Supplementary Online Content

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*. Published online May 26, 2016. doi:10.1001/jamaoncol.2016.1025.

This supplementary material has been provided by the authors to give readers additional information about their work.

Supplementary Online Content

Maas P, Barrdahl M, Joshi A, et al. Breast cancer risk from modifiable and non-modifiable risk factors among Caucasian women in the United States. *JAMA Oncol*. Published online.

TABLE OF CONTENTS

1.0 Overview

We first present a general overview of the methods and then describe details of each step in separate sections. The cohorts used in this analysis were: the European Prospective Investigation into Cancer and Nutrition (EPIC)[1], the Women's Health Initiative (WHI)[2], the Melbourne Collaborative Cohort Study (MCCS)[3], the Nurses' Health Study (NHS)[4], the Women's Health Study (WHS)[5], the American Cancer Society Cancer Prevention Study-II Nutrition Cohort (CPS-II NC)[6], the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO)[7], and the Multiethnic Cohort (MEC)[8].

Following the standard definition[9], the absolute risk of developing breast cancer for a woman of age *a* within *s* years (that is, over time interval [a, a+s]) was defined as

$$R_{a,a+s} = \int_{a}^{a+s} \lambda_0(t)\exp(Z\beta)\exp\left(-\int_{a}^{t}\left(\lambda_0(u)\exp(Z\beta)+m(u)\right)du\right)dt \qquad (1)$$

Formula (1) holds under the assumptions that the risk factors (Z) act in a multiplicative fashion on the baseline hazard function ($\lambda_0(t)$), which is to say that the age-specific hazard ratio parameters (β) remain constant over time. For some risk factors, such as BMI, which is known to have different associations with breast cancer risk among pre-and post-menopausal women, we allowed the hazard-ratio parameters to be different before and after age 50 by including suitable age by risk factor interaction terms. Formula (1) includes age-specific competing hazards of mortality ($m(t)$) to account for the fact that the observable risk of breast cancer is reduced in the presence of competing risks of mortality from other causes.

In brief, the model included "non-modifiable" risk factors other than the PRS (i.e. family history, age at first birth, parity, age at menarche, height, menopausal status, age at menopause), along with "modifiable" risk factors (i.e. BMI, MHT use, level of alcohol consumption, and smoking status). We considered age at first birth and parity, two reproductive risk factors, as non-modifiable as women are unlikely to choose to modify these factors based on breast cancer risk. For all studies information collected at baseline was used to define the risk factors. MHT was defined based on baseline information among postmenopausal women and was categorized as never user or former/current user of estrogen-only type therapy, combined estrogen plus progesterone therapy, or therapy of unknown type. As several of the risk factors had a substantial amount of missing data in some studies (Supplementary Table 1), we built study-specific models for multiple imputation. In the final model, all continuous risk factors other than the PRS were modeled categorically (most in deciles) to allow for the non-linear associations evident from exploratory analysis. The model also included known interactions[10] between menopausal status, BMI and MHT and was adjusted for study and age in categories (<50, 50-<55, 55-<60, 60-<70, 70+ years). Specifically, the interaction allowed the relationship between BMI and breast cancer risk to vary between premenopausal women, postmenopausal ever MHT users, and postmenopausal never MHT users[10].

To incorporate genetic information in the model, we included a total of 92 known susceptibility SNPs (Supplementary Table 2), of which 24 were genotyped in the Breast and Prostate Cancer Cohort Consortium (BPC3) subjects. We first modeled the risk associated with just these 24 SNPs to evaluate possible interactions with other risk factors in the BPC3 data. We derived a polygenic risk score for the 24 SNPs (PRS-24) based on a linear logistic model adjusted for age, cohort, and family history. We conducted an in-depth analysis of possible multiplicative interactions between the PRS-24 and individual risk factors, but did not detect significant evidence of interaction on the logistic scale.

We then modeled the risk of breast cancer associated with all 92 known breast cancer SNPs incorporated through a PRS-92. For the 68 SNPs that were not genotyped in the current study, we simulated the PRS-68 for the BPC3 subjects conditional on their case-control status and family history published estimates[11] of odds-ratios and allele frequencies for the 68 SNPs. The simulation allows building the model based on all 92 SNPs where the odds-ratio for the 68 SNPs that are not genotyped in the current study are informed by external studies. Using the various models for imputation and simulation described above, we created five "complete datasets" that had information on PRS-92, family history and other questionnaire-based risk factors. All estimates provided in this report are obtained by averaging over the results for the five datasets.

Following, we describe the details of several steps.

2.0 Model Building Procedure

In order to select a final multivariate logistic regression model for the association between the PRS-92 and epidemiologic risk factors for breast cancer, we performed a number of analyses, mentioned briefly here and described in detail below. First, we evaluated whether there was heterogeneity in the risk factor associations across the different cohorts. We then evaluated linearity of the relationships between each risk factor and breast cancer risk on the logistic scale. In general, we found non-linearity in the associations and thus chose to take a more nonparametric approach, modeling based on categorical versions of the risk factors. After categorizing the variables, we performed a final modeling step and selected our final model.

2.1 Evaluating Heterogeneity by Study

We evaluated heterogeneity in the associations across different cohorts by fitting logistic regression models for each risk factor adjusted for age and study, creating forest plots of the effect sizes, and testing for heterogeneity in effect size. It is known that the association between BMI and breast cancer risk differs for premenopausal and postmenopausal women and by MHT status. Thus we evaluated heterogeneity in the BMI effect separately in strata defined by menopausal status and ever hormone replacement therapy (MHT) use. We did not find statistically significant heterogeneity in the effects by study, except for the age at first full term birth variable. In that case, all effects were qualitatively consistent in that greater Age at first full term birth (AFFTB) was associated with an increased risk of breast cancer in all cohorts, but with some variability in the estimated effect sizes.

2.2 Evaluating Linearity of Risk Factor Associations

To evaluate whether the effect of each continuous risk factor (height, parity, age at first full-term birth, age at menarche, age at menopause, body mass index, and alcohol consumption) could be modeled in a linear fashion on the logistic scale, we first fit generalized additive models relating case-control status to each continuous variable individually, adjusted for age and cohort[12]. The models allowed us to examine covariate effects using a non-parametric smoothing method to ascertain whether linear modeling was appropriate. When evaluating linearity for the BMI association, we stratified by menopausal status and ever MHT use. We observed non-linear associations for all continuous risk factors, so we chose to build our final model based on categorical versions of the variables.

2.3 Categorizing the Continuous Risk Factors

The main epidemiologic risk factors of interest included some categorical variables (family history, smoking status, and hormone replacement therapy use) and other variables that we initially considered continuous (age, age at menarche, alcohol use, parity, age at first birth, age at menopause, and body mass index, height). After our initial model building exploration, we chose to categorize the continuous variables based on cut points defined by unique values of the deciles. The exceptions to this rule were the categorical parity variable (which we defined in five categories: 0, 1, 2, 3, and 4+ children) and the age variable used for adjustment (which we defined in five categories: <50, 50-<55, 55-<60, 60-<70, 70+). The decile cut points were not all unique, resulting in categorizations with fewer than ten categories. Supplementary Table 4 gives the cut points used to categorize the remaining continuous covariates.

All PRS in the paper are defined as, $PRS = \sum_k \hat{\beta}_k G_k$ where G_k are SNPs taking values 0, 1, 2 and $\hat{\beta}_k$ are the estimated log odds ratio parameters for the association between the SNP and breast cancer, adjusted for family history. PRS-24 includes the 24 SNPs genotyped in BPC3. PRS-92 includes information on an additional 68 SNPs from the Bread Cancer Association Study (BCAC) and GWAS, which have been shown to be associated with breast cancer risk.

3.0 Building Imputation Models

Missingness of the breast cancer risk factors varied by cohort and is presented in Supplementary Table 1. To avoid discarding a large number of subjects who had missing data in at least one of the variables, we built models to impute missing values. We imputed missing values for all risk factors sequentially, starting with the least missing and progressing in order of increasing amount of missing data. We constructed each imputation model conditional

on case-control status, age, cohort, and all complete variables that were significantly associated with the risk factor being imputed. These imputation models also included any significant two-way interactions between the variables included in the model. We compared the associations between case-control status and each covariate adjusted for age and cohort before and after imputation to verify that none of the estimated effects changed by more than 10%. We found that most differed by less than 2% before and after imputation.

The cohorts also had different patterns of missing data for each of the 24 SNPs in the BPC3. Within each cohort, we imputed missing data for each SNP for which there was data, using an imputation model conditional on case-control status, family history, and an interaction between the two. For the SNPs entirely missing in a given cohort, we did not attempt to impute a value for that SNP. Instead, we imputed the missing component of the polygenic risk score, PRS-missing, (rather than each individual SNP) for each cohort. We performed this PRS-missing imputation using the approach described in Supplementary Section 5. Thus, the "empirical PRS-24" was constructed in a cohort-specific manner from a combination of observed SNP information and imputed SNP information, and an imputed PRS-missing component. To add more genetic information to the model, we also included an additional 68 SNPs by generating a PRS-68, described in Supplementary Section 3. We created PRS-92 by simply adding simulated PRS-68 to empirical PRS-24.

Using these imputation methods, we created 5 completed datasets for analysis.

4.0 Examining Interaction between PRS-24 and Epidemiologic Factors

We conducted a number of analyses to examine possible interactions between the empirical PRS-24 (for which we had data to evaluate interaction) and the epidemiologic factors in the BPC3.

First, we fit logistic regression models for each risk factor (modeled continuously to reduce the degrees of freedom and increase the power for detecting interaction) adjusted for age and cohort, within 10 strata defined by the deciles of PRS-24. We created forest plots to look at whether each risk factor association differed across the strata of PRS-24, and performed statistical tests of heterogeneity. We performed this analysis for each of the five datasets completed by imputation and did not find evidence of interaction between PRS-24 and any of the epidemiologic risk factors.

Next, we evaluated interactions between the PRS-24 and each risk factor separately for the lowest extreme of PRS-24, the middle of PRS-24, and the upper extreme of PRS-24 using likelihood ratio tests (LRT). For the "middle" of the PRS-24, we tested whether an interaction between the PRS-24 and the covariate was significant, in a model adjusted for age and cohort. We then fit models with a binary indicator of being in the lower extreme range of PRS-24 (as opposed to the middle) and tested whether an interaction with each covariate was significant; similarly for the upper extreme. We defined the "extremes" of PRS-24 by the 5^{th} and 95^{th} percentiles, leaving the "middle" to include 90% of the PRS-24. To be thorough, we also ran LRTs for interaction in "extremes" defined by the 10^{th} and 90^{th} percentiles of PRS-24, leaving the "middle" to include 80% of the PRS-24. We performed these analyses for each of the five analysis datasets completed by imputation, for both continuous and categorical versions of the risk factors. Examining the results across all these analyses as a whole, we did not find consistent evidence of interaction between PRS-24 and any of the risk factors.

On this basis, we did not include interactions between the PRS-24 and risk factors in our final model. We also performed a goodness-of-fit test for the final model using a novel method that specifically tests for model fit in the extremes of disease risk[13] where small undetected interactions across many variables may cause substantial model misspecification. The goodness-of-fit test was not significant (p-value>0.05) by which we conclude that a model with no interactions between PRS-24 and risk factors fits the BPC3 data adequately.

5.0 Generating Simulated PRS

We simulated PRS-68, i.e. the component of PRS-92 defined by 68 SNPs not genotyped in BPC3, conditional on case-control status and family history of the subjects using the model estimates of the log odds ratios, $\hat{\beta}_{k\ ,}$ and the allele frequencies, f_k, for the SNPs along with an estimate of the log odds ratio for family history, $\hat{\beta}_{fh}$. Briefly, the method requires simulating from:

$$P(\,PRS_i\,|D = d, Family\ H = h) \sim N(\,\mu, \sigma^2_{PRS}),$$

$$\text{where } \mu = d * \sigma^2_{PRS} + \tfrac{1}{2} * h * \sigma^2_{PRS} \quad \text{and} \quad \sigma^2_{PRS} = \sum_k 2\,\hat{\beta}^2_k\, f_k(1 - f_k), \text{ and}$$

where estimates of the log odds ratios, $\hat{\beta}_k$,and the allele frequencies, f_k, for the SNPs were obtained from data available from the Breast Cancer Association Consortium[11,14]. Details of derivation of the distribution of PRS conditional on disease and family history status can be found in Supplementary Methods section of the report by Chatterjee et al[15].
For the 24 SNPs genotyped, we observed that odds-ratios associated with PRS categories and family-history obtained from the above log-normal model tracks well with those obtained empirically within BPC3 (Supplemental Table 3).

6.0 Combining Multiple Imputation Results

We imputed variables with missing values to complete the BPC3 dataset. To account for uncertainty in the values that were imputed, we created five datasets completed by imputation. We then estimated the risk factor associations using each of the five datasets, yielding five log odds ratio estimates: $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$, $\hat{\beta}_4$, and $\hat{\beta}_5$. We combined these results using the standard method to produce a single overall estimate and accompanying confidence interval, which appropriately accounts for the variability in multiple imputations. Specifically, we applied the following formulas for $m = 5$ imputations, and U_k equal to the standard error associated with each $\hat{\beta}_k$estimate:

$$\text{Overall log odds ratio:} \quad \bar{\beta} = \frac{1}{m} \sum_{k=1}^{m} \hat{\beta}_k$$

$$\text{Variance of } \bar{\beta}\text{:} \quad T = \bar{U} + \left(1 + \frac{1}{m}\right) Q$$

This variance formula[16] is based on the overall standard error, $\bar{U} = \frac{1}{m} \sum_{k=1}^{m} U_k$, and the between imputation variance, $Q = \frac{1}{m-1}\sum_{k=1}^{m}\left(\hat{\beta}_k - \bar{\beta}\right)^2$. The resulting overall odds ratios and confidence intervals from applying these formulas to the BPC3 estimates are presented in Figure 1 of the main text.

7.0 Projecting Risk for the US Population

We aimed to use the BPC3 model to project the distribution of absolute risk for the US population. To make such projections, we needed to generate a risk factor distribution that would be representative of the US general population. As a starting point, we use data available from the 2010 National Health Interview Survey (NHIS) on 5879 women with age at interview between 30 and 70. Data on all the required risk factors were available on the NHIS women except for PRS, alcohol consumption, age at menopause and MHT use. Given the information on the known risk factors, we simulated information on the unknown factors using information and models gleaned from a number of other sources. First, data on PRS were generated conditional on family-history using the same model as that we described earlier for simulating PRS-68 within the BPC3 studies. Second, to simulate alcohol and age at menopause, we used controls available from Women's Health Initiative study to build imputation models conditional on all other variables with significant associations with the missing risk factors. Finally, we developed a model to generate a time-dependent profile for use of MHT, taking particular care that it reflect current rate of MHT use in the US population, as follows.

We used data available from the National Health and Nutrition Examination Survey (NHANES) across years 2008, 2010, and 2012 to estimate proportion of women who can be considered ever users of MHT and the type of hormones they use (E-type vs C-type). For simulated profiles who are assigned to be ever MHT users, we then generate a dynamic time-dependent profile of post-menopausal MHT by simulating age-at-initiation and duration using observed distribution of these factors for post-menopausal Caucasian women who were ever MHT users and participated in NHANES 2010, the only year for which such data were available. Following the scheme as above, we generate 5 simulated profiles for each of the 5879 women enrolled in NHIS 2010 leading to a total of 29368 risk profiles that can be considered nationally representative. For each profile, we use the BPC3 model to assign risk over different time intervals taking into account time-dependency of some of the risk factors. In particular, a woman

who was simulated to be an MHT user is categorized as a "current" user only during her simulated interval of use, which is shifted forward by a lag-factor of 5.6 years, the average follow-up time in the BPC3 studies. An MHT user is always categorized as "former" user after the interval of "current" use.

8.0 Distribution of modifiable and non-modifiable risk

We decomposed the overall distribution of breast cancer risk by modifiable and non-modifiable components (Figure 4). We first stratified the population based on deciles of risk scores defined by the non-modifiable risk factors. For each category of non-modifiable risk (x-axis of Figure 4), we then evaluate the absolute risks of the women based on an assigned "average" non-modifiable risk score and a modifiable risk score that is defined by their observed modifiable factors. Boxplots are used to present the variation of risk due to the modifiable factors within each category of non-modifiable risk.

To further quantify how much of the risk of breast cancer can explained by modifiable factors, we calculate the proportion of disease preventable (PDP) by elimination of modifiable risk factors, overall and within strata of the population defined by non-modifiable risk factors.

We define overall PDP as that which, according to our model, would not have arisen if all women in the population had the lowest risk levels for the modifiable breast cancer factors, M_0, instead of their observed levels, M.

Thus, the overall PDP is defined as

$$PDP = \frac{[\mathrm{P(D|M)} - \mathrm{P(D|}\, M_0)]}{\mathrm{P(D)}},$$

which is same as the definition of the population attributable risk[17,18] associated with the modifiable risk factors. To assess how the proportion of preventable cancers are distributed over different strata defined by the non-modifiable factors, we then further decompose PDP as

$$PDP = \frac{\sum_{k=1}^{K} PDP_k}{P(D)} = \frac{\sum_{k=1}^{K}[P(D|M, NM = k) - P(D|M_0, NM = k)]P(NM = k)}{P(D)},$$

where each component of the sum represents the proportion of disease expected to be prevented if the modifiable risk factors were eliminated only within a targeted stratum of the population that is defined by the non-modifiable risk factors. The ratio of PDP for a targeted stratum (tPDP) to the total PDP, defined as

$$ePDP_k = \frac{[P(D|M, NM = k) - P(D|M_0, NM = k)]P(NM = k)}{\sum_{k=1}^{K}[P(D|M, NM = k) - P(D|M_0, NM = k)]P(NM = k)},$$

can be used to assess potential effectiveness of a targeted intervention strategy for risk factor modification. We considered "modifiable factors" (and lowest risk levels) to include: alcohol consumption (0<-0.4 drinks/week), use of hormone replacement therapy (never), body mass index (<21.5), and smoking (never). We examined the impact of modifying each risk factor individually as well as all simultaneously, as measured by the PTC and PPC metrics.

9.0 Tables and Figures

eTable 1. Percentage of subjects with missing information on individual risk factors

Cohort		Sample Size	Family History	Age at Men-arche	Parity	Age at First Birth	Age at Menop	Height	Body Mass Index	Menop Status	MHT Use: Ever	MHT Use: Ever E	MHT Use: Ever C	MHT Use: Current	Alcohol Use	Smoking Status
All	N	37033	35·2	1·9	2·4	6·6	12·6	0·3	1·0	5·5	19·7	38·9	37·1	39·1	2·6	0
	Cases	17171	35·7	2·2	2·4	4·2	12·8	0·3	0·9	5·6	19·5	38·7	36·4	37·0	2·3	0
	Controls	19862	34·7	1·7	2·3	8·6	12·4	0·3	1·1	5·4	19·8	39·1	37·8	40·9	2·9	0
CPS-II NC	N	5923	1·7	1·2	1·8	2·1	1·8	0·6	1·0	1·4	74·7	75·5	75·5	100	6·4	0
	Cases	2558	2·2	1·4	1·9	2·3	2·0	0·6	1·1	1·5	76·1	76·8	76·8	100	5·5	0
	Controls	3215	1·3	1·1	1·7	2·0	1·6	0·6	1·0	1·2	73·6	74·4	74·4	100	7·2	0
EPIC	N	9322	67·0	3·8	7·4	14·7	31·7	0	0	16·4	22·9	50·3	43·2	28·2	0	0
	Cases	4156	63·2	4·5	7·4	4·7	33·8	0	0	17·3	24·5	55·4	46·0	24·7	0	0
	Controls	5166	70·1	3·3	7·4	22·6	29·9	0	0	15·7	21·5	46·2	41·0	31·1	0·1	0
MCCS	N	1695	49·3	0·2	0	0·1	15·2	0·1	0·1	2·2	0·8	100	100	100	0	0
	Cases	930	52·4	0·2	0	0·1	13·4	0	0	4·0	1·2	100	100	100	0	0
	Controls	765	45·6	0·3	0	0	17·3	0·1	0·1	0	0·4	100	100	100	0	0
MEC	N	1091	8·1	0·5	0·7	1·2	2·0	0·1	0·1	2·1	2·4	3·9	3·9	100	3·2	0
	Cases	520	10·8	0·8	1·0	1·0	2·7	0·2	0·2	2·7	3·3	4·4	4·4	100	4·2	0
	Controls	571	5·6	0·2	0·5	1·4	1·4	0	0	1·6	1·6	3·5	3·5	100	2·3	0
NHS	N	4932	0·5	0·5	0	0	9·2	0·1	4·1	2·5	15·0	23·9	23·9	38·9	7·0	0
	Cases	1781	0·7	0·8	0	0	10·1	0·1	4·3	2·2	18·4	17·9	17·9	36·3	7·9	0
	Controls	3151	0·4	0·3	0	0	8·7	0·1	4·0	2·6	13·1	27·3	27·3	40·4	6·5	0
PLCO	N	1767	0·7	0·2	0·1	0·3	1·0	0·2	0·5	1·0	0·6	100	100	1·7	8·8	0
	Cases	787	1·0	0·0	0·0	0·3	1·1	0·1	0·4	1·1	0·4	100	100	1·0	7·4	0
	Controls	980	0·5	0·3	0·2	0·3	0·8	0·3	0·6	0·8	0·8	100	100	2·2	10·0	0
WHI	N	11119	51·1	2·3	0·6	8·3	5·2	0·5	0·8	0	0	0	0	0	0·4	0
	Cases	5920	49·5	2·4	0·8	7·9	5·1	0·5	0·7	0	0	0	0	0	0·5	0
	Controls	6849	52·8	2·2	0·4	8·8	5·3	0·6	0·8	0	0	0	0	0	0·4	0
WHS	N	1334	2·5	0	0	0·1	20·2	0·9	1·3	16·7	4·3	51·3	51·3	100	0	0
	Cases	669	3·3	0	0	0·1	18·2	0·6	0·7	14·8	4·5	48·0	48·0	100	0	0
	Controls	665	1·8	0	0	0	22·1	1·2	2·0	18·6	4·2	54·6	54·6	100	0	0

eTable 2. List of SNPs in PRS-24 and PRS-68

24 SNPs genotyped in BPC3	53 SNPs with relative risks and allele frequencies from BCAC		15 SNPs with relative risks and allele frequencies from COGS[1]
rs11249433	rs75915166	rs11242675	rs12405132
rs1045485	rs554219	rs204247	rs12048493
rs13387042	rs2736108	rs720475	rs72755295
rs4973768	rs2588809	rs9693444	rs6796502
rs10069690	rs10759243	rs6472903	rs13162653
rs10941679	rs11199914	rs2943559	rs2012709
rs889312	rs7072776	rs11780156	rs7707921
rs17530068	rs11814448	rs7904519	rs9257408
rs2046210	rs16857609	rs11820646	rs4593472
rs1562430	rs11552449	rs12422552	rs13365225
rs1011970	rs12662670	rs17356907	rs13267382
rs865686	rs10771399	rs11571833	rs11627032
rs2380205	rs1292011	rs2236007	chr17:29230520:D
rs10995190	rs2363956	rs941764	rs745570
rs1250003	rs2823093	rs17817449	rs6507583
rs2981582	rs17879961	rs13329835	
rs909116	rs616488	rs527616	
rs614367	rs4849887	rs1436904	
rs10483813	rs2016394	rs4808801	
rs3803662	rs1550623	rs3760982	
rs6504950	rs6762644	rs132390	
rs8170	rs12493607	rs6001930	
rs2284378	rs9790517	rs4245739	
rs999737_as	rs6828523	rs6678914	
	rs10472076	rs12710696	
	rs1353747	rs11075995	
	rs1432679		

SNP selection and genotyping described in Supplemental Section 1.
[1]Collabortive Oncological Gene-Environment Study

eTable 3. Comparison of Empirically Estimated and Theoretically derived[1] Odds Ratios for deciles of PRS-24

	Estimated Odds Ratios	
	Empirical	**Theoretical Model**
PGRS Decile 1	1	1
PGRS Decile 2	1·19	1·21
PGRS Decile 3	1·32	1·42
PGRS Decile 4	1·43	1·48
PGRS Decile 5	1·54	1·59
PGRS Decile 6	1·65	1·76
PGRS Decile 7	1·80	1·92
PGRS Decile 8	2·07	2·04
PGRS Decile 9	2·26	2·32
PGRS Decile 10	2·79	2·88
Family History	1·40	1·37

[1]Log-normal model for polygenic risk derived using estimates of odds-ratios and allele frequencies for individual SNPs

Supplementary Table 4. Cut Points Used to Categorize Continuous Variables

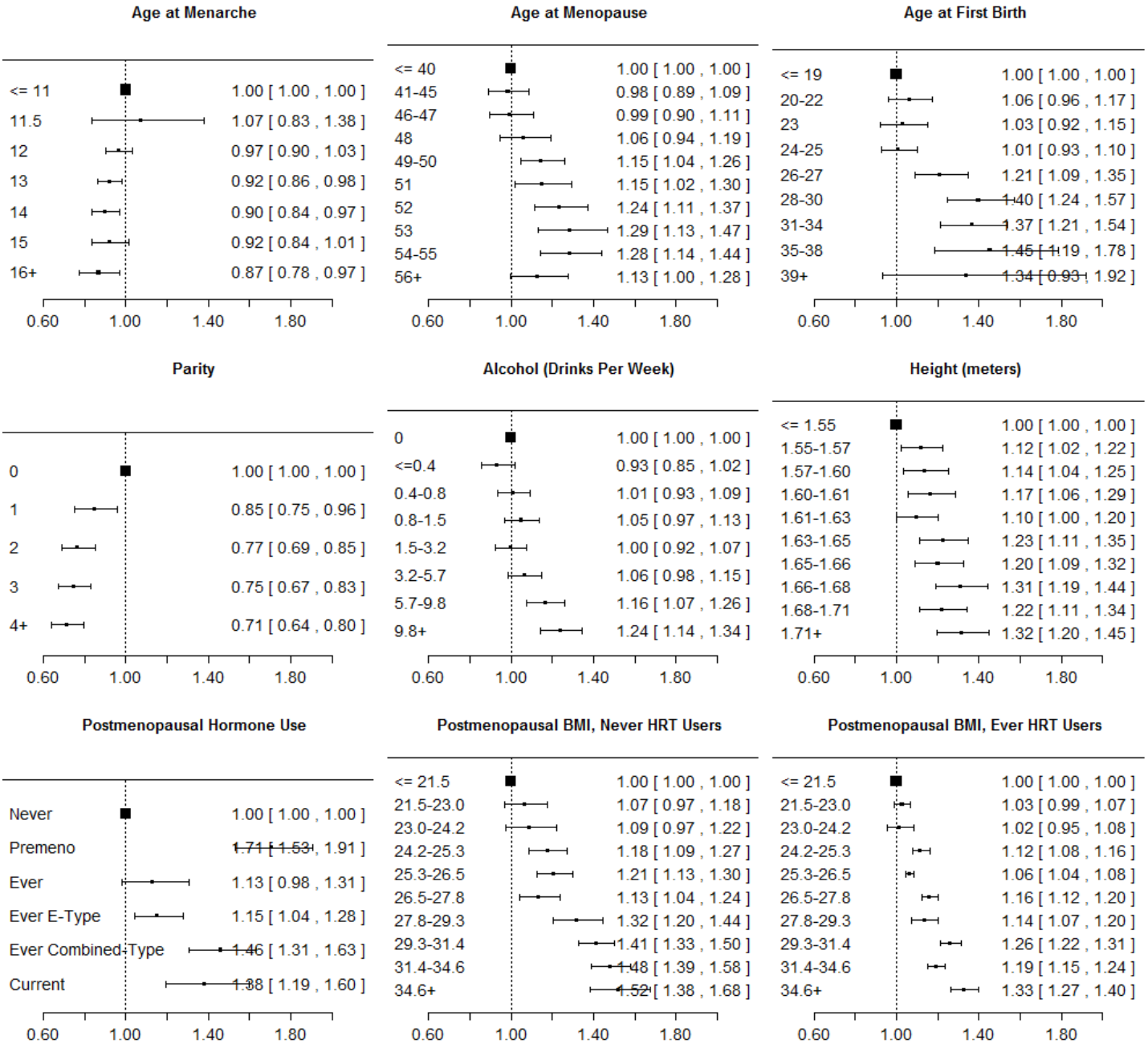
Variable	Scale	Cut Points*								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Age at menarche	**years**	11	11·5	12	13	14	15	--	--	--
Age at first birth[+]	**years**	19	22	23	25	27	30	34	38	--
Age at menop[#]	**years**	40	45	47	48	50	51	52	53	55
Alcohol use	**drinks/week**	0·0	0·4	0·8	1·5	3·2	5·7	9·8	--	--
Body mass index	**kg/m^2**	21·5	23·0	24·2	25·3	26·5	27·8	29·3	31·4	34·6
Height	**meters**	1·55	1·57	1·60	1·61	1·63	1·65	1·66	1·68	1·71

* Each category includes right end point, does not include left end point; mathematically (x,y].

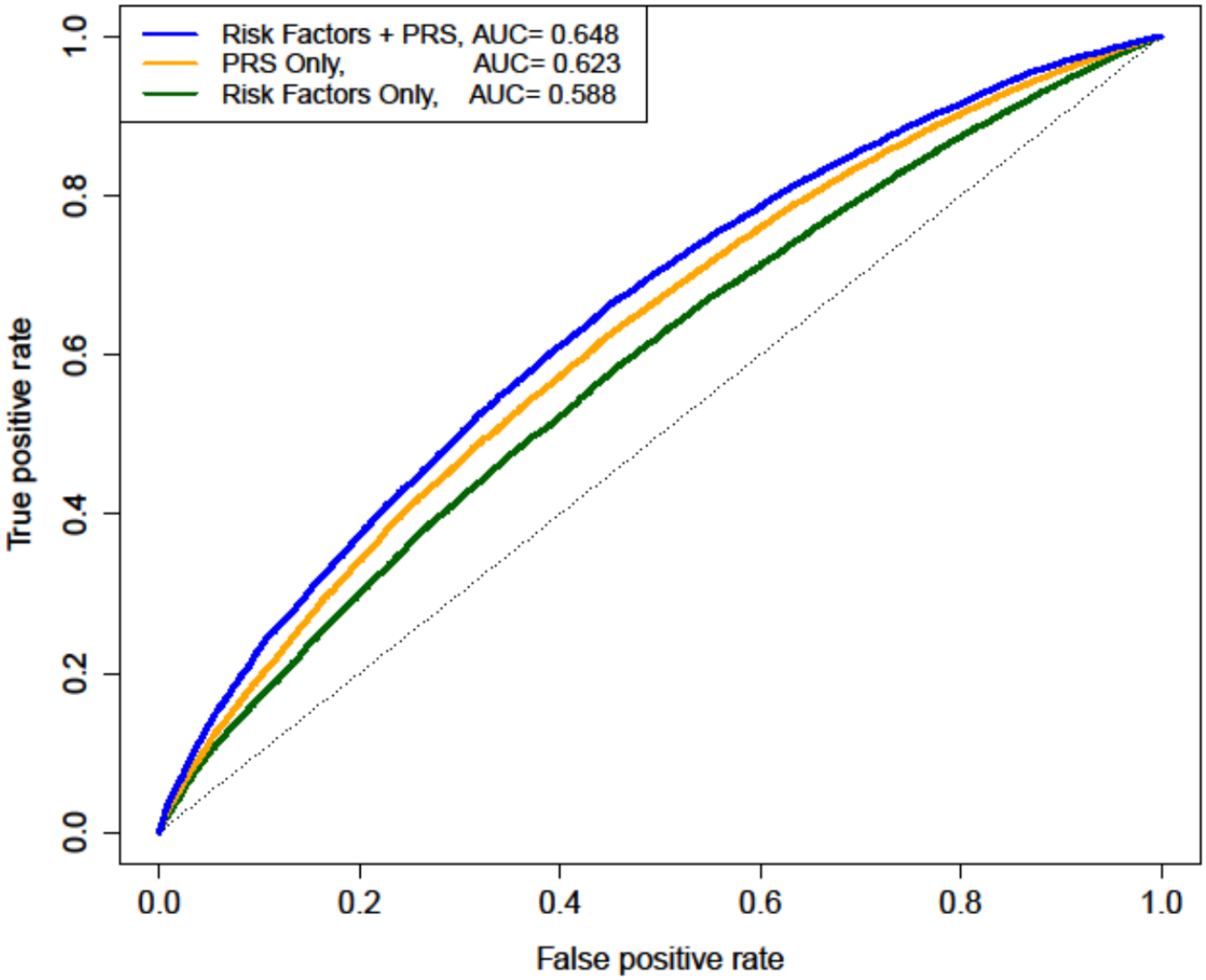
[+] Individuals with no children were assigned to category 1.

[#] Premenopausal individuals were assigned to category 1.

eFigure 1. Odds Ratios for Categorical Covariates in Full Model



eFigure 2. ROC Curves for Models



10.0 References

1. Riboli E, Hunt KJ, Slimani N, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. Public Health Nutr 2002;5:1113-24.
2. Anderson GC, S.; Freedman, L. S.; Furberg, C.; Henderson, M.; Johnson, S. R.; Kuller, L.; Manson, J.; Oberman, A.; Prentice, R. L.; Rossouw, J. E. Design of the Women's Health Initiative Clinical Trial and Observational Study. Controlled Clinical Trials 1998;19:61-109.
3. Giles GG ED. The Melbourne Collaborative Cohort Study. IARC Sci Publ 2002:69-70.
4. Colditz GA, Hankinson SE. The Nurses' Health Study: lifestyle and health among women. Nat Rev Cancer 2005;5:388-96.
5. Rexrode K, Lee I, Cook N, Hennekens CH, Burning JE. Baseline characteristics of participants in the Women's Health Study. Womens Health Gend Based Med 2000;9:19-27.
6. Calle EE, Rodriguez C, Jacobs EJ, et al. The American Cancer Society Cancer Prevention Study II Nutrition Cohort: rationale, study design, and baseline characteristics. Cancer 2002;94:2490-501.
7. Hayes RB, Reding D, Kopp W, et al. Etiologic and early marker studies in the prostate, lung, colorectal and ovarian (PLCO) cancer screening trial. Controlled Clinical Trials 2000;21:349-55.
8. Kolonel LN, Henderson BE, Hankin JH, et al. A Multiethnic Cohort in Hawaii and Los Angeles: Baseline Characteristics. Am J Epidemiol 2000;151:346-57.
9. Gail MH, Brinton LA, Byar DP, 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.
10. Morimoto LM, White E, Chen Z, et al. Obesity, body size, and risk of postmenopausal breast cancer: the Women's Health Initiative (United States). Cancer causes & control : CCC 2002;13:741-51.
11. Michailidou K, Hall P, Gonzalez-Neira A, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nature genetics 2013;45:353-61, 61e1-2.
12. Hastie T, Tibshirani R. Generalized Additive Models. 1986:297-310.
13. Song M, Kraft P, Joshi AD, Barrdahl M, Chatterjee N. Testing calibration of risk models at extremes of disease risk. Biostatistics 2014.
14. Michailidou K, Beesley J, Lindstrom S, et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nature genetics 2015;47:373-80.
15. Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock SJ, Park JH. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nature genetics 2013;45:400-5, 5e1-3.
16. Rubin DB. Multiple Imputation for Nonresponse in Surveys: John Wiley & Sons, Inc.; 2008:154-201.
17. Cole P, MacMAHON B. Attributable risk percent in case-control studies. British journal of preventive & social medicine 1971;25:242-4.
18. Bruzzi P, Green SB, Byar DP, Brinton LA, Schairer C. Estimating the population attributable risk for multiple risk factors using case-control data. Am J Epidemiol 1985;122:904-14.