여성.

학 번: 2018-31469

Acknowledgements

This study was conducted with bioresources from National Biobank of Korea, the Korea Disease Control and Prevention Agency, Republic of Korea (KBN-2021-504). This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HR14C0003).

Author's Contributions

Jihye Choi: Formal Analysis, writing – original draft, writing – review & editing. Tae-woong Ha: Data curation, methodology, formal Analysis, writing – original draft, writing – review & editing. Hye-Mi Choi: Data curation, methodology, writing – review & editing. Han-Byeol Lee: Conceptualization, methodology, writing – review & editing. Hee-Chul Shin: Conceptualization, methodology, writing – review & editing. Woosung Chung: Conceptualization, methodology, supervision, writing – review & editing. Wonshik Han: Conceptualization, methodology, supervision, writing – review & editing.