Polygenic Score to Understand Cancer Etiology and Predict Cancer Risks

POLYGENIC SCORE TO UNDERSTAND CANCER ETIOLOGY AND PREDICT CANCER RISKS

CHI GAO

A Dissertation Submitted to the Faculty of

The Harvard T.H. Chan School of Public Health

in Partial Fulfillment of the Requirements

for the Degree of *Doctor of Science*

in the *Department of Epidemiology*

Harvard University

Boston, Massachusetts.

May 2020

Dissertation Advisor: Dr. Peter Kraft Chi Gao

Polygenic Score to Understand Cancer Etiology and Predict Cancer Risks

Abstract

Genetics have been an important risk factor for cancer. The information we learned from genome-wide association studies (GWAS) provide researchers with tools and new approach to better understand cancer epidemiology. In this dissertation, I present three projects using GWAS discoveries to understand cancer etiology and infer cancer risks.

Chapter 1 uses GWAS information as an instrument variable to estimate the causal relationship between adiposity measures at different life stages (at birth, during childhood, at adulthood) and risk of breast, ovarian, prostate, colorectal and lung cancers via Mendelian Randomization analysis. We found that the genetic predicted adult BMI was inversely associated with breast cancer risk but positively associated with ovarian, lung and colorectal cancer risk.

Chapter 2 evaluates the performance of a synthetic breast cancer risk prediction model utilizing both classical risk factors of breast cancer and common genetic variants in form of polygenic risk score (PRS). We validated the model using Nurses Health Study and Nurses Health Study II. We found that adding PRS greatly improved the performance of risk prediction models and of all three models validated, the model with both classic risk factor and PRS performed the best.

Chapter 3 investigates the joint effect of PRS and pathogenic mutation in nine breast cancer predisposition genes using population based cohort studies in CAnceR RIsk Estimates Related to Susceptibility (CARRIERS) consortium. We also estimated 5-year and lifetime

absolute risk using the final model built from penalized regression. We found that PRS is associated with breast cancer in carriers of pathogenic variant as well as in non-carriers but there was no significant difference between these effect (odds ratio associated with one standard deviation change in PRS). More importantly, we found that PRS can be particularly important for managing risk of carriers of pathogenic variants in moderate penetrance cancer predisposition genes such as *ATM* and *CHEK2*.

Together, the projects presented in this dissertation demonstrated three approaches to utilize genetic information to understand cancer in the post-GWAS era. We hope that these findings could shed light to the underlying genetic architecture of cancer and could contribute to future studies of building breast cancer risk prediction models and generating effective screening guidelines.

Table of Contents

Listing of Figures

Listing of Tables

Acknowledgements

I would like to give my foremost gratitude to my advisor, Dr. Peter Kraft, for his mentorship, guidance, and support. I could not wish for a better advisor than Pete and it has been such a memorable and enjoyable life experience working with him. I became a better researcher because of Pete and I have learned so much from him. He taught me to be academically rigorous and taught me to focus on key insights in solving complex research problems. He mentored me on both the micros and macros of genetic epidemiology, the technicalities of running statistical models and the art of storytelling in science. I would like to thank him for always being so supportive, for providing me the space and freedom to grow both in academia and in life, and for always being so kind and generous to me. Pete is truly a role model in my life and I wish that I could become a person like him in the years to come: insightful, decent, and reliable.

I would like to express my greatest appreciation to my research committee members: Dr. Liming Liang, Dr. Sara Lindstroem, and Dr. Rulla Tamimi. Liming has been a tremendous mentor in statistical methods. He provided me new thoughts on research methods and always offered great insights during my committee meetings. I also benefited a lot from the journal club he organized. To Sara, whom I worked closely during my first two years at Harvard Chan. Many thanks for introduce me to the scientific community in the genetic epidemiology field. If it weren’t for such a positive and pleasant experience working with Sara, I probably would not have continued on as a doctoral student at Harvard Chan. To Rulla, who is my go-to person whenever I have questions about the Nurses Health Studies. Rulla’s advice and kindness offered me great support, especially during the last few years of my graduate study.

I am forever grateful to have worked with many talented researchers and collaborators all over the country during my time at the Harvard Chan. I would like to thank Dr. Chirag J. Patel, who provided great insight and advice for my first project of Mendelian Randomization Study. Dr. Montserrat Garcia-Closas, Dr. Nilanjan Chatterjee, Dr. Parichoy Choudhury and Ms. Amber Wilcox have been such a wonderful team to work with for my second project. I have learned a lot from them regarding risk prediction models and validation analyses. To Dr. Eric Polley, Dr. Fergus Couch, Dr. Susan Domchek, Dr. Kate Nathanson, Dr. David Goldgar on the CARRIERS consortium. Many thanks to them for their support and mentorship on my third project. They have taught me how to be a good collaborator, how to communicate effectively across disciplines, and how to translate research findings with clinical importance. To Dr. Lori Chibnik, whom I worked closely with for Epi 215 class. Lori's class on modeling and imputation method for missing data is one of my favorites throughout my graduate school and the material I learned in that class were radically applied in my second and third dissertation projects.

I would also like to thank my college advisor Dr. Nicholas Horton who introduced me to the world of statistics and my mentor Dr. Sekar Kathiresan who brought me into genetic epidemiology and bioinformatics that had significant influenced my research trajectory.

Next, I would like to thank my colleagues and my friends. It has been an honor and great pleasure to have known amazing colleagues at the Program in Genetic Epidemiology and Statistical Geneics and the Epidemiology department. Many thanks to Dr. Xia Jiang, Dr. Jihye Kim, and Dr. Howard Liu for your help and support to keep my sanity intact throughout the past few years.

My sincere gratitude to all my friends at Harvard for your friendship, loyalty, and continued support throughout these years. You have made my time at Harvard so colorful and memorable. There are so many to name here, but I would like to specifically pay thanks to Dr. Siyuan Ma, who provided me unwavering mental support and helped me greatly in programming. To Dr. Xiaoyu Li, Dr. Qian Di, Dr. Jeremiah Zhe Liu, and Dr. Zilin Li for their friendship and kindness. To my HCSSA friends, Dr. Liuchuan Tong, Dr. Ximei Sun, Dr. Xizhu Zhao, Dr. Guannan Qu, Mr. Martin Jin, Mr. Jialiang Wang, Ms Xiao Zhang, Mr. Yichu Jin, Ms. Lucy Liu, Mr. Jiawei Zhou who have been brothers and sisters to me and whom I have shared many laughs, drinks and tears with. To Dr. Thida Ong and Ms. Monica Zhong whose friendship and unparalleled support meant a world to me.

I would genuinely like to thank Mr. Xian Gong, for making me laugh, for giving me a shoulder to cry on, for telling me that I will make it when I stressed out, and for being there for me when I needed the most.

Lastly—and most importantly – I would like to extend my deepest gratitude to my father and my mother, Mr. Qin Gao and Ms. Dongyun Wang. I could have easily written another thesis to convey my heartfelt appreciation towards their unconditional support, trust and love. They are the reason I become who I am today. I would like to thank them for raising me with all the love and kindness in the world, for teaching me to be empathetic, genuine, and hard-working, for being my source of courage and wisdom, and for always believing in me even when I start to doubt myself. I would like to thank them for all they've been to me, for all they've done for me, and for all that they are. This dissertation is dedicated to them.

For my loving parents and grandparents.

You are my anchor
and
my inspiration.

Introduction

Cancer develops when cells divide and grow uncontrollably. Epidemiological studies have identified many risk factors for cancer including radiation, alcohol drinking, obesity, tobacco uses, hormone levels et cetera(1). Family history and genetic composition have also been seen as important risk factors of cancer(2). For instance, family history contributes greatly to increased risk of prostate cancer(3, 4). A meta-analysis of 33 epidemiological studies found that subjects with a first-degree relative family history of prostate cancer were at approximately 2.5 folds increase of lifetime risk(5). Family history also plays a critical role in other cancers etiology such as breast(6) and colorectal cancer(7).

While family history is an important cancer risk factor, it has limitation in distinguish genetic from non-genetic contributions as some family members also share similar lifestyles and exposures. To better understand the genetic component of a given cancer within families, several types of studies have been performed in the past few decades: a) twin studies to implicate the strong heredity in cancer susceptibilities. Analyses from the Nordic Twin Study of Cancer (NorTwinCan) found significant estimates of heritability of 57% for prostate cancer, 31% for breast cancer, and 58% for skin melanoma(8). b) segregation studies simulated under various genetic models to identify inheritance patterns of cancer(9). Past segregation studies have favored a highly penetrance, autosomal dominant genetic model for breast cancer(10-12). c) linkage analysis to identify cancer predisposition genes. Using large breast cancer pedigrees, linkage analyses mapped high-penetrance breast cancer genes *BRCA1(13)* and *BRCA2(14)*.

Linkage and segregation analyses work best for mendelian diseases where high penetrant variants segregate according to clear patterns within families. However, for common, complex disease such as cancer, the genetic risk is comprised of multiple alleles with no single allele being fully deterministic for driving tumorigenesis. Hence, to identify alleles associated with complex diseases, research focus has shifted from highly penetrant alleles clustered within families to more common variant present in larger, unrelated populations. Initial efforts to identify modestly penetrant alleles relied on resequencing candidate genes predicted to play a role in cancer risks. Some convincing findings have been reported for some cancers(15), however, only few associations were robustly validated in independent cohorts. For instance, the genes for androgen receptors attracted significant attention given its known role in prostate carcinogenesis but extensive variant annotation across genes in prostate cancer cases and controls found no inherited variant associated with risk(16). This suggests that a less biased and more robust approach is needed to identify common alleles associated with complex diseases which leads to the era of genome-wide association studies (GWAS).

For the past two decades, several advances in genetic research made the implementation of GWAS possible, including the sequencing of the human genome, the publication of International HapMap Project and the 1000 Genomes Project, the availability of high-throughput genotyping, and the development of statistical methods to interpret and impute massive amount of genetic data. GWAS scan the genome for polymorphisms, usually single nucleotide polymorphisms (SNPs) that are associated with disease of interest. To date, multiple GWAS have been reported for many of the major cancers in European populations, including breast cancer(17-20), prostate cancer(21-23), lung cancer(24-27), colorectal cancer(28-31), gastric cancer(32, 33), ovarian cancer(34), and pancreatic cancer(35-37).

Most of these common variants (allele frequency >1%) identified by GWAS are associated with modest increase in disease risk, with odds ratios (ORs) generally less than 1.5(38). Individually, these SNPs may not be informative in evaluating the risk of developing cancer, but the collective use of large numbers of common variants to create a polygenic risk score (PRS) have demonstrated abilities to modify and individualize cancer risk estimates(39-41). The recent work by Mavaddat et al. found that one standard deviation change of a breast cancer PRS is associated with a 1.61 folds increase of breast cancer, and that the lifetime risk of overall breast cancer for the top percentile of the PRS is 32.6%(42). PRS can also be useful in further stratifying the risk among carriers of a pathogenic variant. Take breast cancer as an example, Kuchenbaecker et al. found a statistically significant association between a PRS and breast cancer risk among *BRCA1* and *BRCA2* variant carriers (HR:1.14, 95%CI: 1.11- 1.17 for *BRCA1* carriers, and HR: 1.22, 95%CI: 1.17-1.28 for *BRCA2* carriers)(43).

In this dissertation, I present three projects to investigate the important role of PRS in understanding cancer etiology and cancer risks. The first project is a mendelian randomization study using PRS as proxies for adiposity to examine the causal relationship between adiposity at different life stages (birth weight, childhood obesity, adult BMI and WHR) and risk of breast, ovarian, prostate, colorectal and lung cancers. The results of this project may help us understand the underlying genetic architecture of the five cancers studied. The second project performs validation analysis of a synthetic breast cancer prediction model utilizing both PRS and classic risk factors from questionnaire data. This project demonstrates the contribution of PRS in improving performance and accuracy of breast cancer risk prediction models. The third project investigated the joint effect of PRS and rare pathogenic variant in breast cancer predisposition genes in the general population. Findings from this project can help better develop risk prediction model incorporating both common and rare variants of breast cancer. It also shed light into developing more individualized breast cancer prevention and screening strategies.

CHAPTER 1

Mendelian Randomization study of adiposity-related traits and risk of breast, ovarian, prostate, lung and colorectal cancer

1.1 Introduction

Obesity influences risk for many chronic diseases such as cancer, cardiovascular disease and diabetes(44). Observational studies have found associations between body mass index (BMI) and various cancer types including increasing risk of postmenopausal breast (45), colorectal(46), endometrial(47), and pancreatic cancer(48, 49) and decreasing risk of lung cancer and premenopausal breast cancer (50). However, the mechanisms underlying the contribution of obesity to cancer risk remains poorly understood. It is also unclear whether these associations between obesity and cancer in observational studies are causal. For instance, the observed increased risk of lung cancer among individuals with low BMI may be due to residual confounding by smoking or weight loss resulting from chronic lung disease(51).

Recent studies have also found time-dependent associations between assessment of adiposity and subsequent cancer risk. Higher adiposity at young ages is inversely associated with both pre- and postmenopausal breast cancer(52). In contrast, higher adult BMI is positively associated with postmenopausal breast cancer risk(45, 53, 54). Evidence also suggests that childhood obesity may be associated with ovarian cancer independent of adult BMI(55). These findings demonstrate a dynamic relationship between adiposity and cancer development during different time frames of life that requires a deeper investigation.

Elevated waist-to-hip ratio (WHR), representing a higher abdominal fat distribution, is associated with multiple hormonal and metabolic changes including insulin resistance and hyperinsulinemia that may increase risk of chronic disease such as cancer(56-58). Previous studies examining WHR and breast cancer risk indicated a positive association, which remained positive after adjusting for BMI(54, 59). Some studies also suggest that measures of abdominal adiposity are more predictive of colorectal cancer than BMI(60, 61). Thus, further investigations on the contribution of WHR to cancer risk may improve our understanding of the relationship between body fat distribution, obesity, and cancerogenesis.

Mendelian randomization (MR) is a technique that uses genetic predictors of risk factors as instrumental variables to assess the possible causal associations between risk factors and diseases(62). As genetic variants are fixed at conception and generally independent of confounders, such an approach seeks to eliminate potential reverse causality and reduce confounding bias(63, 64). To our knowledge, there has not been any large-scale MR study assessing the potential causal relationship between obesity across different life stages and risk of multiple cancers.

In this study, we performed MR analysis to estimate the causal relationship between adiposity at different life stages (birth weight, childhood BMI, adult BMI and WHR) and risk of breast, ovarian, prostate, colorectal and lung cancers. We leveraged the results of recently published large-scale genome-wide association studies (GWAS) of adiposity-related traits to define a genetic score for each trait. We then assessed the associations between these scores and risks of five cancers from the Genetic Associations and Mechanisms in Oncology (GAME-

ON) Consortium, which include 51 537 cancer cases and 61 600 controls from 32 participating studies.

1.2 Materials and Methods

The GAME-ON post-GWAS initiative

The Genetic Associations and Mechanisms in Oncology (GAME-ON) Initiative is a network of cancer-specific consortia engaged in GWAS and post-GWAS research. It includes five cancer-specific consortia: DRIVE (breast), CORECT (colorectal), ELLIPSE (prostate), FOCI (ovarian) and TRICL (lung) (Table 1.1). GWAS data from 32 studies (all European ancestry) contributing to the GAME-ON consortium were imputed using the 1000 Genomes reference panel (phase I version 3). Studies contributed summary statistics only to cancer- specific meta-analyses. Further information regarding imputation and analyses can be found in Fehringer et al.(65) and Zhang et al.(66).

Table 1.1: Participants and Studies Included in the Genetic Associations and Mechanisms in Oncology (GAME-ON) consortium by Cancer Site and Subtype.

Cancer Type	**Cancer subtype**	**Cases**	**Controls**	**GWAS studies**
Breast	**All**	15,748	18,084	11
	ER-negative	4939	13,128	8
Colorectal	**All**	5,100	4,831	6
Lung	**All**	12,160	16,838	6
	Adenocarcinoma	3,718	15,871	6
	Squamous	3,422	16,015	6
Ovarian	**All**	4,369	9,123	3
	Clear-cell	356	9,123	3
	Endometrioid	715	9,123	3
	Serous	2,556	9,123	3
Prostate	**All**	14,160	12,724	6
	Aggressive	4,450	12,724	6
Total	**All**	51,537	61,600	32

Identification of SNPs associated with birth weight, childhood obesity and adult BMI and WHR.

To calculate the genetic scores, we considered SNPs that were genome-wide significant ($p < 5x10^{-8}$) in the largest GWAS to date for each trait as follows: a) 7 SNPs of birth weight from Horikoshi et al.(67), b) 15 SNPs of childhood BMI from Felix et. al.(68), c) 77 SNPs of adult BMI from Locke et al. (SNPs from primary meta-analysis of European-descents only) (69) and d) 14 SNPs of adult WHR from Heid et al. (70). All GWAS were restricted to individuals of European ancestry. For all identified SNPs, we obtained the chromosome and position, the nearest gene, the risk allele, and trait-specific association estimates and standard errors reported in the papers above. For each SNP, we also extracted cancer-specific effect estimates and p-values from the GAME-ON consortium (Supplementary Table 1.1).

Several SNPs associated with birth weight, childhood BMI, adult BMI and WHR were not found in GAME-ON data for ovarian endometrioid cancer subtype, lung cancer, and colorectal cancer. For these SNPs, proxy SNPs (r^2>0.9, 1000 Genomes Northern and Western European population) were used in the analysis instead (Supplementary Table 1.2) There were no overlaps (lead SNPs within 250kb) among the GWAS-identified loci for different adiposity-related traits except childhood BMI and adult BMI, for which we found ten overlap regions: *SEC16B, TNNI3K, FTO, MC4R, TMEM18, TFAP2B, OLFM4, ADCY3, GPR61/GNAT2, GNPDA2* (Supplementary Figure 1.1).

Statistical Analysis

We conducted MR analyses to estimate the association between adiposity-related traits and cancer using summary genetic association statistics, as described in Burgess et al.(71). Specifically, the ratio estimate ($\hat{\beta}$) of the effect of a risk factor (X) on disease outcome (Y) using genetic variants k=1,...,K can be calculated as $\hat{\beta}= \frac{\sum_k X_k Y_k \sigma_{Y_k}^{-2}}{\sum_k X_k^2 \sigma_{Y_k}^{-2}}$ where X_k is the per-allele effect of SNP k with the risk factor, Y_k is the per-allele change in the log odds ratio for the cancer being tested, and σ_{Yk}^2 is the standard error for Y_k. The summary statistics X_k, Y_k and σ_{Yk}^2 are taken from the GWAS for the risk factor and for cancer, respectively. The standard error of $\hat{\beta}$ is given by: se($\hat{\beta}$) = $\sqrt{\frac{1}{\sum_k X_k^2 \sigma_k^{-2}}}$[16, 21]. Under certain assumptions(72), the ratio estimate $\hat{\beta}$ can be interpreted as the causal log odds ratio of cancer risk associated with one unit change in the adiposity-related traits (birth weight, childhood BMI, adult BMI, and WHR).

Since some cancers demonstrate etiologic heterogeneity by histologic subtype or clinical characteristics, we also conducted the following cancer-specific subgroup analyses: estrogen receptor negative (ER-) breast cancer; clear cell, endometrioid and serous ovarian cancer; adenocarcinoma and squamous lung cancer; and aggressive prostate cancer (defined as a Gleason score of ≥8, a disease stage of 'distant', a prostate-specific antigen level of >100 ng/ml or death from prostate cancer(73). In addition, sensitivity analyses were performed excluding the overlap loci between childhood BMI and adult BMI. One key assumption for MR analysis is no pleiotropic effect. Thus, Egger regression was performed to evaluate directional pleotropic effect for adult and childhood BMI (74) to provide effect estimates after adjusting for potential pleiotropic effects. The intercept from Egger regression provides a test for directional pleiotropy (the average direct effects of adiposity-increasing variants increase [or decrease] cancer risk). Under the assumption that the SNPs' direct effects on cancer risk are independent of their association with body mass index, Egger regression provides an unbiased estimate of the causal effect of genetically predicted BMI on cancer. Unless otherwise noted, all p-values are unadjusted for multiple testing.

1.3 Results

We estimated the associations between adiposity-related genetic scores and risk of five cancers (Table 1.1). Figures comparing results across cancers are in Supplementary Figure 1.2.

Breast cancer

The risk of breast cancer decreased with increasing genetic score for childhood BMI (OR=0.71 per s.d. increase in childhood BMI; 95%CI: 0.60, 0.80; $p=6.5x10^{-5}$), and also with increasing genetic score for adult BMI (OR=0.66 per s.d. increase in adult BMI; 95%CI: 0.57, 0.77; $p=2.5\times10^{-7}$) (Table 1.2). Similar associations were found for ER negative breast cancer (OR=0.69, 95%CI: 0.53, 0.98, $p=5.8\times10^{-3}$ for childhood BMI; OR=0.59, 95%CI: 0.46, 0.75, $p=2.0\times10^{-5}$ for adult BMI). We did not observe an association between the genetic score for birth weight and breast cancer and observed an inverse association between the genetic score for WHR and breast cancer risk (OR=0.73; 95%CI: 0.54, 1.00; p=0.05).

Ovarian cancer

The estimated association between the genetic scores for higher adult BMI is associated with increased risk of overall ovarian cancer. One standard deviation increase in genetically predicted adult BMI was associated with 35% increased risk of ovarian cancer (OR=1.35, 95%CI:

1.05,1.72; p=0.017). We did not find strong evidence of associations between genetically predicted birth weight, childhood BMI, or WHR and ovarian cancer risk.

Lung cancer

We observed a positive association between genetically predicted adult BMI and overall lung cancer (OR=1.27, 95%CI: 1.09, 1.49; p-value=$2.9x10^{-3}$)(Table 1.2). This association appeared restricted to squamous cell lung cancer (OR=1.54, 95%CI: 1.20, 1.96; p=$6.6x10^{-4}$), as we found no strong evidence for association with lung adenocarcinoma (OR=0.93, 95%CI:0.73,1.19, p=0.59). We also did not find strong evidence for association between either genetically predicted birth weight or childhood BMI and lung cancer risk.

Prostate Cancer

We found a positive association between the genetic score for birth weight and aggressive prostate cancer (OR=1.63 per s.d. unit increase in birth weight, 95%CI: 1.03, 2.57, p = 0.037). No strong evidence was found for associations between prostate cancer and any other adiposity measures.

Colorectal Cancer

We found an increase in risk of colorectal cancer per s.d. increase of genetically predicted adult BMI (OR=1.39, 95%CI: 1.06, 1.82, p = 0.016). No associations were found between birth weight, childhood BMI or waist-hip-ratio and colorectal cancer risk.

Table 1.2: Mendelian randomization odds rations (ORs) of birth weight, childhood obesity, adult BMI, and waist-hip-ratio across five different cancer types obtained using summary data from GAME-ON consortium.

		Birth Weight		**Childhood BMI**		**Adult BMI**		**Waist-hip-ratio**	
		OR (95%CI)	**p-value**	**OR (95%CI)**	**p-value**	**OR (95%CI)**	**p-value**	**OR (95%CI)**	**p-value**
Breast Cancer	*All*	1.22 (0.93, 1.60)	0.15	0.71 (0.60, 0.80)	$6.5x10^{-5}$	0.66 (0.57, 0.77)	$2.5x10^{-7}$ *	0.73 (0.53,1.00)	0.051
	ER_negative	1.01 (0.66, 1.53)	0.98	0.69 (0.53, 0.98)	0.0058	0.59 (0.46, 0.75)	$2.0x10^{-5}$ *	0.74 (0.45, 1.21)	0.23
Ovarian Cancer	*All*	1.07 (0.69, 1.65)	0.75	1.07 (0.82, 1.39)	0.62	1.35 (1.05,1.72)	0.017	1.19 (0.73, 1.94)	0.50
	Clear_cell	2.75 (0.82, 9.30)	0.10	1.45 (0.68, 3.09)	0.34	1.68 (0.84, 3.36)	0.14	1.31 (0.32, 5.30)	0.71
	Endometrioid	0.79 (0.33, 1.92)	0.60	1.47 (0.86, 2.52)	0.16	1.34 (0.80, 2.26)	0.26	1.03 (0.38, 2.84)	0.95
	Serous	0.85 (0.50, 1.45)	0.56	0.91 (0.65, 1.26)	0.56	1.30 (0.97, 1.76)	0.089	1.34 (0.73, 2.46)	0.34
Prostate Cancer	*All*	1.33 (0.96, 1.82)	0.082	1.01 (0.83, 1.22)	0.91	1.01 (0.84, 1.21)	0.97	1.02 (0.72, 1.46)	0.90
	Aggressive	1.63 (1.03, 2.57)	0.037	1.10 (0.83, 1.45)	0.49	1.11 (0.85, 1.44)	0.44	1.19 (0.71, 1.98)	0.51
Lung Cancer	*All*	0.93 (0.70, 1.23)	0.64	1.01 (0.85, 1.2)	0.90	1.27 (1.09, 1.49)	$2.9x10^{-3}$	1.15 (0.80, 1.66)	0.46
	Adenocarcinoma	0.95 (0.62, 1.46)	0.83	0.90 (0.69, 1.19)	0.47	0.93 (0.73, 1.19)	0.59	0.90 (0.51, 1.58)	0.71
	Squamous	0.99 (0.64, 1.52)	0.94	1.08 (0.82, 1.43)	0.57	1.54 (1.20, 1.96)	$6.6x10^{-4}$ *	1.33 (0.75, 2.36)	0.33
Colorectal Cancer	*All*	0.69 (0.44, 1.10)	0.12	1.20 (0.90, 1.59)	0.21	1.39 (1.06, 1.82)	0.016	1.29 (0.75, 2.22)	0.35

* denotes analyses that have p<0.001 after Bonferroni Correction for 48 tests

BMI SNP rs12016871 has been merged into rs9581854 and thus rs9581854 was used for analysis instead

Overlap in adiposity SNP scores

None of the pairs of adiposity-trait SNP scores overlap (within 250kb) except childhood BMI and adult BMI, which overlap at ten loci: *SEC16B, TNNI3K, FTO, MC4R, TMEM18, TFAP2B, OLFM4, ADCY3, GPR61/GNAT2, GNPDA2*. To assess the specificity of the observed associations between childhood and adult BMI and cancer risk, we repeated the analyses after removing the SNPs from the overlapping loci. The associations remained between adult BMI and breast and lung cancer, whilst the associations between childhood BMI and breast was attenuated after removing the overlapping loci (Table 1.3).

Egger Regression

With the possible exception of genetically predicted childhood BMI and breast cancer risk, the Egger regression did not reveal any strong directional pleiotropic effect on the risk estimation of genetically predicted adult BMI/childhood BMI/WHR/birth weight on various cancers (Table 1.4). All estimated intercept from the Egger regression are near zero. The effect estimates from the Egger Regression are generally in the same direction as the estimates from the MR analysis and larger in magnitude, except for lung cancer. We detect no strong pleiotropic effect on the risk estimation of genetically predicted adult BMI and lung cancer (intercept=0.011, p=0.057) but found no positive association between the BMI score on lung cancer in the Egger regression analysis (OR=0.90, 95% C.I. 0.51-1.29; p=0.59).

Table 1.3: Mendelian randomization odds ratios (ORs) of childhood BMI and adult BMI across five different cancer types obtained using summary data from GAME-ON consortium, excluding overlap loci (*SEC16B, TNNI3K, FTO, MC4R, TMEM18, TFAP2B, GNAT2, OLFM4, ADCY3, GNPDA2*)

		Childhood BMI		**Adult BMI**	
		OR (95%CI)	**p-value**	**OR (95%CI)**	**p-value**
Breast Cancer	*All*	1.05 (0.74,1.48)	0.80	0.75 (0.62, 0.92)	4.7x10^{-3} *
	ER_negative	1.17 (0.68, 2.03)	0.57	0.66 (0.49, 0.91)	0.011
Ovarian Cancer	*All*	0.58 (0.34,1.01)	0.053	1.26 (0.93,1.72)	0.14
	Clear_cell	0.70 (0.15,3.25)	0.69	1.44 (0.60,3,43)	0.42
	Endometrioid	0.67 (0.22,2.03)	0.47	0.84 (0.43,1.64)	0.61
	Serous	0.54 (0.27,1.06)	0.07	1.43 (0.98,2.10)	0.062
Prostate Cancer	*All*	1.29 (0.88,1.87)	0.19	1.09 (0.86,1.37)	0.48
	Aggressive	1.32 (0.77,2.29)	0.32	1.24 (0.89,1.73)	0.20
Lung Cancer	*All*	0.90 (0.63,1.28)	0.55	1.41 (1.16,1.73)	6.8x10^{-4} *
	Adenocarcinoma	1.06 (0.62,1.83)	0.83	1.00 (0.74,1.36)	0.99
	Squamous	0.66 (0.38,1.14)	0.13	1.73 (1.27,2.38)	5.3x10^{-4} *
Colorectal Cancer	*All*	0.85 (0.48,1.50)	0.57	1.36 (0.96,1.92	0.08

* denotes analyses that have p<0.001 after Bonferroni Correction for 48 tests

Table 1.4. Effect estimates from Egger regression for adult BMI, childhood BMI, birth weight, and WHR

Adult BMI		Egger regression					
	MR OR	Intercept	Standard error	p	OR_egg	Standard Error	p
Breast cancer	0.66 (0.57, 0.77)	0.0035	0.0056	0.53	0.59	0.1949	0.0076
Ovarian Cancer	1.35 (1.05,1.72)	-0.0093	0.0088	0.29	1.80	0.3082	0.054
Prostate Cancer	1.01 (0.84, 1.21)	0.0096	0.0066	0.15	0.74	0.2324	0.19
Lung Cancer	1.27 (1.09, 1.49)	0.011	0.0057	0.057	0.90	0.2000	0.59
Colorectal	1.39 (1.06, 1.82)	0.0082	0.0098	0.40	1.08	0.3317	0.82

Childhood BMI		Egger regression					
	MR OR	Intercept	Standard Error	p	OR_egg	Standard Error	p
Breast cancer	0.71 (0.60, 0.80)	0.048	0.0274	0.026	0.34	0.2078	0.0017
Ovarian Cancer	1.07 (0.82, 1.39)	-0.053	0.0436	0.12	2.44	0.3271	0.10
Prostate Cancer	1.01 (0.83, 1.22)	-0.020	0.0332	0.42	1.38	0.2462	0.42
Lung Cancer	1.01 (0.85, 1.2)	-0.0015	0.0877	0.95	1.04	0.2076	0.92
Colorectal	1.20 (0.90, 1.59)	-0.020	0.1483	0.41	1.63	0.3464	0.22

Table 1.4. Effect estimates from Egger regression for adult BMI, childhood BMI, birth weight, and WHR (CONTINUED)

WHR		Egger regression					
	MR OR	Intercept	Standard Error	p	OR_egg	Standard Error	p
Breast cancer	0.73 (0.53,1.00)	0.0048	0.0263	0.85	0.63	0.8307	0.58
Ovarian Cancer	1.19 (0.73, 1.94)	-0.037	0.0424	0.38	3.67	1.3153	0.32
Prostate Cancer	1.02 (0.72, 1.46)	0.046	0.0310	0.14	0.25	0.9747	0.15
Lung Cancer	1.15 (0.80, 1.66)	-0.017	0.0316	0.60	1.97	1.0440	0.52
Colorectal	1.29 (0.75, 2.22)	-0.068	0.0458	0.14	10.38	1.4318	0.10

Birth weight		Egger regression					
	MR OR	Intercept	Standard Error	p	OR_egg	Standard Error	p
Breast cancer	1.22 (0.93, 1.60)	0.040	0.0300	0.18	1.75	0.5831	0.34
Ovarian Cancer	1.07 (0.69, 1.65)	0.069	0.0480	0.15	3.46	0.9274	0.18
Prostate Cancer	1.33 (0.96, 1.82)	0.0043	0.0346	0.90	0.82	0.6856	0.77
Lung Cancer	0.93 (0.70, 1.23)	0.0011	0.0307	0.97	1.10	0.6000	0.88
Colorectal	0.69 (0.44, 1.10)	-0.026	0.0510	0.96	1.38	0.9950	0.75

Associations between individual adiposity-related SNPs and cancer risk

Figure 1.1 illustrates SNP-specific associations with risk of breast (top left), ovarian (top right), colorectal (bottom left), and lung cancer (bottom right) versus the documented associations between each SNP and adult BMI. After excluding potential outliers (rs1558902 and rs17024393 for breast and ovarian cancer; rs17105752 for lung cancer), the MR analysis still show strong evidence for association between predicted adult BMI and cancer (for breast cancer, OR: 0.69 per s.d. increase in BMI, 95%CI: 0.58, 0.82, p=3.0x10^{-5}; for ovarian cancer, OR: 1.32 per s.d. increase in BMI; 95%CI: 1.01, 1.74, p=0.041; for lung cancer, OR: 1.30 per s.d. increase in BMI, 95%CI: 1.10, 1.52, p=1.5x10^{-3}).

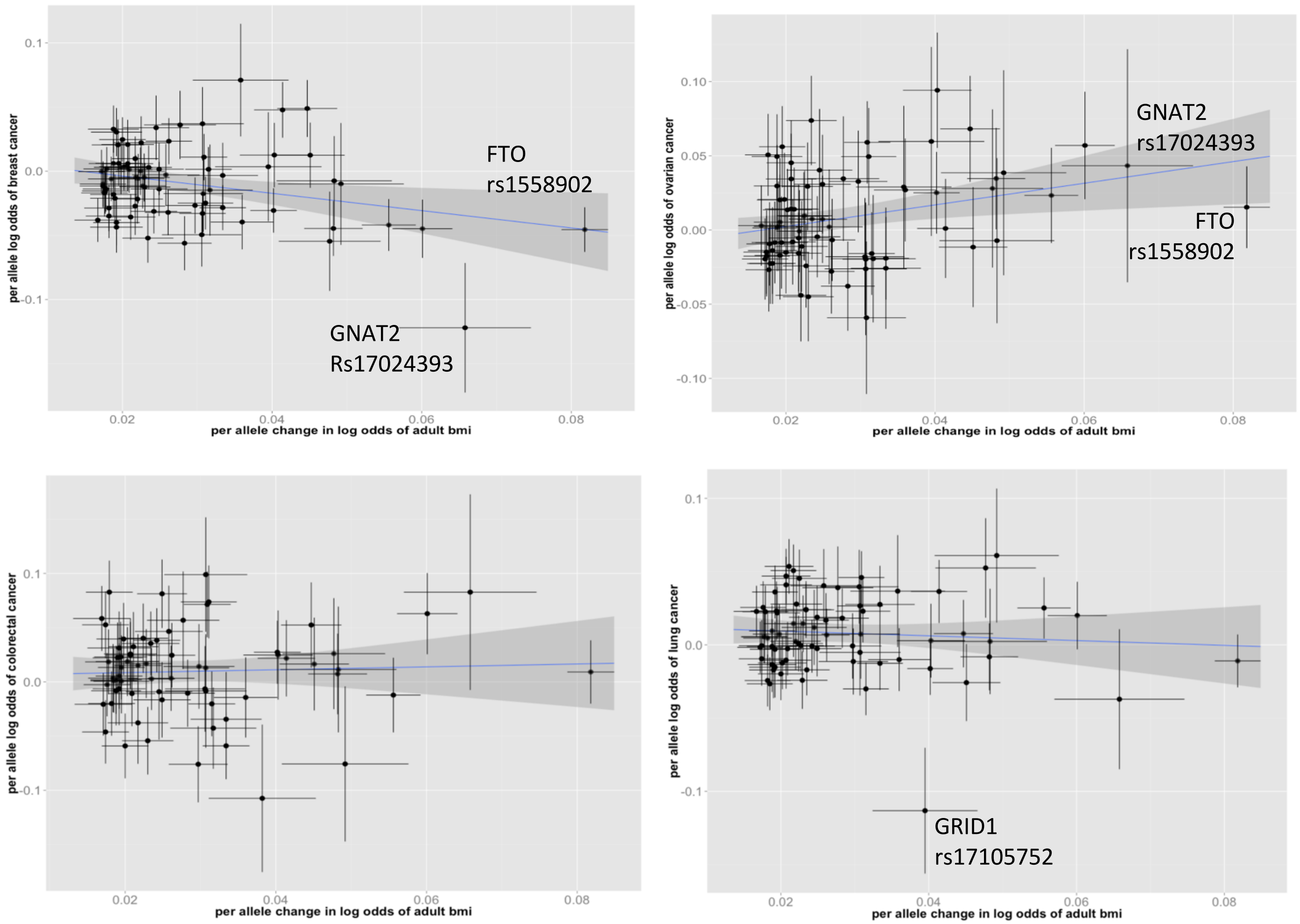


Figure 1.1: Scatterplot of SNP-specific effects for the associations with adult BMI and a) breast cancer, b) ovarian cancer risk, c) colorectal cancer, d) lung cancer for all 77 BMI-associated SNPs. SNP-specific vertical and horizontal bars correspond to standard errors for the breast/ovarian/colorectal/lung cancer association and BMI association respectively. The shaded region corresponds to 95%CI of the association between BMI and cancer risk.

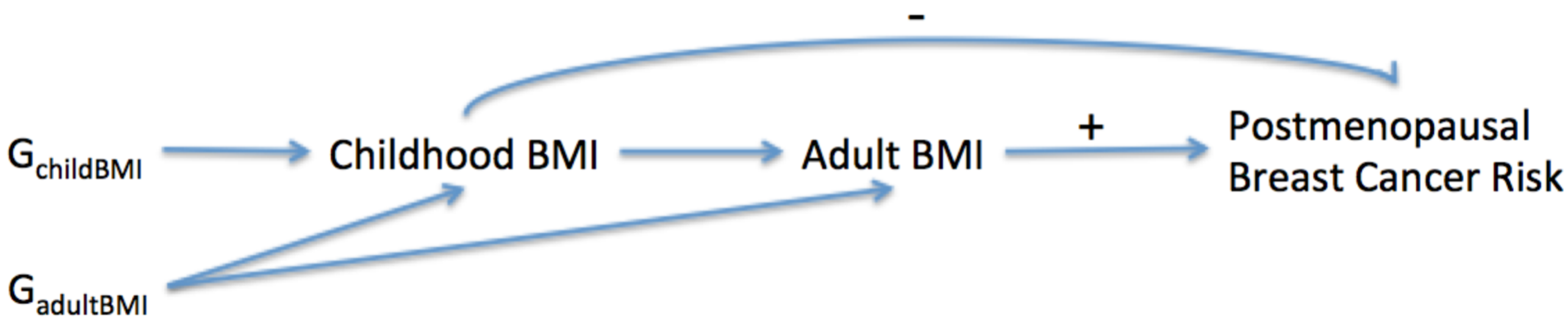


Figure 1.2: DAG demonstrating one potential explanation of how genetic variants influence postmenopausal breast cancer risk

1.4 Discussion

In this study, we found an inverse association between the genetic scores for childhood BMI and adult BMI and risk of both overall and ER-negative breast cancer. Further, the genetic score for adult BMI was associated with increased risk of ovarian, lung, squamous lung, and colorectal cancer.

Consistent with our results, observational studies have shown an inverse association between higher childhood BMI and both premenopausal and postmenopausal breast cancer(52, 75, 76). In contrast to our findings, observational studies have found that higher adult BMI was positively associated with postmenopausal breast cancer (77, 78), this includes a recent instrumental variables analysis using offspring BMI as an instrument for parental BMI (79). However, we found decreased risk of breast cancer with higher adult BMI genetic score, even though the majority of women that contributed to our analysis were postmenopausal (62%). We did not have access to summary statistics stratified by menopausal status but findings from a recent MR analysis of a large data set from the Collaborative Oncological Gene-Environment Study (COGS) are consistent with our study. The MR estimate from that study for 5kg/m^2 increase in BMI was 0.65 (95% CI: 0.56-0.75; p=3.32x10^{-10}) for overall breast cancer. This inverse association was consistent across both pre- and post- menopausal women: OR: 0.44, 95%CI: 0.31, 0.62, p=9.91x10^{-8} for premenopausal women, and OR: 0.57, 95%CI: 0.46, 0.71, p=1.88 x10^{-6} for postmenopausal women(80).

Thus, at first sight, our results might suggest that increasing adult BMI is associated with reduced postmenopausal breast cancer risk, contradicting the epidemiological evidence. There are several possible explanations for this discrepancy. One hypothesis to explain this is illustrated in the causal graph in Figure 1.2. The positive association between observed adult BMI and postmenopausal breast cancer in observational studies may be driven by adult weight gain, which has been linked to increased postmenopausal breast cancer risk (81). This weight gain could be due to environmental factors that are not captured by genetic risk scores (82). The effects of the BMI-associated SNPs on breast cancer risk may be mediated through their effects on BMI in childhood and young adulthood, which have been shown to be inversely associated with postmenopausal breast cancer risk (as shown in Figure 1.2 by a negative sign)(52, 75, 76). It is also possible that the adult BMI genetic score is a stronger instrumental variable for early life BMI as compared to later life BMI that is largely determined by environment, and that the inverse association of early life BMI with breast cancer may counterbalance the association with BMI later in life.

Consistent with our hypothesis, an observational study examining the association between weight change across the life-course and breast cancer risk in the Nurses Health Study (77,232 women from 1980-2012) found that weight at age 18 was inversely associated with both pre-and postmenopausal incidence of breast cancer. In contrast, adult weight gain was positively associated with both pre and post-menopausal breast cancer risks(83).

Three of the four strongest (largest effect size) adult BMI SNPs are also associated with childhood BMI. In sensitivity analyses excluding overlapping loci from the adult and childhood BMI scores, we still observed an inverse association with breast cancer for the genetic score for adult BMI (OR:0.75, 95%CI: 0.62, 0.92, p=4.7x10^{-3}). But the association between childhood BMI score and breast cancer was attenuated (Table 1.3). However, we found the genetic instrument for adult BMI was associated with childhood BMI (and vice versa, Supplementary Table 1.5) even after removing the overlapping loci. This suggests care is required when interpreting these results. The association between predicted adult BMI and breast cancer risk may reflect effects on a pathway distinct from childhood BMI, or it may simply reflect the shared genetics of early- and later-life BMI.

We found that a genetic risk score predicting higher BMI was associated with increased risk of lung cancer overall and lung squamous carcinoma in particular. Studies have found obesity to be associated with high insulin resistance(84) which is positively associated with lung cancer risk(85), suggesting the observed positive associations may be mediated by insulin resistance. Multiple studies have reported an inverse relationship between BMI and lung cancer among smokers but no or a weakened association among never smokers(50, 86-88). These results may be due to residual confounding, reverse causation, or effect modification by smoking(51, 87, 88). We did not have access to individual level genetic and smoking data for this study, so our Mendelian Randomization estimate of the effect of body mass index on cancer risk should be interpreted with care: it represents an average of the effects across smoking status. (83% of the participants in the lung cancer GWAS were ever smokers.) Future work in the large OncoArray Network will be able to perform stratified analysis by smoking status(89).

Another concern with our MR analyses on adult BMI and lung cancer risk is that some BMI-associated SNPs are associated with neurological response and stress related behavior that affect smoking (69, 90, 91). To assess whether our results were driven by pleiotropic effects, we performed additional analysis excluding SNPs that are associated with smoking initiation or schizophrenia (rs1191560, rs11030104(69)). We still observe a positive association between genetically predicted adult BMI and lung cancer (OR = 1.25; 95%CI: 1.07,1.47; p=6.0x10^{-3}). It is also worth noting that although we detect limited directional pleiotropy for the association between predicted adult BMI and lung cancer risk, we found positive association between the genetically predicted adult BMI and lung cancer risk in the MR Egger regression analysis (p=0.59). This could be due to bias caused by other type of pleiotropy or lack of statistical power.

Our MR results showed an increased risk in ovarian cancer with increasing adiposity measures across different life stages; this is consistent with previous observational studies(92, 93). Obesity in adolescence is associated with increased risk of ovulatory infertility that may increase risk of ovarian cancer(94). In addition, obesity is also associated with an increased level of insulin-like growth factor 1(IGF-1)which increases cell proliferation and modulates synthesis and bioavailability of sex steroids hormones that are involved in ovarian cancer etiology(95, 96). The opposite risk profiles between breast and ovarian cancer also suggest that adiposity determined by genetic variants has different underlying mechanisms in relation to breast versus ovarian cancer carcinogenesis.

Our analyses suggest that adult BMI is associated with increased risk of colorectal cancer, consistent with the published epidemiological literatures. Keimling et. al. found a 14% increase in colorectal cancer risk per s.d. increase in BMI(97). A recently published MR study also found that genetically influenced BMI was associated with higher risk of colorectal cancer (OR: 1.50 per 5kg/m^2 increase, 95%CI: 1.13, 2.01)(98). The mechanisms linking adiposity and colorectal cancer are not yet fully understood. One possible explanation is that obese individuals have higher leptin secretion from the white adipose tissue, and the binding of leptin to its receptor in the colon epithelium activates biological pathways implicated with colorectal cancer(99).

Although there is evidence that genetically predicted BMI is associated with breast and lung cancer, the underlying mechanisms remain unknown. There are many factors that can influence both adiposity and cancer risks such as physical activity, mental stress, insulin resistance, and exposure to hormones secreted by adipose tissue. Further studies incorporating these factors might provide a better understanding of the mechanism underlying the relationship between adiposity and cancer risk. As data on SNP-specific function emerges, future studies can also carefully categorize SNPs by their functionality, and perform MR analysis for different groups of SNPs. This will allow us to parse out specific set of SNPs and further evaluate which pathway(s) are of importance in the adiposity-cancer association. In addition, gene-environmental interaction can also provide additional insights in understanding the mechanism underlying adiposity and cancer risk. Although not feasible in the GAME-ON data, in the newly completed OncoArray data where we have individual data on menopausal status, hormone therapy, reproductive factors for breast cancer, and smoking status for lung cancer, we will be able to perform gene-environment interaction analysis in the near future(89).

Our study has several limitations. The summary-level statistics approach does not allow us to perform analyses stratified by covariates such as menopausal or smoking status. The summary statistics also did not permit us to explore the non-linearity of the association between obesity and cancer risk, which have been observed in a previous study(50). We note that nonlinearity does not invalidate the test of association, although it may complicate the interpretation of the effect estimate(100). Finally, the statistical power is limited by both the proportion of the adiposity risk factors explained by the genetic instruments and the sample size in the cancer genetic association studies (101), and this is particularly an issue for analyses of rare cancer subtypes.

MR analyses are only valid under a few strong assumptions(72, 102): a) valid association between SNPs and risk factors; b) SNPs are not associated with other confounders of the risk factors and outcome; c) SNPs only affect the outcome through their effect on the risk factors (no pleiotropic effects). The second and third assumptions are the most concerning and requires careful interpretation. For b), population stratification may be a source of confounding but the original studies saw little evidence for such bias and all have appropriately controlled for it. Assumption c) raises the most concern, especially for relationship between genetically predicted adult BMI and breast cancer risk. As noted before, the association between the genetic instrument for adult BMI and childhood BMI (and vice versa) makes the associations between these instruments and breast cancer difficult to distinguish. This is a situation where the InSIDE (Instrument Strength Independent of Direct Effects) assumption—the direct effect of a SNP on cancer risk is uncorrelated with its association with trait of interest—does not hold(74). There are other reasons why assumption c) might not hold. For example, two SNPs known to be associated with breast cancer are near the *FTO* gene, raising the possibility that obesity-related variants may affect cancer risk through other pathways (103). To test for and correct for bias due to pleiotropy where the InSIDE assumption holds, we performed Egger regression for all traits investigated (Table 1.4 and Supplementary Table 1.4). Egger regression show limited evidence for any directional pleiotropic effects influencing associations between genetically predicted adiposity traits and the cancer studied here.

Despite these issues, our study also has several important strengths. Many studies examining BMI and cancer risk in the past are susceptible for recall bias, confounding and reverse causation(104), none of which are concerns of MR studies. In addition, we used summary statistics from the largest meta-analyses of primary GWAS of these cancer types to date, which improves our power of detecting real causal effects. Moreover, by comparing results across cancer types, we are able to demonstrate specificity of the association between genetic markers of adiposity and particular cancers.

In summary, we found associations between genetic scores for higher adult BMI and increased risk of lung, colorectal and ovarian cancers. Additionally, we observed an inverse association of both genetically predicted childhood BMI and adult BMI with breast cancer. Given the strength of the epidemiological and biological studies linking obesity after menopause with increased risk of breast cancer, this highlights the need for caution when interpreting the results of MR analyses. Our study supports the hypothesis of dynamic relationships between genetic variation underlying obesity and different cancer risks throughout life. To better interpret the complexity of the relationship between adiposity and breast cancer, future investigations that effectively distinguish childhood versus adulthood obesity need to be undertaken. In addition, MR studies stratifying by menopausal status or smoking status can add additional insight in understanding the relationship between adiposity and breast or lung cancer risk.

CHAPTER 2

Validation of breast cancer risk prediction models in cohorts with long-term follow up

2.1 Introduction

Breast cancer is the most common cancer diagnosed in women in developed countries worldwide, with an estimate of over 252,710 new cases diagnosed and 40,610 deaths in United States in 2017(105). The five-year survival rate of breast cancer can be dramatically improved by almost 3.7 fold when comparing localized versus distant breast cancer(106), making early detection and effective screening of breast cancer especially important in clinical care(107, 108). More importantly, stratification of women according to the risk of developing breast cancer could improve risk reduction and screening strategies by targeting those most likely to benefit(109).

Both genetic and lifestyle factors are implicated in the aetiology of breast cancer. In the past decades, epidemiological research have identified many lifestyle and environmental risk factors of breast cancer, including menstrual and reproductive history, use of hormone therapy, anthropometry, and alcohol consumption etc(110). Although each risk factor explains a modest proportion of the variation in disease risk, when combined together, they could have a substantial effect on breast cancer risk, suggesting an important utilization in risk prediction(111).

The development of genome-wide association studies (GWAS) have led to the identification of common susceptibility loci of breast cancer marked by single nucleotide polymorphisms (SNPs)(18, 20). These SNPs have only a small effect size but cumulatively explain substantial variation in risk, implying potential utility for breast cancer risk prediction(39). In fact, several studies have reported modest utility of SNPs for improving the discriminatory accuracy of breast cancer risk prediction models(112-115). In addition, Mavaddat et al. found that a polygenic risk score (PRS) defined using common risk-susceptible SNPs can be useful for providing substantial breast cancer risk stratification(39).

In a recently published paper, Maas et al built a prediction model incorporating a PRS of 77 identified breast cancer SNPs and known breast cancer risk factors such as BMI and menopausal hormone therapy use(116). Evidence have also shown that models utilizing both life risk factor and a PRS defined by known SNPs can provide better risk stratification of breast cancer(117). However, very few validation studies have been carried out in independent populations to further assess the generalizability of these models.

External validation uses data on new participants, independent of the ones used for model development, to examine whether the model's prediction is reliable in individuals from populations similar to but distinct from those used to train the model. External validation is essential for any model to be broadly adopted. External validation usually uses two measures to evaluate model performance: discrimination and calibration. Discrimination indicates how well the prediction model distinguishes cases versus controls, while calibration tests for how the predicted probability of risk matches with the actual observed risk.

In addition, when validating 5-year absolute risk models, the accuracy of the estimate of discrimination and calibration is often limited by the number of cases diagnosed within the first five years from the baseline. For rare disease, this number can be small. Further, limiting the follow-up time to the first five years after baseline ignore the remaining follow-up, which can be substantial. For example, there is over 21 years of follow-up after baseline (blood draw) for the Nurses Health Study (NHS) and 15 years of follow-up for Nurses Health Study II (NHSII). Using the subsequent follow up by combining data from non-overlapping time windows can potentially increase validation sample size, enabling analysis within studies with limited number of outcome but a long follow up time.

In our analysis, we assessed and evaluated a synthetic breast cancer risk prediction model published by Montserrat Garcia-Closas et al. based on published estimates of risk parameters and a PRS using the most recently identified 313 breast cancer SNPs(20). To evaluate the model performance, we applied the model to data collected prospectively in both the NHS and NHSII using the Individualized Coherent Absolute Risk Estimator (iCARE) software. Three prediction models were assessed: one with only classical risk factor in the full cohort of Nurses; one with only PRS in the nested case-control study; and one with both classical risk factor and PRS in the nested case-control study of NHS and NHSII. We also describe a procedure that uses the subsequent follow up by combining data from non-overlapping time windows. We show by simulation that this procedure produces unbiased and more precise estimates of risk calibration and observed ten-year risks within expected risk categories.

2.2 Materials and methods

Study Population in the Validation cohort

The Nurses' Health Studies (NHS and NHSII) are prospective cohort studies of women with updated exposure assessment for a broad range of classic risk factors, endogenous hormones and DNA, in relation to risk of cancer. NHS has 121,700 women aged 30–55 enrolled in 1976 and NHSII has 116,000 women of 25–42 years of age enrolled in 1989. Both cohorts of women were asked to complete detailed questionnaires regarding their diet, classic epidemiological risk factors and disease outcome on bi-annual basis(118). Overall breast cancer cases (both in-situ and invasive) were identified either through self-report or by querying population based registries, followed by confirmation in the form of medical records and biopsy reports in Nurses Studies. Cases that occur within the first year of follow up must be excluded to remove any potentially prevalent cases from analyses.

Blood samples were collected from 32,826 cohort members in NHS in 1989 and from 29,240 women in NHS II in 1997. Cheek samples were also collected from 33,100 cohort members in NHS in 2002 and from 29,700 women in NHS II in 2005. We have GWAS data on a total of 18,531 women in NHS and 8285 women in NHS II—these data were generated as part of GWAS of 15 complex traits (including case-control studies of breast cancer, pancreatic cancer, ovarian cancer, colon cancer, endometrial cancer, cardiovascular disease, type 2 diabetes, gout, venous thromboembolism, PTSD etc.(119). They were genotyped using five GWAS arrays including Affymetrix 6.0, IlluminaArray, Illumina OmniExpress, HumanCore, and OncoArray. Imputation was performed using 1000 Genomes Project ALL Phase I Integrated Release Version 3 as the reference panel(119). The classic risk factor only prediction model was evaluated in both the full blood sub-cohort and the nested samples where genetic information is available for NHS and NHSII. Models with PRS and life risk factors + PRS were evaluated in the nested case-control group within the full cohort.

Polygenic Risk Score

The effects of cancer susceptibility variants on breast cancer were combined into a polygenic risk score (PRS). The PRS for any individual i was defined as the sum of the number of risk alleles across k variants weighted by the effect size of each variant:

$PRS_i=\beta_1 x_{1i} + \ldots + \beta_k x_{ki}$

where x_{ki} is the genotype of person i for variant k, expressed as the number of effect alleles (0, 1, or 2), and βk is the per-allele log risk ratio (odds ratio [OR] or hazard ratio [HR]) associated with the effect allele of SNP k. Using the largest published GWAS on breast cancer risk to date, the PRS used in this analysis was generated using 313 breast cancer risk-associated SNPs. (Supplementary Table 2.1)

Risk Prediction Model

In this analysis, we focused on a synthetic model established by Garcia-Closas et al. based on published estimates of risk parameters of breast cancer and the assumption of multiplicative gene-environment interaction(117). The model included a polygenic risk score (PRS), nonmodifiable risk factors other than the PRS (ie: family history, age at first birth, parity, age at menarche, height, menopausal status, and age at menopause), along with modifiable risk factors (ie: body mass index [BMI; calculated as weight in kilograms divided by height in meter squared], post-menopausal hormone (PMH) use, and level of alcohol consumption. The model is primarily for risk prediction of women with European ancestry. It used two sets of relative risk estimates from large published studies for women less than 50 years of age and 50 years of age or greater(117).This age stratification accounts for modification of the relative risks for BMI, family history, and benign breast cancer (BBD) , and the age-dependent distributions of several risk factors.

Validation Analysis and Simulation

Due to the small number of women who are <50 years old in NHS, we only performed validation analysis for women age ≥50 years old in NHS. In addition, since blood sample and cheek sample were collected at different times, the validation was first performed separately for blood samples taken in 1990 and for cheek sample taken in 2002; the results of these analyses were then meta-analyzed together for the final output in NHS (women ≥50 years old). Risk factor data were pulled from questionnaires corresponding to the time of DNA collection (1997 for blood and 2002 for cheek). Similarly, in NHS II, for women younger than 50 years old, analysis was first performed separately for blood samples taken in 1997 and for cheek samples taken in 2005; these results were then meta-analyzed. For women older than 50 years old in NHSII, the analysis was carried out using 2005 as the baseline for both blood and cheek samples in 2005, because of small number of women >= 50 at the blood collection in 1997. Table 2.1 A detailed summary of risk factor distributions for women older than 50 as well as younger than 50 in both NHS and NHSII.

Table 2.1: Risk factor distribution by validation cohorts

		NHS >=50	**NHS II >=50**	**NHS II <50**
Baseline Risk Factors(% in the full cohort)		N=58,163	N=36,323	N=37,847
Age at menopause				
	<40	4.2%	2.8%	--
	40-<45	6.7%	7.2%	--
	45-<50	25.0%	21.1%	--
	50-<55	56.6%	39.3%	--
	≥55	7.3%	29.5%	--
Age at menarche				
	≤10	6.4%	7.8%	6.9%
	11	16.2%	16.7%	16.0%
	12	26.2%	30.5%	30.3%
	13	30.8%	28.0%	27.8%
	14	12.4%	10.1%	10.8%
	15	4.5%	3.9%	4.5%
	≥16	3.5%	3.2%	3.6%
Parity				
	Nulliparous	5.6%	23.1%	23.4%
	1 birth	6.5%	17.9%	16.5%
	2 births	26.7%	50.2%	50.5%
	3+ births	61.2%	8.8%	9.6%
Age at first birth				
	<20	0.8%	8.5%	6.0%
	20-24	51.6%	28.1%	21.8%
	25-29	37.9%	38.1%	43.1%
	≥30	9.7%	25.3%	29.1%
BMI				
	<25	48.2%	43.2%	54.5%*
	25-<30	32.8%	29.8%	25.5%
	≥30	18.9%	27.0%	20.0%
Height (cm)				
	Mean (sd)	163.8 (0.0026)	163.5(0.0063)	165.1 (0.035)
Oral contraceptive use				
	Never	52.2%	11.5%	14.0%
	Ever	47.8%	88.5%	86.0%*

Table 2.1: Risk factor distribution by validation cohorts (CONTINUED)

Alcohol intake (g/day)				
	None	40.1%	34.9%	37.1%
	<5	35.1%	36.3%	41.8%
	5-14	16.4%	18.2%	16.0%
	15-24	5.7%	6.4%	3.0%
	25-34	1.7%	1.4%	1.2%
	35-44	0.0%	1.9%	0.5%
	≥45	0.0%	0.9%	0.4%
Hormone replacement therapy				
	Never	28.5%	34.8%	--
	Former	41.1%	29.4%	--
	Current	30.4%	35.8%	--
	Estrogen-only users	39.6%	40.5%	--
	Combined type users	60.4%	59.5%	--
Breast cancer family history				
	No	86.3%	89.7%	91.0%
	Yes	13.7%	10.3%	9.0%
History of benign breast disease				
	No	64.1%	53.1%	64.0%
	Yes	35.9%	46.9%	36.0%

*: slightly different definition in the women <50 prediction model

We used Individualized Coherent Absolute Risk Estimation (iCARE) R package developed by Chatterjee et al. to perform validation analyses(120). Conditional age-specific incidence rates given risk factors were assumed to follow a Cox proportional hazards model(121), and age was used as the timescale in the incidence modeling. The five-year absolute risk of breast cancer was estimated using the relative risk estimates from published literature, age-specific breast cancer rates from the US National Cancer Institute-Surveillance, Epidemiology, and End Results Program(NCI-SEER), and data on competing hazards for mortality available from the Center for Disease Control (CDC) WONDER database(122). Both discrimination and calibration analyses were performed using this R package.

We used inverse probability weighting (IPW) to account for the non-random sampling of participants with GWAS data when calculating observed 5-year incidence rates. Logistic regression models were used to estimate the probability of an individual being selected as a control, adjusting for age and follow-up time. The inverse of this probability was used as a weight to balance the contribution of controls so that the distribution matched the underlying cohort. The cases were assigned to a weight of 1 assuming that all cases will be included in the cohort.

To take advantage of the long follow up time in the NHS (22 years since blood draw), we also performed validation analyses among the participants in NHS blood cohort but at three different baselines: 1990, 1995 and 2000. At each of the baseline, only samples that were free of breast cancer were used. Covariates were updated at each baseline. We ran validation analysis at each baseline individually as well as in a combined synthetic cohort, where we concatenated the baseline cohorts creating a larger validation population.

The combined synthetic cohort approach assumes that breast cancer outcomes from any individual who contributes to multiple baseline cohorts are conditionally independent across baselines, and therefore the usual Fisher information estimates of the variance in observed incidence rates is valid. If the model is mis-specified, for instance, by excluding a risk factor that affects incidence in a non-collapsible fashion, then the usual variance estimates is no longer valid and can over or under-estimate the variance in the predicted incidence rates. To assess the impact of this model mis-specification on estimation of observed incidence rate (and the ratio of the observed versus expected events) and on tests of calibration, we simulated cohorts of 100,000 individuals assuming it will follow the true hazard for breast-cancer incidence

$$\lambda_0(t)e^{\beta_G G+\beta_X X}$$

, where G represents known, observed risk factors and X represents unknown, unobserved risk factors. The baseline incidence $\lambda_0(t)$ is modeled using the incidence rate at 30 and 70 years old non-Hispanic white women in US (27.3 and 451.5 per 100,000). All women entered the cohort at age 30 and were followed up for 40 years. To perform calibration analyses on these simulated cohorts, we calculated predicted 10-year rates using the assumed (mis-specified) hazard:

$$\lambda_0(t)e^{\beta_G G}$$

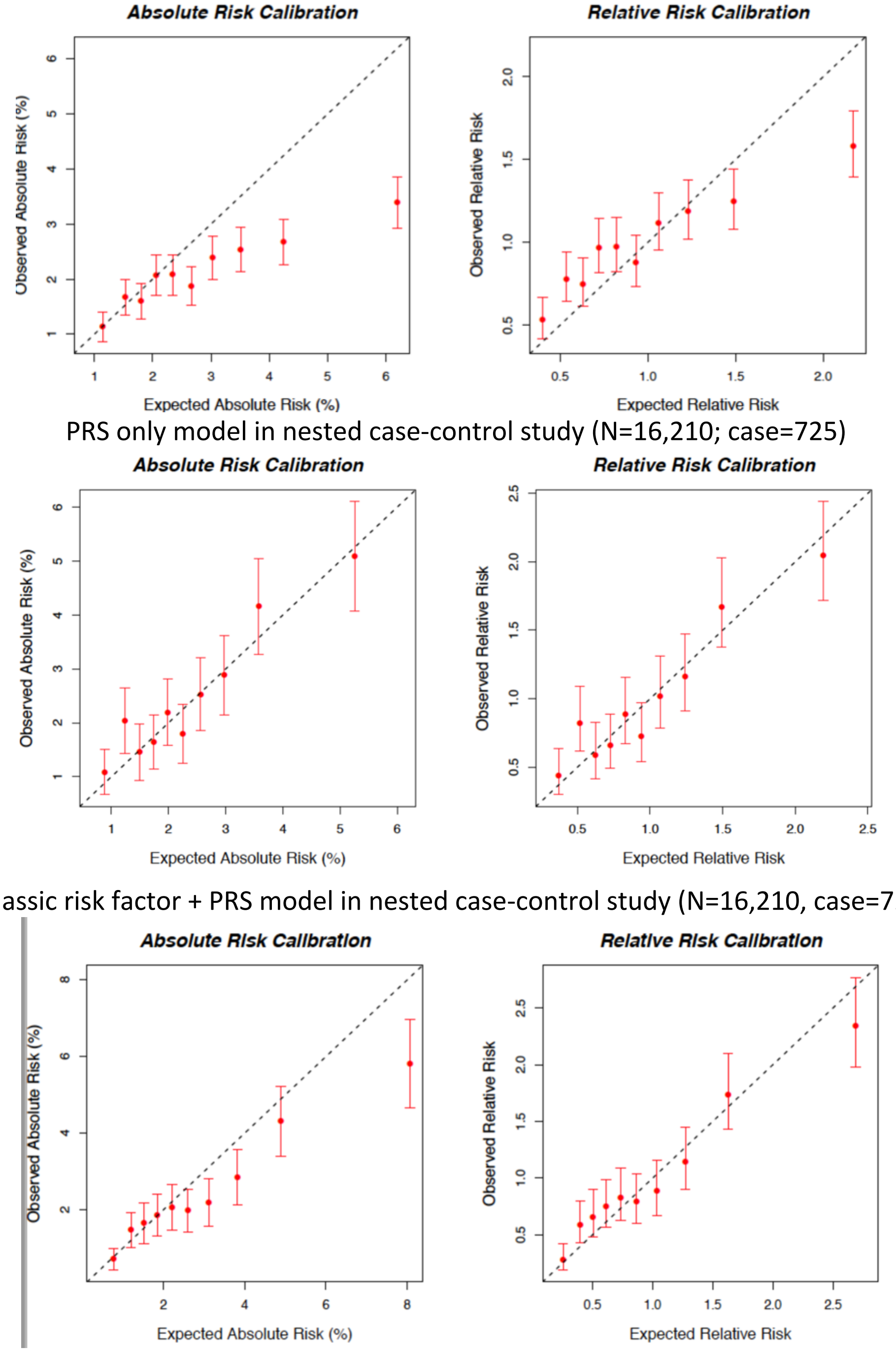
The observed 10-year incidence was calculated using the number of events divided by the number of at-risk people in that time interval in presence of G and X. We tested range of β_G from 1, 1.25, 1.5, 2, 2.5, and 3, and a range of β_X from 1, 1.25, and 1.5. For each combination of β_G and β_X, we calculated the observed and expected incidence of breast cancer within each of the risk deciles using six methods: just using age group 30-40, age group 40-50, age group 50-60, and age group 60-70, pooled analysis across the four age groups in the synthetic cohort, and meta-analysis of the four age groups' specific results. The observed rate was plotted against the expected rate across each of the decile for each method. We also calculated the Hosmer-Lemeshow goodness of fit tests for all six methods, using both model-based and robust variance estimates. To assemble impact of model misspecification in the model-based variance estimate, we also calculated the variance of the observed rate for each replication as well as the mean of the variance of the observed rate across all replications (Table 2.3).

2.3 Results

We evaluated the performance of three prediction models in both NHS and NHSII using the first five year of follow up time.

1. Classic risk factors only model [including age at menopause, age at menarche, partiy , age at first birth, BMI, heigh, oral contraceptive use, alcohol intake, hormone replacement therapy, family history of breast cancer and history of benign breast disease] predicting 5-year risk of breast cancer in women greater than 50 years old and younger than 50 years old respectively, in the full cohort (58,163 women >=50 in NHS, 36,323 women <50 in NHSII, and 37,847 women >=50 in NHSII)
2. PRS only model predicting 5-year risk of breast cancer in women greater than 50 years old and younger than 50 years old respectively, in the nested case-control study (16,210 women >=50 in NHS, 5,578 women <50 in NHSII, and 5,127 women >=50 in NHSII)
3. Classic risk factor and PRS model predicting 5-year risk of breast cancer in women greater than 50 years old and younger than 50 years old respectively, in the nested case-control study (16,210 women >=50 in NHS, 5,578 women <50 in NHSII, and 5,127 women >=50 in NHSII)

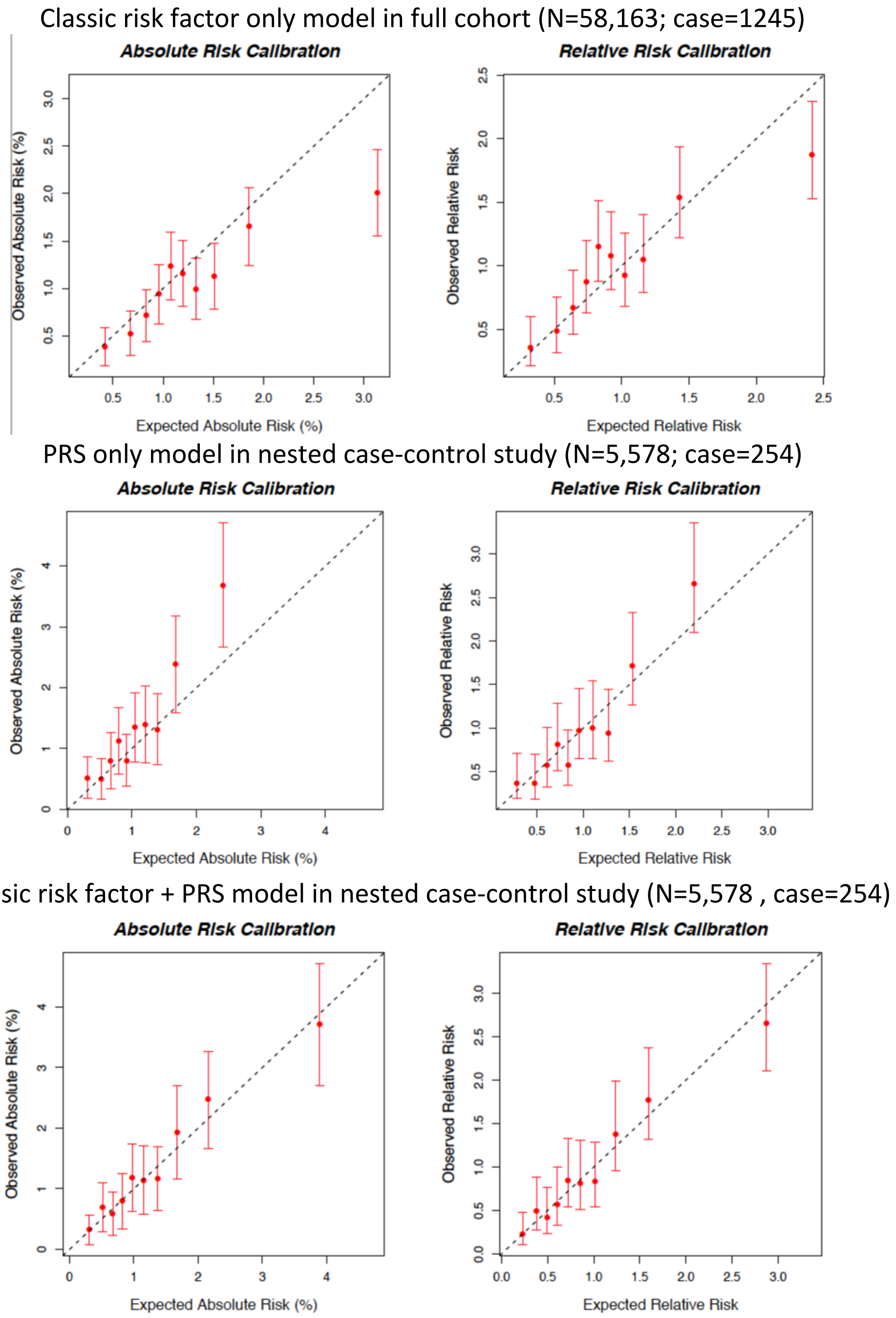
The relative risks were well calibrated for models incorporating PRS (models 2 and 3) among women older than 50 years old in NHS, for all three models among women older than 50 years in NHS II, and for all three models among women younger than 50 years in NHS II (Figure 2.1-2.3). In NHS, the absolute risk calibration of the classic risk factor only model showed over-estimation (observed risk > predicted risk) at the highest risk decile. In NHS II, the absolute risk calibration at the highest risk decile showed over-estimation for classic risk factor only model but under-estimation for PRS only model. Comparing across all three prediction, the model incorporating both classic life risk factors and PRS had the best calibration both on the relative and absolute risk scale in NHS and NHS II.


Classic risk factor only model in full cohort (N=58,163; case=1245)

PRS only model in nested case-control study (N=16,210; case=725)

Classic risk factor + PRS model in nested case-control study (N=16,210, case=725)

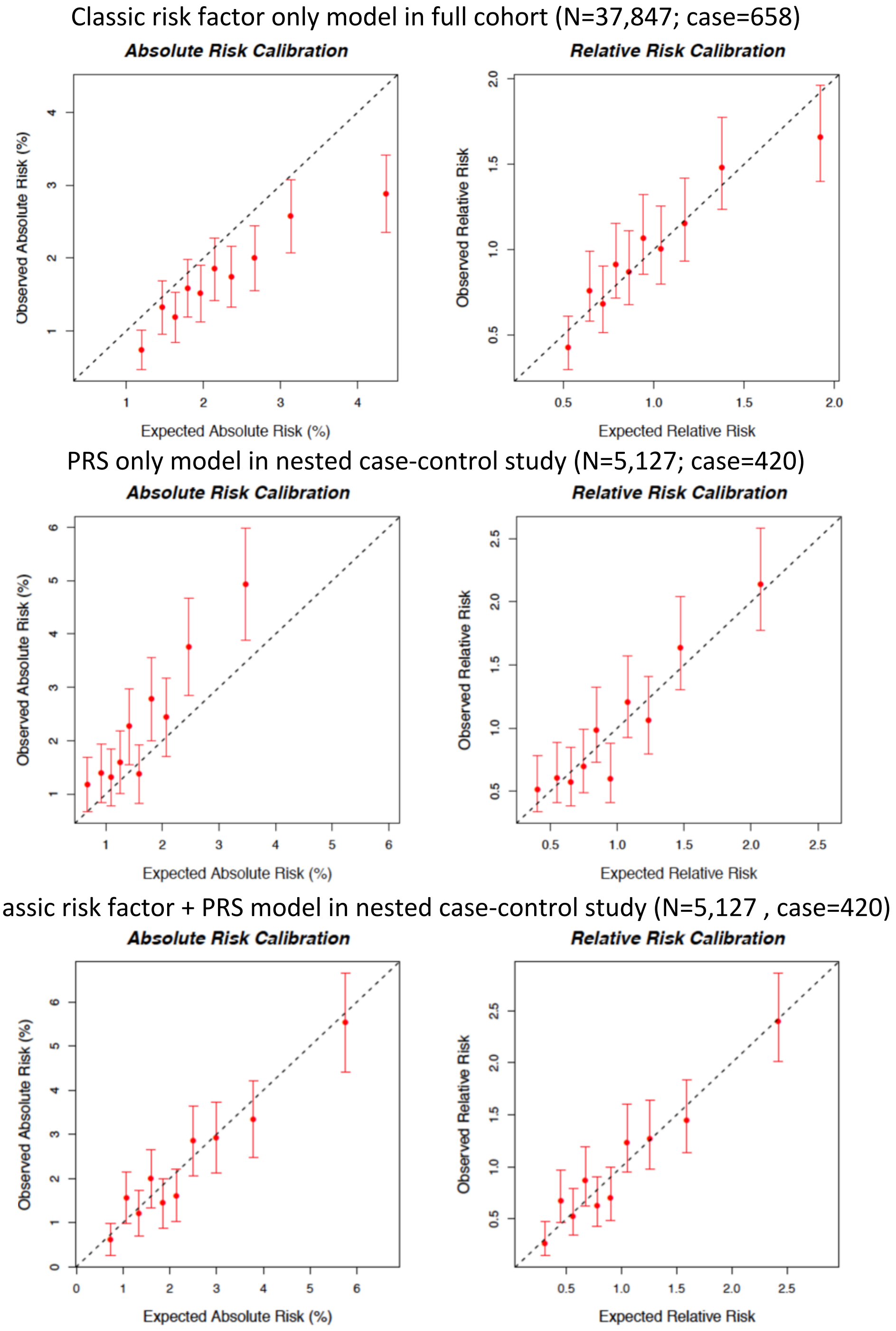
Figure 2.1: Validation output for 5-year risk prediction model in NHS, women >=50


Classic risk factor only model in full cohort (N=58,163; case=1245)

PRS only model in nested case-control study (N=5,578; case=254)

Classic risk factor + PRS model in nested case-control study (N=5,578 , case=254)

Figure 2.2: Validation output for 5-year risk prediction model in NHSII, women <50



Classic risk factor only model in full cohort (N=37,847; case=658)

PRS only model in nested case-control study (N=5,127; case=420)

Classic risk factor + PRS model in nested case-control study (N=5,127 , case=420)

Figure 2.3: Validation output for 5-year risk prediction model in NHSII, women >=50

For both age groups, adding the 313-SNP PRS to the classical risk factors substantially improved overall risk discrimination (Table 2.2). Age-adjusted AUC (95% CI) was 0.58 (0.56 to 0.60) for risk factor only model, 0.63 (0.60 to 0.65) for PRS only model, and 0.65 (0.63 to 0.68) for model with both life risk factor and PRS in NHS women ≥50 years old. Similar improvement was seen for NHS II women ≥50years old as well as women < 50 years old (Figure 2.2&2.3): AUC improved from 0.60 (0.58 to 0.62) to 0.65 (0.63-0.68) for women ≥50 years old and improved from 0.62 (0.59-0.65) to 0.69 (0.65 to 0.72) for women <50 years old.

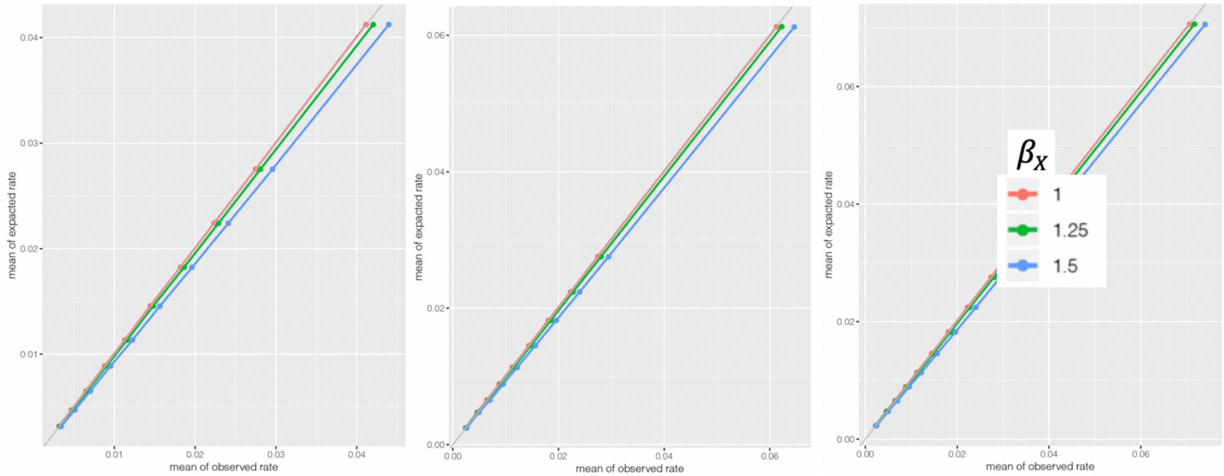
Table 2.2: AUC from all three prediction models among women older than 50 years old and among women younger than 50 years old in both NHS and NHS II.

	Women >50 in NHS	**Women < 50 in NHS II**	**Women > 50 in NHS**
Classic risk factor only model	0.58 (95%CI: 0.56, 0.60)	0.62 (95%CI: 0.59, 0.65)	0.60 (95%CI: 0.58, 0.62)
PRS only model	0.63 (95%CI: 0.60, 0.65)	0.67 (95%CI: 0.64, 0.71)	0.63 (95%CI: 0.60, 0.66)
Classic risk factor + PRS model	0.65 (95%CI: 0.63, 0.68)	0.69 (95%CI: 0.65, 0.72)	0.65 (95%CI: 0.63, 0.68)

We also performed validation analysis using data from three non-overlapping baseline time windows (1990, 1995, 2000) in NHS, both individually at each baseline and combining across three time windows. As shown in the Supplementary Figure 2.1, it was notable that by utilizing data from a baseline closer to the reference incidence year (such as 1995 and 2000), the calibration of absolute risk improved. In addition, by increasing our sample size via combining across three baselines (Supplementary Figure 2.1.d and 2.1.e), we achieved more precision in our estimation, indicated by a tighter confidence interval by nearly 50%.

In the simulation analysis across 1000 replicates, there was small differences at the tail between the mean of expected incidence rate (estimated only using β_G) and the mean observed incidence rate (estimated using both β_G and β_X) across all ranges of β_G from 1.25 to 3

(Figure 2.4). Such difference, although statistically significant in the goodness of fit test when β_X was 1.5, was very small in absolute value (Table 2.3), suggesting that our estimation of the variance of β_G, even in the presence of X, could still be valid.



β_G=1.5 β_G=2 β_G=3

Figure 2.4: Mean of the expected rate versus the mean of observed rate over 1000 replicates in the simulation study, across different range of β_G. β_G: the effect between Y and the measured exposure of interest G; β_X: the effect between Y and hypothetically unmeasured risk factor X. The G represents the known, observed risk factors and X represents unknown, unobserved risk factors.

Table 2.3: the variance of observed incidence rate for each replicate and the mean of variance across 1000 replicates for the 1st and 10th risk decile, and the rate of type I error from the goodness of fit tests in the simulation study from the meta-analysis of all four age groups' results

β_G	β_X	**Var(obs)_1st decile**	**E(Var(obs))_ 1st decile**	**Var(obs)_10th decile**	**E(Var(obs))_ 10th decile**	**Rate of Type I error from G.O.F. tests**
1.5000	**1.0000**	2.65E-04	2.61E-04	8.76E-04	9.08E-04	0.063
	1.2500	2.73E-04	2.64E-04	9.41E-04	9.18E-04	0.115
	1.5000	2.68E-04	2.72E-04	9.00E-04	9.40E-04	0.99
2.5000	**1.0000**	1.89E-04	1.82E-04	8.31E-04	8.75E-04	0.078
	1.2500	1.90E-04	1.84E-04	8.42E-04	8.82E-04	0.114
	1.5000	1.96E-04	1.90E-04	8.62E-04	9.01E-04	0.984
3.0000	**1.0000**	1.70E-04	1.64E-04	8.62E-04	8.93E-04	0.053
	1.2500	1.67E-04	1.66E-04	8.56E-04	9.00E-04	0.106
	1.5000	1.77E-04	1.71E-04	8.81E-04	9.16E-04	0.977

2.4 Discussion

Our study assessed the calibration of a breast cancer risk prediction models(20) incorporating questionnaire based factors and PRS in NHS and NHS II. We assessed the performance of models designed for women who are younger than 50 years old and women who are older than 50 years old. Among the three models we evaluated (classic risk factor only model, PRS only model, classic risk factor + PRS model), the integrated model with 313-SNP PRS and classical risk factors had the best performance. Specifically, the integrated models showed improvement on discrimination and good calibration especially on the relative risk scale.

There is some overprediction (predicted risk > observed risk) at the highest risk decile of the absolute risk calibration especially for classic risk factor only models among women ≥50 years old in NHS. Another validation study assessing the same model across a wide range of populations including studies in the U.S., UK, and Australia(123) also saw some miscalibration of absolute risk at the highest risk decile, although there was no systematic under- or over-prediction across different studies. This suggests that the slight miscalibration in the absolute risk scale are likely due to random variation or differences between study populations (e.g., wide range of study time periods or differences in risk factor distributions or disease rates), rather than a reflection of intrinsic model properties. The mis-calibration on the absolute risk scales may also due to the effect estimates used in these prediction models were drawn from a synthetic model which is not mutually adjusted. Future prediction models using more accurate and well-adjusted effect estimates of risk factors may further improve absolute risk calibration. Overall, the relative risk calibration was much better than the absolute risk calibration especially in NHS. This may due to the differences between the validation population of NHS

(1990) and the reference population taken from SEER 2008-2012. This highlights the importance of absolute risk validation across multiple study populations, particularly using cohorts similar to the target populations, both in chronologic years of study and underlying risk(123).

We explored the potential of using samples from cohorts with long follow-up but at three non-overlapping baselines. Comparing the validation results at different baseline time point (Supplementary Figure 2.1), the model performance did change slightly over time. Specifically, the calibration of absolute risk became better when the baseline population was closer to the reference population. In addition, by combining the three baseline cohorts, the model precision improved greatly as shown by the narrower confidence interval. This suggests that it is important to be aware of the time frame we used for validation studies and that the difference between the validation cohort baseline and the reference population baseline can yield different model performances. It is also possible that the small difference in model performance over time is due to random variation. In that case, researchers can take advantage of long follow-up time and combine cohorts from non-overlapping baselines together, which can greatly improve precision.

Such combination across different baseline can produce valid results as long as our model is not mis-specified. To test that, we ran simulation analysis to examine 1) how different the estimation of the outcome incidence rate can be in the presence of X (the unobserved and unknown risk factors)? 2) how accurate our variance estimation of the effect size can be in presence of X?. As shown in figure 2.4, the expected rate estimated using G (the known, observed risk factors) was very similar to the observed rate using both G and X. There was slight deviation from the diagonal line as the effect size of X increases but such differences were expected due to non-collapsibility. In addition, the small difference between the variance of the observed rate and the mean of the variances of observed rate across 1000 replications provide some confidence in our method combining cohort across baselines (Table 2.3).

Our study does have some limitations. First, our model was designed for testing only among women with European ancestry. We cannot infer the utility of the prediction model in population other than ones of European ancestry and alternative models have only been evaluated in relatively small studies(124-126). In a recent work done by our group, we evaluated the same synthetic prediction model in a Korean population(127). We found significant overestimation of risk for women older than 50 years old and that recalibrating the model using Korean incidence rate, mortality rates and risk factor distributions could improve model performance(127). We expect that incorporating RRs from large population-based studies in Korea (rather than from studies of European ancestry) can further improve model performance. Secondly, our risk models do not adequately capture risk for women with strong family history or carrying high-risk variant in breast cancer predisposition genes. Future studies could integrate with family-based models, such as BOADICEA model(128) as well as effect estimates of rare mutation in BRCA1/BRCA2. Other risk factors such as mammographic density should also be considered. Thirdly, although iCARE can be used for risk predictions over any time period, our current study only evaluated the five-year risk prediction, and further work is needed to evaluate longer-term predictions used by some clinical guidelines. Finally, we only tested the risk of overall breast cancer (both invasive and in situ) rather than subtype specific cancer (ie: estrogen receptor positive and negative cancer). It has been shown from past literature(129-131) that subtype specific tumors have very different risk profile and prognosis and hence future work on subtype-specific breast cancer is needed to obtain more precise screening and risk modeling strategies.

In summary, we presented extensive validation of a breast cancer risk prediction model integrating both classical risk factors and genetic factors in form of PRS. We showed that adding PRS can substantially improve model performance which can be useful in future risk-stratified prevention and screening strategies. We also demonstrated that when model performance does not vary much over time, we can take advantage of long follow-up time by concatenating cohorts from non-overlapping baselines to boost precision.

CHAPTER 3

The combined effect of polygenic risk score and pathogenic mutations in breast cancer predisposition genes in the general population

3.1 Introduction

Breast cancer is the most common cancer among women in the United States(132). Primary prevention such as tamoxifen can greatly reduce breast cancer risk, but also has side effects such as hot flashes, blood clots and uterine cancer(133). Early detection of breast cancer with screening can help detect cancer early and hence improve survival rates, but it may also result in overdiagnosis, over treatment, and increased medical cost(134). Hence, it is particularly important to effectively stratify women according to their risk of developing breast cancer and provide a more personalized approach to identify women most likely to benefit from these prevention and screening strategies(107, 108) and the best timing for these interventions. For instance, women who are at particularly high risk at young ages may initiate MRI screening at an earlier age.

Pathogenic variants detected in multi-gene cancer predisposition panels are increasingly used to counsel women regarding their risk for breast cancer. In the past few decades, germline genetic testing has evolved substantially due to advances in genetic sequencing techniques and bioinformatics, enabling rapid and efficient detection of genetic variation(135). However, our understanding of how to transform the genetic information of a woman into actionable clinical recommendation still needs improvement. Pathogenic variants in high penetrance genes such as *BRCA1* and *BRCA2* are well studied but the clinical implications of variants in moderate penetrance genes (e.g. *CHEK2*, *ATM*) remain unclear.

Common variants (SNPs) found through genome-wide association studies (GWAS) have also shown to be associated with elevated breast cancer risks(38). The risk conferred by each individual SNP is small and not useful in risk prediction, however the combined effect of multiple SNPs in the form of a polygenic risk score (PRS) can achieve substantial effects(107-109, 136). The most recent PRS study by Mavaddat et al. found that a one standard deviation change in PRS increases the odds of breast cancer risk by 61% (OR=1.61, 95%CI: 1.57-1.65), and the lifetime risk of overall breast cancer in the highest percentile of PRS was 32.6%.

With information available for both the pathogenic variants in breast cancer predisposition genes and the common variants as PRS, a key question to understand is how pathogenetic variant and PRS interact: will the effect of PRS be different among carriers of pathogenic variants versus non-carriers? Can PRS further stratify risk of breast cancer among carriers of pathogenic variants? Previous work found that PRS modified breast cancer risk in women with pathogenic variant in *BRCA1* or *BRCA2*(137), but the joint effects of pathogenic variants and PRS have not been studied in samples drawn from the general population. There is also no published study evaluating the effect of PRS among women with pathogenic variants in genes other than *BRCA1/2*(137) and *CHEK2(138)*.

In this study, we evaluated the combined effect of polygenic risk score and pathogenic variants in nine established breast cancer predisposition genes in the general population using 26,798 cases and 26,127 controls from 12 population based case-control studies. We evaluated the performance of an overall breast cancer PRS as well as an ER negative specific PRS. We also estimated 5-year and lifetime absolute risk of developing breast cancer across percentiles of PRS for carriers of pathogenic variants as well as non-carriers.

3.2 Materials and methods

Study Population

The study consists of subject from nine cohorts and three population-based case-control studies in the CAnceR RIsk Estimates Related to Susceptibility" (CARRIERS) consortium. The nine cohort studies are Cancer Prevention Study II (CPSII)(139), Cancer Prevention Study 3 (CPS3)(140), California Teachers Study (CTS)(141), Multiethnic Cohort (MEC)(142), Mayo Mammography Health Study (MMHS)(143), Nurses Health Study (NHS)(144), Nurses Health Study II (NHSII)(145), Women's Health Initiative (WHI)(146), and the SISTER study(147). The three population-based case-controls studies are the Women's Circle of Health Study (WCHS)(148), Mayo Clinical Breast Cancer Study (MCBCS)(149), and Wisconsin Women's Health Study (WWHS)(150). Cases were identified via self-report and confirmed by reviews of medical records or were identified through registry linkage. Controls from the CPSII, CTS, MEC, MCBCS, NHS, NHSII, WCHS, WHI, and WWHS were matched to cases by age. CPSIII, SISTER and MMHS utilized a case-cohort design, where the controls were breast-cancer-free members of reference sub-cohort.

In total, we analyzed 52,925 non-Hispanic European-ancestry individuals (26,127 controls and 26,798 cases). The number of cases and controls with respect to five age groups (age≤40, 40-50, 50-60, 60-70, >70) and family history status of 1st degree relatives including mom, sisters, and dad (yes or no) for each individual study is shown in Table 3.1.

Sequencing of rare variant in 9 cancer predisposition genes

Genomic DNA samples were subjected to multiplex amplicon-based analysis of 746 target regions covering all coding regions and consensus splice sites from 37 cancer predisposition genes using a QIAseq (QIAGEN) custom panel(151). The QIAseq protocol was optimized for high-throughput robotic processing of DNA samples and validated as previously described(152). Libraries were individually bar-coded by dual indexing and sequenced in pools of 768 on a HiSeq4000. Median sequence read depth was about 200X.

Nine genes were evaluated in this study: *ATM, BARD1, BRCA1, BRCA2, BRIP1, CDH1, xsCHEK2, NF1,* and *PALB2*. These genes were selected because of their common inclusion on clinical hereditary cancer genetic testing panels and because of previous reports suggesting associations with breast cancers(153-156). In addition, genotypes on 138 common variants were generated by sequencing the regions flanking these variants. The common variants included the 77 SNPs in previously published PRS by Mavadatt et al. (or proxies in high linkage disequilibrium with these SNPs) as well as other SNPs that were found to be associated with breast cancer or breast cancer subtypes in subsequent GWAS and fine-mapping studies. [make sure to reference to the main analysis paper once that is done by Fergus]

Table 3.1: Case control number by age group, and by family history status(1st degree relative) in CARRIERS consortium

				Age group					Family History of BC	
Study	**Total #**	**Case control status**		**<=40**	**41-50**	**51-60**	**61-70**	**>70**	**No**	**Yes**
CPS 3	2739	**Control**	1397	362	413	482	140	0	1207	190
			51.0%	25.9%	29.6%	34.4%	10.0%	0.0%	86.4%	13.6%
		Case	1342	45	294	782	218	3	1004	338
			49.0%	3.4%	22.0%	58.2%	16.2%	0.2%	74.8%	25.2%
CPS II	7762	**Control**	3843	0	9	338	1559	1937	3291	552
			49.5%	0.0%	0.2%	8.8%	40.6%	50.4%	85.6%	14.5%
		Case	3919	0	10	343	1592	1974	3159	760
			50.5%	0.0%	0.3%	8.8%	40.6%	50.4%	85.6%	14.5%
CTS	3910	**Control**	1917	25	175	586	713	418	1669	248
			49.6%	1.3%	9.2%	30.6%	37.2%	21.8%	87.1%	12.9%
		Case	1992	20	204	598	706	464	1649	343
			50.5%	1.0%	10.2%	30.0%	35.4%	23.3%	82.8%	17.2%
MCBCS	6926	**Control**	3152	181	585	899	849	638	2522	630
			45.5%	5.7%	18.6%	28.5%	26.9%	20.2%	80.0%	20.0%
		Case	3774	251	810	1079	1004	630	2860	914
			54.5%	6.7%	21.5%	28.6%	26.6%	16.7%	75.8%	24.2%
MEC	1772	**Control**	893	0	14	188	363	328	798	95
			50.4%	0.0%	1.6%	21.1%	40.6%	36.7%	89.4%	10.6%
		Case	879	0	25	192	325	337	729	150
			49.6%	0.0%	2.8%	21.8%	37.0%	38.3%	82.9%	17.1%
MMHS	1395	**Control**	1131	67	358	275	254	177	941	190
			81.1%	5.9%	31.7%	24.3%	22.5%	15.6%	83.2%	16.8%
		Case	264	0	24	50	80	110	192	72
			18.9%	0.0%	9.1%	18.9%	30.3%	41.7%	72.7%	27.3%
NHS	4285	**Control**	2303	0	45	380	983	895	1953	350
			53.7%	0.0%	2.0%	16.5%	42.7%	38.9%	84.8%	15.2%
		Case	1982	0	50	353	825	754	1525	457
			46.3%	0.0%	2.5%	17.8%	41.6%	38.0%	76.9%	23.1%

Table 3.1: Case control number by age group, and by family history status(1st degree relative) in CARRIERS consortium (CONTINUED)

NHS II	2239	**Control**	1355	26	512	750	67	0	1177	178
			60.5%	1.9%	37.8%	55.4%	4.9%	0.0%	86.9%	13.1%
		Case	884	16	303	497	68	0	685	199
			39.5%	1.8%	34.3%	56.2%	7.7%	0.0%	77.5%	22.5%
SISTER	3599	**Control**	1561	64	369	610	432	86	0	1561
			43.4%	4.1%	23.6%	39.1%	27.7%	5.5%	0.0%	100.0%
		Case	2038	9	279	659	730	361	0	2038
			56.6%	0.4%	13.7%	32.3%	35.8%	17.7%	0.0%	100.0%
WCHS	1120	**Control**	571	105	172	226	68	0	467	104
			51.0%	18.4%	30.1%	39.6%	11.9%	0.0%	81.8%	18.2%
		Case	549	56	171	195	105	22	415	134
			49.0%	10.2%	31.1%	35.5%	19.1%	4.0%	75.6%	24.4%
WHI	9529	**Control**	4535	0	5	591	1888	2051	3819	716
			47.6%	0.0%	0.1%	13.0%	41.6%	45.2%	84.2%	15.8%
		Case	4994	0	6	710	2138	2140	3958	1036
			52.4%	0.0%	0.1%	14.2%	42.8%	42.9%	79.3%	20.7%
WWHS	7650	**Control**	3469	194	815	1297	1163	0	2947	522
			45.3%	5.6%	23.5%	37.4%	33.5%	0.0%	85.0%	15.0%
		Case	4181	238	1056	1569	1233	85	3267	914
			54.7%	5.7%	25.3%	37.5%	29.5%	2.0%	78.1%	21.9%
Total	52925	**Control**	26127	1024	3474	6621	8479	6530	20791	5336
			49.4%	3.9%	13.3%	25.3%	32.5%	25.0%	79.6%	20.4%
		Case	26798	635	3233	7026	9024	6880	19423	7375
			50.6%	2.4%	12.1%	26.2%	33.7%	25.7%	72.6%	27.4%

Polygenic Risk Score (PRS)

We filtered the list of 138 SNPs by linkage disequilibrium in a stepwise fashion, firstly removing all SNPs with $r^2>0.2$ with the smallest p-value (based on the largest published GWAS of overall breast cancer(20)), then removing all SNPs with $r^2>0.2$ with the second-most significant remaining SNP, and so on. A total of 105 independent ($r^2<0.2$) common variants were used to construct the final polygenic risk score (PRS) (Supplemental Table 3.1). For For any individual i, the PRS was calculated as the sum of the number of risk alleles across 105 variants weighted by the effect size of each variant:

$$PRS_i=\beta_1 x_{1i} + \ldots + \beta_k x_{ki.}$$

where x_{ki} is the genotype of person i of variant k, encoded as the number of effect alleles (0, 1, or 2), and βk is the per-allele log risk ratio associated with the effect allele of SNP k. The primary overall breast cancer PRS used in this analysis used effect estimates from the largest published breast cancer GWAS(20) (Supplementary Table 3.1). To construct ER negative specific PRS, we used a hybrid method to obtain the effect size, in which ER- effect sizes of the SNPs were used if the p-value from the heterogeneity test (ER positive versus ER negative disease) was <0.05, and effect sizes of overall breast cancer were used otherwise. Both the overall breast cancer PRS and ER- specific PRS were standardized to a mean of 0 and standard deviation of 1.

Model Fitting

We fitted a baseline model using logistic regression, with overall breast cancer (including both invasive and in-situ) as the outcome and the following explanatory variables: nine indicator variables denoting carriers status of pathogenic variant for each of the breast cancer predisposition genes, PRS as a continuous variable, age in five categories (age <=40, 41-

50, 51-60, 61-70, >70) and an indicator variable for family history of breast cancer. Age was defined as the age of diagnosis for cases and age of baseline/age of matching date for controls. Family history of breast cancer was defined as the family history of 1st degree relative including mother, sisters, daughters and father. Missing values in age (0.9% missing) and family history (3.3% missing) were replaced using conditional draw imputation as implemented in the MICE R package(157). Because of well-established modification of *BRCA1*, *BRCA2*(137) and PRS(42) effects by age, we included product interaction terms between ordinally coded age categories and carriers status of *BRCA1* and *BRCA2*, and PRS. In addition to the mutually adjusted baseline model, we evaluated the PRS effect modification on pathogenic variant in each individual gene in a simple logistic regression without adjusting for variants in other genes. We also tested whether the effect of PRS differ comparing non-carriers versus carriers of pathogenic variants in any of the nine genes(pvalue<0.005 after Bonferroni correction of multiple testing).

To assess whether the discriminating ability of our model improved by allowing the effect of the PRS to change by pathogenic variant status, age and family history, we performed L_1 penalized logistic regression using the glmnet R package(158). All covariates in the baseline model were pre-selected for inclusion. Additional covariates included all the other possible interactions between variant in predisposition genes and age, variant in predisposition genes and family history, variant in individual predisposition gene and PRS, PRS and family history, any variant in any of the predisposition genes and PRS, any variant in any of the predisposition genes and family history, any variant in any of the predisposition genes and age. The final model was chosen by 10-fold cross validation maximizing the AUC as a function of the L1 penalty. An ER-negative specific PRS was used to model ER negative breast cancer.

Absolute Risk Estimation

Using the log odds ratios from the final model and external estimates of breast cancer incidence and competing mortality, we estimated 5-year and lifetime absolute risk of developing breast cancer (both invasive and in-situ). The 5-year and lifetime absolute risk of a woman starting at age *a* was estimated using the following formula(159):

$$\int_{a}^{a+\tau} \lambda_0(t)\exp(\beta' Z)\exp\left(-\int_{a}^{t}[\lambda_0(u)\exp(\beta' Z)+m(u)]du\right)dt$$

Here τ represents the time window of interest; Z represents the risk factors of breast cancer; β represents the relative risk parameters; $\lambda_0(t)$ is the baseline hazard function and m(t) is the age-specific morality rate. The marginal age-specific disease incidence was obtained from the SEER registry 2008-2012, and the competing mortality rate was obtained from CDC WONDER database 2008-2012.

3.3 Results

The best fitting risk model included pathogenic variant status for nine genes (*BRCA1, BRCA2, ATM, CHEK2, BARD1, BRIP1, CDH1, NF1, PALB2*), PRS, and age interaction for *BRCA1*, *BRCA2* and PRS, adjusted for study, age and family history but did not include any PRS-by-pathogenic variant interaction terms. Thus, in our final model, the relative risk gradient associated with per unit change in PRS among carriers of pathogenic variants was similar to that among non-carriers. Holding every other covariates constant, a one standard deviation change in the PRS was associated with 1.61x (95%CI: 1.54, 1.70) change in the odds of overall breast cancer for women who are younger than 40 years old (Table 3.2).

Table 3.2: Adjusted OR and its 95%CI of overall and ER negative breast cancer at different age group. Overall breast cancer PRS is used for overall breast cancer ER- specific PRS is used for ER- breast cancer. The OR is calculated from our best fitting model, adjusting for the following the explanatory variables: nine indicator variables denoting carriers status of pathogenic variant for each of the breast cancer predisposition genes, PRS as a continuous variable, age in five categories and an indicator variable for 1st degree family history.

OR (95%CI)	**Overall Breast Cancer**					**ER – Breast Cancer**				
	<=40	**41-50**	**51-60**	**61-70**	**>70**	**<=40**	**41-50**	**51-60**	**61-70**	**>70**
PRS*	1.62 (1.54, 1.70)	1.58 (1.52, 1.63)	1.54 (1.50, 1.57)	1.50 (1.47, 1.53)	1.46 (1.42, 1.51)	1.47 (1.15, 1.87)	1.47 (1.15, 1.86)	1.46 (1.15, 1.86)	1.46 (1.16, 1.85)	1.46 (1.16, 1.84)
BRCA1*	15.19 (8.29, 29.39)	9.31 (6.13, 14.1)	5.70 (4.14, 7.86)	3.49 (2.28, 5.36)	2.14 (1.12, 4.07)	54.92 (25.6, 125.1)	33.19 (19.8, 55.7)	20.06 (13.7, 29.5)	12.12 (7.24, 20.3)	7.33 (3.33, 16.1)
BRCA2*	16.47 (9.15, 31.12)	10.91 (7.19, 16.5)	7.22 (5.44, 9.59)	4.78 (3.52, 6.50)	3.17 (1.99, 5.03)	12.75 (5.29, 30.6)	11.91 (6.54, 21.7)	11.13 (7.45,16.64)	10.40 (6.92, 15.6)	9.72 (5.28, 17.9)
ATM	1.87 (1.56, 2.49)	1.87 (1.56, 2.49)	1.87 (1.56, 2.49)	1.87 (1.56, 2.49)	1.87 (1.56, 2.49)	1.20 (0.62, 2.13)	1.20 (0.62, 2.13)	1.20 (0.62, 2.13)	1.20 (0.62, 2.13)	1.20 (0.62, 2.13)
CHEK2	2.37 (1.95, 2.90)	2.37 (1.95, 2.90)	2.37 (1.95, 2.90)	2.37 (1.95, 2.90)	2.37 (1.95, 2.90)	1.19 (0.70, 1.92)	1.19 (0.70, 1.92)	1.19 (0.70, 1.92)	1.19 (0.70, 1.92)	1.19 (0.70, 1.92)
PALB2	3.49 (2.40, 5.21)	3.49 (2.40, 5.21)	3.49 (2.40, 5.21)	3.49 (2.40, 5.21)	3.49 (2.40, 5.21)	7.82 (4.39,13.80)	7.82 (4.39, 13.8)	7.82 (4.39,13.80)	7.82 (4.39, 13.8)	7.82 (4.39, 13.8)
BARD1	1.59 (0.98, 2.59)	1.59 (0.98, 2.59)	1.59 (0.98, 2.59)	1.59 (0.98, 2.59)	1.59 (0.98, 2.59)	2.93 (1.24, 6.34)	2.93 (1.24, 6.34)	2.93 (1.24, 6.34)	2.93 (1.24, 6.34)	2.93 (1.24, 6.34)
BRIP1	1.47 (0.99, 2.21)	1.47 (0.99, 2.21)	1.47 (0.99, 2.21)	1.47 (0.99, 2.21)	1.47 (0.99, 2.21)	1.61 (0.63, 3.54)	1.61 (0.63, 3.54)	1.61 (0.63, 3.54)	1.61 (0.63, 3.54)	1.61 (0.63, 3.54)
CDH1	5.83 (1.84, 25.8)	5.83 (1.84, 25.8)	5.83 (1.84, 25.8)	5.83 (1.84, 25.8)	5.83 (1.84, 25.8)	5.71 (0.26, 60.8)	5.71 (0.26, 60.8)	5.71 (0.26, 60.8)	5.71 (0.26, 60.8)	5.71 (0.26, 60.8)
NF1	1.96 (0.82, 5.10)	1.96 (0.82, 5.10)	1.96 (0.82, 5.10)	1.96 (0.82, 5.10)	1.96 (0.82, 5.10)	0.83 (0.13, 2.89)	0.83 (0.13, 2.89)	0.83 (0.13, 2.89)	0.83 (0.13, 2.89)	0.83 (0.13, 2.89)

*: model includes ordinal age interaction

PRS-by-pathogenic variant interactions for each individual gene were not statistically significant(Supplemental Table 3.7), confirming our results from the best fitting, mutually adjusted model above. But the effect of PRS by pathogenic variant in any of the nine genes was statistically significant (p=0.0002). The results by forcing a PRS-by-pathogenic variant in any genes interaction in our final model can be found in Supplementary Table 3.8.

The effect of PRS on breast cancer risk decreased with age: the OR of overall breast cancer per standard deviation change in PRS decreased from 1.61 (95%CI: 1.54, 1.70) among women <= 40 years old to 1.46 (95%CI: 1.42, 1.51) among women who were older than 70 years old. The OR of overall breast cancer for each age group with respect to their PRS (10th percentile, median, 90th percentile) and variant carrier status could be found in Supplementary Table 3.2. Comparing the 90th percentile of PRS to the 10th percentile of PRS, the OR of breast cancer was 3.41, 3.19, 3.00, 2.81, 2.63-folds increase for women who are <= 40 years old, 41-50 years old, 51-60 years old, 61-70 years old, and >70 years old respectively (Supplementary Table 3.2).

We also examined the association between ER negative disease and ER negative specific PRS and the overall breast cancer PRS. The OR of ER negative breast cancer was 1.47 (95%CI: 1.15, 1.86) for one standard deviation(s.d.) change in the ER negative PRS for women <= 40 years old. By comparison, the OR for ER negative breast cancer for one s.d. change in overall breast cancer was 1.20 (95%CI: 1.07, 1.35) (Supplementary Table 3.4). The strength of the association between the ER negative PRS and ER negative breast cancer declined with age, but the magnitude of such change was small.

The estimated lifetime absolute risk by age 80 years in the 10th, 50th, and 90th percentile of overall breast cancer PRS by variant carrier status is shown in Table 3.3. As expected, pathogenic variants in breast cancer predisposition genes greatly increase the lifetime risk of breast cancer. Carriers of pathogenic variants in high penetrance genes like *BRCA1* and *BRCA2*, had much higher lifetime risk than carriers of variant in moderate penetrance genes such as *CHEK2* and *ATM*. The lifetime risk of breast cancer of non-carriers in the 10th and 90th percentile of PRS were 6.7% and 18.2% for women without family history and 9.1% and 23.9% for women with 1st degree family history of breast cancer. Going from the 10th percentile to the 90th percentile of PRS, the estimated lifetime risk of women without family history of breast cancer ranged from 12.8% to 32.1% for *ATM* carriers, 15.2% to 37.3% for *CHEK2* carriers, and 21.4% to 48.9% for *PALB2* carriers, 10.5% to 27.1% for BARD1 carriers, 9.8% to 25.4% for BRIP1 carriers, 23.8% to 32.1% for NF1 carriers, suggesting that PRS significantly help to describe the breast cancer risk gradient among carriers of pathogenic variant in moderate penetrance genes.

Table 3.3: Lifetime absolute risk for different mutation carriers with respect to different PRS percentile and family history status.

	No Family History			Family History		
OR	**10th% PRS**	**median PRS**	**90th% PRS**	**10th% PRS**	**median PRS**	**90th% PRS**
non-carrier	0.067	0.111	0.182	0.091	0.148	0.239
ATM carrier	0.128	0.205	0.323	0.170	0.268	0.409
CHEK2 carrier	0.153	0.241	0.373	0.201	0.312	0.467
PALB2 carrier	0.215	0.332	0.491	0.280	0.419	0.593
BRCA1 carrier	0.288	0.438	0.627	0.369	0.540	0.729
BRCA2 carrier	0.348	0.514	0.703	0.439	0.619	0.794
BARD1 carrier	0.105	0.170	0.271	0.140	0.223	0.348
BRIP1 carrier	0.098	0.158	0.254	0.131	0.209	0.327
CDH1 carrier	0.329	0.482	0.660	0.416	0.584	0.754
NF1 carrier	0.128	0.204	0.321	0.170	0.266	0.407

Table 3.4: The % of total population identified by PRS for lifetime risk of breast cancer >20%, given their variant status and family history

Lifetime risk (to age 80) of BC	**No Family History**	**Family History**
non-carrier	6.06%	21.2%
ATM carrier	52.5%	79.8%
CHEK2 carrier	69.7%	89.9%
PALB2 carrier	92.9%	98.0%
BRCA1 carrier	98.0%	99.0%
BRCA2 carrier	99.0%	99.0%
BARD1 carrier	32.2%	61.6%
BRIP1 carrier	26.3%	54.5%
CDH1 carrier	98.9%	98.9%
NF1 carrier	51.5%	78.8%

The US National Comprehensive Cancer Network (NCCN)(160, 161) recommends beginning MRI screening for women with a lifetime risk greater than 20%. Table 3.4 shows the percentage of women who have greater than 20% of lifetime risk based on their PRS, stratified by carrier status and family history of breast cancer. Most (>90%) carriers of pathogenic variant in *BRCA1, BRCA2,* and *PALB2* have >20% lifetime risk. However, for *ATM* and *CHEK2* carriers, only 52.5% and 69.7% are above the threshold without a first degree relative family history of breast cancer and 79.8% and 89.9% with a family history. This suggests that even if a woman is a carrier of pathogenic variant in *ATM* or *CHEK2*, her lifetime risk may be below the 20% lifetime risk threshold and thus may potentially avoid additional intervention at her early ages, depending on her PRS.

We also estimated 5-year absolute risk of developing breast cancer across different percentile of PRS for women at age 40 and age 60 respectively (Figure 3.1). For 40 years old women, the estimated 5-year risk of breast cancer for *BRCA1* or *BRCA2* carriers were significantly larger than that of *CHEK2/ATM/PALB2* carriers and noncarriers regardless of their family history status.-Of note, many women with pathogenic variant in CHEK2 and ATM, particularly those in the lowest 50% of PRS with no first degree relative of breast cancer, have a low 5 year risk at age 40.

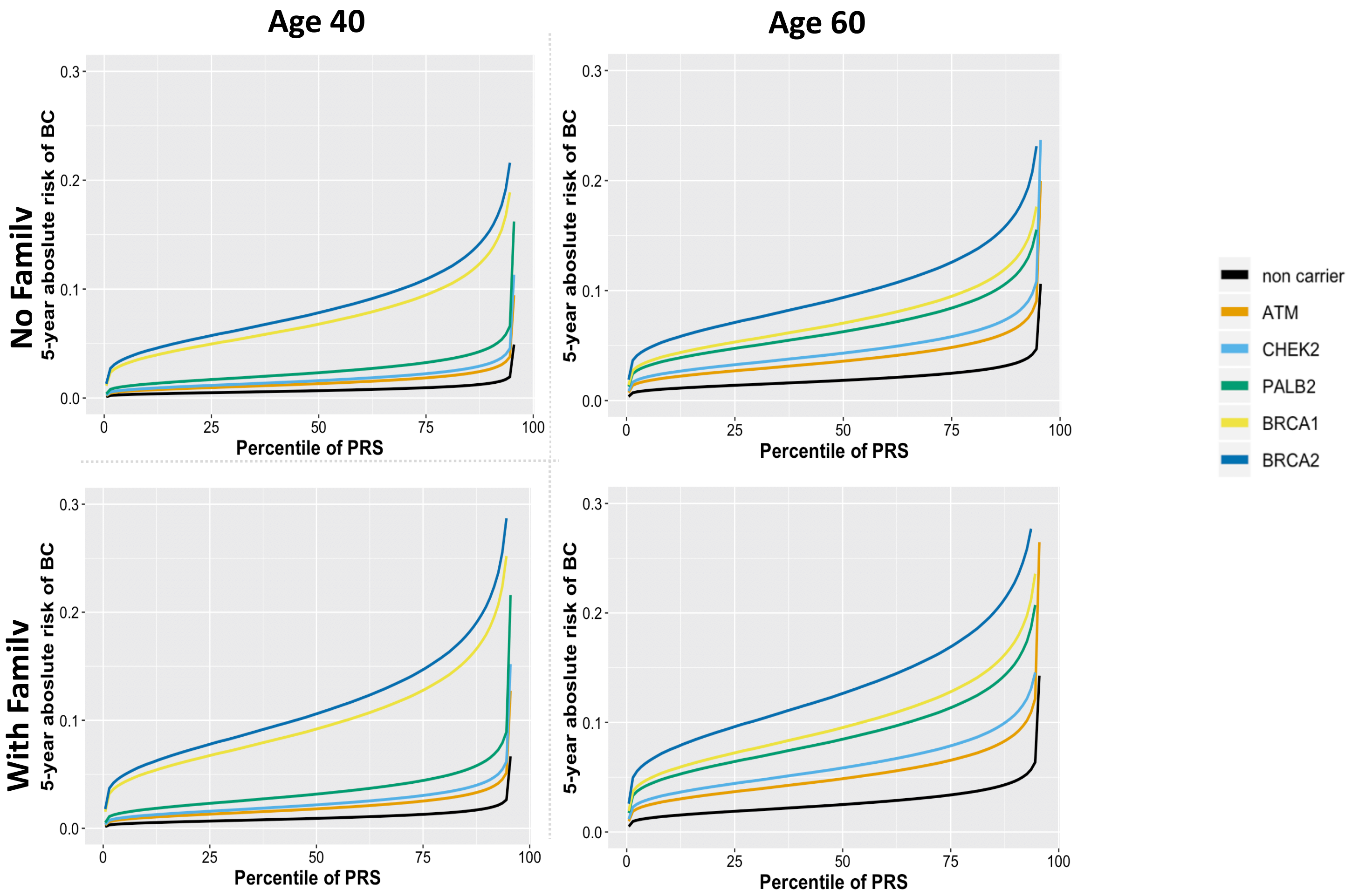


Figure 3.1: 5-year absolute risk of breast cancer across 1%-95%cetile of PRS for different mutation carriers at age 40 and 60, with and without family history. PRS is standardized with a mean of 0 and standard deviation of 1.

3.4 Discussion

In our large, population-based case-control study, we jointly evaluated the association between PRS and pathogenic variant in nine breast cancer predisposition genes and risk of breast cancer. The relative risk gradient associated with per unit change in PRS among carriers of pathogenic variants was similar to that among non-carriers. In addition, we have shown that PRS could be particularly important for estimating breast cancer risk among carriers of pathogenic variants in moderate penetrance genes such as *CHEK2* and *ATM*, enabling more precise approach for MRI screening strategy and breast cancer risk management.

Both common variants in form of PRS and rare variants in cancer predisposition genes contribute to breast cancer risk. Consistent with prior studies(20, 39, 136, 137), our results also showed that the odds ratio for PRS, *BRCA1*, and *BRCA2* decreased with increasing age. The analysis of PRS-by-variant interaction for each individual gene did not show any significant results (Supplementary Table 3.7). However, the direction of effect was consistent with past literatures, indicating a decreasing effect of BRCA1 and BRCA2 on breast cancer risk as PRS increased. The effect of pathogenic variant in BRCA2 is slightly larger than that of BRCA1 (Table 3.2) which may be due to random chance in our dataset and may due to the fact that there were more BRCA2 carriers than BRCA1 carriers in our samples.

Breast MRI is recommended for women with a lifetime risk of breast cancer of 20%-25%(160, 162). Our results suggest that PRS can help delineate which women with pathogenic variants in moderate penetrance genes fall above or below this level of risk. For instance, *ATM* carriers at the 10th percentile of PRS have an estimated lifetime risk of breast cancer of 12.8% which is similar to population average(163). Utilization of PRS could have clinical impact as it can stratify risk among these carriers in order to create a targeted screening and more personalized prevention strategies. In addition, the addition of PRS in women with CHEK2 and ATM pathogenic variants, may help determine when to initiate screening by examining the 5 year risk of breast cancer.

We also showed that ER negative PRS was more accurate in predicting ER negative breast cancer, suggesting that subtype specific PRS could eventually be used to target screening or preventive interventions that are specific to particular subtypes although absolute risk estimates of ER negative breast cancer are low outside of BRCA1 mutations.

Prior work found that the OR of breast cancer associated with per unit change in PRS among *BRCA1* and *BRCA2* variant carriers recruited from cancer genetics clinics was slightly smaller than the OR in the general population (137). The difference between those findings and ours may be due to the smaller number of *BRCA1* and *BRCA2* variant carriers in our study, and hence smaller power to detect subtle differences in PRS ORs between carriers and non-carriers. The differences may also be due to differences in ascertainment (high-risk individuals versus the general population) or analysis (retrospective survival analysis versus prospective logistic regression). Another prior study examined the combined effect of PRS and *CHEK2* variant carriage and they found the effect gradient by PRS was similar in carriers vs non-carriers, consistent with our results(138). Although our study had smaller number of *BRCA1*, *BRCA2* and *CHEK2* carriers compared to previous studies (Supplementary Table 3.5 & 3.6), our study is the first to evaluate the joint effect of PRS and pathogenic variant in nine different breast cancer predisposition genes in the general population. We were able to examine breast cancer predisposition genes other than *BRCA1, BRCA2* and *CHEK2*.

Our study also has certain limitations. First, the PRS was calculated based on 105 SNPs whereas the most recent PRS has been updated to include 313 SNPs(20). Future studies should perform using the updated PRS incorporating more SNPs. Second, our study only used women with non-Hispanic European ancestry. PRS constructed specifically for European ancestry has been found to be less precise for other ancestry groups such as African Americans(164). A multi-ethnic cohort can shed further light in understanding the genetic contribution to breast cancer risk in other ethnicities. In addition, we have limited numbers of ER negative breast cancer cases which may limit our statistical power in examining subtype specific effect estimates. Although we are one of the largest studies to study the combined effect of PRS and rare variant on breast cancer risk, an even larger sample size could potentially provide more power in detecting interactions between PRS and pathogenic variants in breast cancer predisposition genes, as well as increased precision modeling risk in the tails of the PRS among carriers.

As many multigene testing panels becomes readily available and the cost of genotyping and sequencing goes down, women can obtain their genetic information for both rare variants in breast cancer predisposition genes and common variants. Hence, future guidelines and prediction models should increasingly consider the joint usage of both common and rare variants. Our study shows that when common variants are jointly analyzed as PRS, they can contribute significantly to the risk prediction of rare variant carriers of moderate penetrance genes, suggesting future breast cancer risk prediction models should include both PRS and rare variants to provide a more precise and personalized estimate of risk for variant carriers. Further studies (such as simulated screening studies to assess surveillance strategies) are also needed to validate the effect estimates of our final model and try to understand the clinical implication of using both rare variant in genes in addition to BRCA and PRS.

Appendix

Supplemental Materials for Chapter 1

Supplemental Table 1.1: trait-specific and cancer-specific effect of lead SNPs and proxy SNPs in birth weight, childhood obesity, and adult BMI. Trait beta, se, and nearest gene were obtained from published GWAS studies; cancer-specific beta and p-value were obtained from GAME-ON consortium studies. bw=birth weight; c_bmi=childhood BMI; BMI= adult BMI; WHR=waist-hip-ratio; EA=effect allele; BC=breast cancer; OC=ovarian cancer; PC=prostate cancer; LC=lung cancer; CC=colorectal cancer; b=beta; p=p-value

Trait	**SNP**	**EA**	**Trait-beta**	**Trait-se**	**Trait-gene**	**BC_b**	**BC_p**	**OC_b**	**OC_p**	**PC-b**	**PC_p**	**LC_b**	**LC_p**	**CC_b**	**CC_p**
bw	rs1801253	G	0.041	0.007	ADRB1	-0.009	0.628	0.012	0.694	0.003	0.889	-0.008	0.670	-0.028	0.395
bw	rs1042725	T	0.047	0.005	HMGA2	0.011	0.537	0.007	0.810	0.008	0.705	-0.012	0.504	-0.021*	0.483*
bw	rs9883204	C	0.059	0.006	ADCY5	0.004	0.824	0.026	0.408	-0.009	0.728	-0.017	0.406	-0.089	0.008
bw	rs900400	C	0.072	0.006	CCNL1	-0.005	0.765	-0.023	0.406	0.022	0.289	0.004	0.823	-0.007	0.823
bw	rs724577	C	0.042	0.006	LCORL	0.024	0.208	-0.011	0.712	0.032	0.159	-0.027	0.165	0.027*	0.43*
bw	rs4432842	C	0.034	0.006	5q11.2	0.023	0.228	0.054	0.066	0.002	0.930	0.024	0.215	-0.030	0.363
bw	rs6931514	G	0.050	0.006	CDKAL1	0.042	0.031	-0.010	0.748	0.035	0.096	0.011	0.578	0.016	0.639
c_bmi	rs7550711	T	0.105	0.019	*GPR61*	-0.128	0.012	0.059	0.457	0.025	0.690	-0.040	0.408	-0.083	0.360
c_bmi	rs543874	G	0.077	0.009	*SEC16B*	-0.045	0.038	0.035	0.311	0.014	0.602	-0.008	0.720	0.007	0.844
c_bmi	rs12041852	G	0.046	0.007	*TNNI3K*	-0.031	0.072	-0.005	0.864	-0.001	0.966	NA	NA	0.039	0.193
c_bmi	rs7132908	A	0.066	0.008	*FAIM2*	-0.004	0.812	-0.027	0.329	0.052	0.007	-0.030	0.095	-0.020	0.498
c_bmi	rs12429545	A	0.076	0.010	*OLFM4*	-0.003	0.896	-0.026	0.528	0.002	0.934	0.028	0.297	-0.035	0.429
c_bmi	rs1421085	C	0.059	0.007	*FTO*	-0.045	0.009	0.015	0.589	-0.021	0.329	-0.011	0.543	0.009	0.755
c_bmi	rs8092503	G	0.045	0.008	*RAB27B*	-0.039	0.058	-0.071	0.029	-0.028	0.211	-0.026	0.213	-0.005	0.898
c_bmi	rs6567160	C	0.05	0.008	*MC4R*	-0.042	0.038	0.023	0.468	-0.023	0.364	0.025	0.229	-0.012	0.724
c_bmi	rs13387838	A	0.139	0.025	*ADAM23*	-0.161	0.156	-0.038	0.781	-0.083	0.376	0.150	0.121	0.005	0.963
c_bmi	rs11676272	G	0.068	0.007	*ADCY3*	-0.033	0.056	-0.017	0.549	0.009	0.662	0.022	0.205	0.017	0.559
c_bmi	rs4854349	C	0.09	0.009	*TMEM18*	-0.042	0.063	0.049	0.175	0.023	0.374	0.014	0.536	0.062	0.094

Supplemental Table 1.1: trait-specific and cancer-specific effect of lead SNPs and proxy SNPs in birth weight, childhood obesity, and adult BMI. (CONTINUED)

c_bmi	rs13130484	T	0.067	0.007	*GNPDA2*	-0.030	0.082	0.020	0.479	-0.026	0.175	-0.017	0.354	0.028	0.341
c_bmi	rs987237	G	0.062	0.009	*TFAP2B*	0.043	0.055	0.057	0.101	-0.028	0.214	0.000	0.997	0.021	0.578
c_bmi	rs13253111	A	0.042	0.007	*ELP3*	-0.004	0.831	0.003	0.902	0.023	0.196	0.022	0.210	0.009	0.763
c_bmi	rs3829849	T	0.041	0.007	*LMX1B*	0.067	0.000	-0.031	0.283	-0.017	0.417	0.002	0.902	-0.018	0.559
BMI	rs17024393	C	0.066	0.009	GNAT2	-0.122	0.016	0.043	0.581	0.020	0.742	-0.037	0.438	0.083	0.360
BMI	rs543874	G	0.048	0.004	SEC16B	-0.045	0.038	0.035	0.311	0.014	0.602	-0.008	0.720	0.007	0.844
BMI	rs2820292	C	0.020	0.003	NAV1	0.006	0.721	0.056	0.040	0.012	0.593	0.022	0.217	0.013	0.652
BMI	rs657452	A	0.023	0.003	AGBL4	-0.011	0.520	-0.024	0.390	-0.005	0.811	-0.001	0.978	NA	NA
BMI	rs11583200	C	0.018	0.003	ELAVL4	-0.015	0.388	-0.027	0.341	-0.014	0.535	0.006	0.760	0.058*	0.051*
BMI	rs3101336	C	0.033	0.003	NEGR1	-0.028	0.104	-0.019	0.493	-0.008	0.706	-0.013	0.491	-0.059	0.055
BMI	rs12566985	G	0.024	0.003	FPGT-TNNI3K	-0.031	0.070	-0.005	0.867	-0.001	0.965	-0.009*	0.96*	0.038	0.196
BMI	rs12401738	A	0.021	0.003	FUBP1	-0.036	0.045	0.014	0.618	-0.022	0.265	0.054	0.004	-0.032	0.311
BMI	rs11165643	T	0.022	0.003	PTBP2	-0.005	0.779	-0.001	0.976	-0.004	0.851	0.024*	0.177*	0.035*	0.229*
BMI	rs17094222	C	0.025	0.004	HIF1AN	-0.014	0.515	0.031	0.355	0.042	0.119	0.019	0.386	-0.017	0.632
BMI	rs11191560	C	0.031	0.005	NT5C2	-0.017	0.561	-0.059	0.250	-0.007	0.833	0.007	0.810	-0.008*	0.872*
BMI	rs7903146	C	0.023	0.003	TCF7L2	-0.052	0.007	0.074	0.014	0.011	0.654	0.024	0.216	NA	NA
BMI	rs7899106	G	0.040	0.007	GRID1	0.004	0.935	0.060	0.349	-0.017	0.735	-0.113	0.008	-0.107*	0.115*
BMI	rs12286929	G	0.022	0.003	CADM1	0.010	0.572	-0.016	0.570	-0.018	0.345	0.051	0.004	0.015	0.602
BMI	rs11030104	A	0.041	0.004	BDNF	0.048	0.028	0.001	0.975	-0.030	0.256	0.036	0.091	-0.022	0.538
BMI	rs2176598	T	0.020	0.004	HSD17B12	0.003	0.882	0.021	0.512	0.025	0.248	0.007	0.718	-0.040	0.235
BMI	rs3817334	T	0.026	0.003	MTCH2	0.023	0.194	-0.007	0.809	0.000	0.985	0.007	0.707	-0.025	0.410
BMI	rs4256980	G	0.021	0.003	TRIM66	0.001	0.944	-0.008	0.781	-0.008	0.714	-0.003	0.884	-0.011	0.720
BMI	rs11057405	G	0.031	0.006	CLIP1	0.037	0.197	-0.026	0.558	0.026	0.536	-0.005	0.861	0.099	0.061
BMI	rs7138803	A	0.032	0.003	BCDIN3D	0.002	0.932	-0.016	0.569	0.045	0.018	-0.030	0.094	0.020	0.498

Supplemental Table 1.1: trait-specific and cancer-specific effect of lead SNPs and proxy SNPs in birth weight, childhood obesity, and adult BMI. (CONTINUED)

BMI	rs9581854	T	0.030	0.005	MTIF3	-0.059	0.008	-0.002	0.945	0.016	0.546	-0.011	0.614	0.014	0.717
BMI	rs12429545	A	0.033	0.005	OLFM4	-0.003	0.896	-0.026	0.528	0.002	0.934	0.028	0.297	0.035	0.429
BMI	rs10132280	C	0.023	0.003	STXBP6	-0.012	0.514	-0.045	0.134	-0.008	0.718	0.014	0.457	-0.054	0.081
BMI	rs12885454	C	0.021	0.003	PRKD1	0.003	0.874	0.045	0.112	0.006	0.775	0.047	0.012	0.025	0.422
BMI	rs11847697	T	0.049	0.008	PRKD1	-0.010	0.835	0.039	0.577	-0.037	0.506	0.061	0.183	0.076	0.290
BMI	rs7141420	T	0.024	0.003	NRXN3	0.003	0.862	0.007	0.786	-0.005	0.812	-0.017	0.329	-0.003	0.925
BMI	rs3736485	A	0.018	0.003	DMXL2	-0.017	0.360	0.051	0.067	-0.011	0.549	0.026	0.154	0.003	0.929
BMI	rs16951275	T	0.031	0.004	MAP2K5	-0.025	0.228	0.049	0.131	-0.045	0.079	0.023	0.283	-0.074	0.028
BMI	rs12446632	G	0.040	0.005	GPRC5B	0.013	0.616	0.094	0.015	-0.026	0.413	0.003	0.910	0.025	0.548
BMI	rs2650492	A	0.021	0.004	SBK1	0.021	0.281	0.014	0.632	-0.022	0.354	0.041	0.032	-0.026	0.434
BMI	rs3888190	A	0.031	0.003	ATP2A1	0.011	0.546	0.059	0.033	-0.013	0.558	0.046	0.011	-0.071	0.018
BMI	rs9925964	A	0.019	0.003	KAT8	-0.040	0.023	0.020	0.474	-0.043	0.020	0.036	0.046	-0.005	0.860
BMI	rs758747	T	0.023	0.004	NLRC3	0.022	0.283	0.010	0.751	0.050	0.045	0.045	0.021	NA	NA
BMI	rs1558902	A	0.082	0.003	FTO	-0.046	0.009	0.015	0.578	-0.022	0.312	-0.011*	0.543*	-0.009	0.755
BMI	rs1000940	G	0.019	0.003	RABEP1	0.031	0.104	-0.017	0.565	-0.024	0.258	-0.003	0.871	0.003*	0.913*
BMI	rs12940622	G	0.018	0.003	RPTOR	-0.035	0.043	-0.023	0.409	-0.010	0.600	-0.024	0.170	-0.020	0.501
BMI	rs1808579	C	0.017	0.003	C18orf8	-0.038	0.025	0.003	0.912	0.040	0.060	0.023	0.193	0.0006*	0.985*
BMI	rs7243357	T	0.022	0.004	GRP	-0.027	0.227	-0.005	0.880	0.009	0.747	0.014	0.526	0.038	0.318
BMI	rs6567160	C	0.056	0.004	MC4R	-0.042	0.038	0.023	0.468	-0.023	0.364	0.025	0.229	-0.012	0.724
BMI	rs17724992	A	0.019	0.004	PGPEP1	0.021	0.301	-0.008	0.785	0.008	0.720	0.023	0.251	-0.023	0.481
BMI	rs29941	G	0.018	0.003	KCTD15	-0.029	0.111	-0.014	0.635	-0.009	0.706	0.005	0.804	NA	NA
BMI	rs2075650	A	0.026	0.005	TOMM40	-0.003	0.906	0.002	0.956	0.062	0.050	0.040	0.107	-0.047	0.270
BMI	rs2287019	C	0.036	0.004	QPCTL	-0.040	0.063	0.027	0.437	-0.002	0.952	-0.010	0.639	-0.014	0.696
BMI	rs3810291	A	0.028	0.004	ZC3H4	-0.056	0.008	-0.038	0.208	-0.002	0.932	0.018	0.371	0.011	0.735

Supplemental Table 1.1: trait-specific and cancer-specific effect of lead SNPs and proxy SNPs in birth weight, childhood obesity, and adult BMI. (CONTINUED)

BMI	rs2121279	T	0.025	0.004	LRP1B	0.034	0.178	0.040	0.328	-0.024	0.425	0.012	0.643	0.009	0.837
BMI	rs1528435	T	0.018	0.003	UBE2E3	-0.014	0.434	-0.009	0.738	0.006	0.770	0.005	0.764	-0.018	0.538
BMI	rs7599312	G	0.022	0.003	ERBB4	-0.022	0.257	-0.044	0.157	-0.001	0.977	0.002	0.901	0.0312*	0.354*
BMI	rs10182181	G	0.031	0.003	ADCY3	-0.033	0.054	-0.020	0.471	0.009	0.637	0.027	0.132	0.013	0.661
BMI	rs11126666	A	0.021	0.003	KCNK3	0.006	0.752	0.034	0.261	0.024	0.286	-0.010	0.592	NA	NA
BMI	rs1016287	T	0.023	0.003	FLJ30838	-0.005	0.794	0.029	0.321	0.020	0.360	-0.024	0.216	0.017*	0.598*
BMI	rs11688816	G	0.017	0.003	EHBP1	0.000	0.984	-0.020	0.474	0.058	0.001	-0.002	0.922	NA	NA
BMI	rs13021737	G	0.060	0.004	TMEM18	-0.045	0.049	0.057	0.116	0.025	0.348	0.020	0.387	0.063	0.093
BMI	rs16851483	T	0.048	0.008	RASA2	-0.008	0.829	-0.007	0.897	0.002	0.954	0.002	0.949	-0.011	0.845
BMI	rs1516725	C	0.045	0.005	ETV5	0.013	0.622	-0.012	0.775	-0.013	0.647	-0.026	0.327	0.016	0.701
BMI	rs6804842	G	0.019	0.003	RARB	-0.006	0.727	-0.008	0.759	-0.013	0.539	-0.027	0.141	0.001	0.972
BMI	rs2365389	C	0.020	0.003	FHIT	0.025	0.154	-0.015	0.586	-0.005	0.816	-0.020	0.270	-0.059	0.048
BMI	rs3849570	A	0.019	0.003	GBE1	0.033	0.076	0.049	0.082	0.029	0.161	-0.001	0.953	-0.002	0.952
BMI	rs13078960	G	0.030	0.004	CADM2	-0.027	0.223	0.033	0.340	0.011	0.671	-0.001	0.973	-0.076	0.030
BMI	rs13107325	T	0.048	0.007	SLC39A8	-0.055	0.158	0.028	0.600	0.001	0.979	0.052	0.122	-0.026	0.612
BMI	rs11727676	T	0.036	0.006	HHIP	0.071	0.106	0.029	0.596	0.083	0.046	0.037	0.337	NA	NA
BMI	rs10938397	G	0.040	0.003	GNPDA2	-0.031	0.077	0.025	0.363	-0.025	0.193	-0.016	0.368	0.027	0.349
BMI	rs17001654	G	0.031	0.005	SCARB2	-0.050	0.046	-0.018	0.649	0.052	0.084	0.040	0.116	-0.007	0.870
BMI	rs2112347	T	0.026	0.003	POC5	-0.032	0.072	-0.028	0.326	-0.004	0.868	0.017	0.362	-0.003	0.913
BMI	rs9400239	C	0.019	0.003	FOXO3	0.006	0.753	0.001	0.960	0.046	0.048	-0.013	0.473	-0.008	0.797
BMI	rs13191362	A	0.028	0.005	PARK2	0.036	0.174	0.035	0.410	0.014	0.658	0.039	0.165	-0.057	0.207
BMI	rs205262	G	0.022	0.004	C6orf106	-0.006	0.765	-0.011	0.718	0.029	0.177	0.028	0.155	0.0187*	0.556*

Supplemental Table 1.1: trait-specific and cancer-specific effect of lead SNPs and proxy SNPs in birth weight, childhood obesity, and adult BMI. (CONTINUED)

BMI	rs2033529	G	0.019	0.003	TDRG1	-0.021	0.259	0.002	0.934	-0.018	0.418	-0.017	0.374	0.022	0.494
BMI	rs2207139	G	0.045	0.004	TFAP2B	0.049	0.028	0.068	0.057	-0.016	0.509	0.008	0.739	0.052	0.182
BMI	rs1167827	G	0.020	0.003	HIP1	0.004	0.850	0.014	0.622	-0.037	0.070	-0.012	0.490	NA	NA
BMI	rs2245368	C	0.032	0.006	PMS2L11	-0.015	0.679	-0.019	0.652	-0.017	0.596	NA	NA	-0.043	0.255
BMI	rs17405819	T	0.022	0.003	HNF4G	0.000	0.999	0.009	0.754	0.005	0.807	0.001	0.961	-0.040	0.209
BMI	rs2033732	C	0.019	0.004	RALYL	-0.044	0.029	0.005	0.868	-0.027	0.267	-0.015	0.462	-0.006	0.848
BMI	rs6477694	C	0.017	0.003	EPB41L4B	-0.010	0.575	-0.015	0.597	-0.018	0.324	-0.010	0.603	0.053	0.089
BMI	rs1928295	T	0.019	0.003	TLR4	-0.018	0.293	0.030	0.282	0.032	0.108	0.010	0.587	-0.021*	0.474*
BMI	rs10733682	A	0.017	0.003	LMX1B	-0.012	0.500	-0.018	0.515	0.018	0.420	-0.001	0.974	0.046	0.114
BMI	rs4740619	T	0.018	0.003	C9orf93	0.002	0.915	-0.023	0.405	0.009	0.644	0.023	0.202	-0.083	0.004
BMI	rs10968576	G	0.025	0.003	LINGO2	0.002	0.936	0.007	0.808	-0.008	0.704	-0.003	0.893	0.081	0.010
WHR	rs984222	G	0.034	0.003	TBX15-WARS2	-0.030	0.085	0.010	0.716	-0.027	0.161	NA	NA	0.011*	0.700*
WHR	rs1011731	G	0.028	0.003	DNM3-PIGC	-0.031	0.070	0.025	0.359	0.020	0.346	0.004	0.831	-0.015*	0.613*
WHR	rs4846567	G	0.034	0.004	LYPLAL1	0.032	0.092	0.063	0.038	0.026	0.214	0.029	0.147	-0.037	0.252
WHR	rs718314	G	0.030	0.004	ITPR2-SSPN	0.005	0.806	0.050	0.107	0.023	0.312	0.000	0.995	-0.037*	0.271*
WHR	rs1443512	A	0.031	0.004	HOXC13	-0.022	0.284	-0.024	0.470	0.028	0.248	-0.004	0.853	0.057*	0.095*
WHR	rs10195252	T	0.033	0.003	GRB14	-0.019	0.282	0.026	0.348	0.043	0.019	0.001	0.954	0.001	0.982
WHR	rs4823006	A	0.023	0.003	ZNRF3-KREMEN1	0.029	0.094	-0.055	0.045	0.020	0.310	0.006	0.724	-0.007	0.815
WHR	rs6784615	T	0.043	0.007	NISCH-STAB1	0.034	0.373	-0.048	0.399	-0.067	0.097	0.054	0.163	-0.007	0.921
WHR	rs6795735	C	0.025	0.003	ADAMTS9	0.023	0.176	0.016	0.555	-0.019	0.341	-0.004	0.838	0.018*	0.547*
WHR	rs6861681	A	0.022	0.004	CPEB4	-0.056	0.002	-0.029	0.316	-0.014	0.552	-0.012	0.538	-0.031	0.321

Supplemental Table 1.1: trait-specific and cancer-specific effect of lead SNPs and proxy SNPs in birth weight, childhood obesity, and adult BMI. (CONTINUED)

WHR	rs9491696	G	0.042	0.003	RSPO3	-0.014	0.398	0.020	0.460	0.018	0.401	NA	NA	0.037	0.204
WHR	rs6905288	A	0.036	0.003	VEGFA	-0.011	0.595	-0.032	0.281	-0.039	0.061	-0.018	0.370	0.033	0.262
WHR	rs1294421	G	0.028	0.003	LY86	-0.007	0.684	0.002	0.937	0.014	0.476	0.019	0.284	NA	NA
WHR	rs1055144	T	0.040	0.004	NFE2L3	-0.042	0.053	-0.011	0.756	-0.066	0.008	0.003	0.897	0.045	0.233

"*" denoates estimates obtained from the proxy SNP

Supplemental Table 1.2: A summary of final number of SNPs included in the analysis. OV= overall; ER-=ER negative; CC=Clear-cell type; EN=Endometroid type; S= Serous type; AG= Aggressive type; AD= Adenocarcinoma; SQ= Squamous type

	Breast Cancer		**Ovarian Cancer**				**Prostate Cancer**		**Lung Cancer**			**Colorectal Cancer**
	OV	ER -	OV	CC	EN	S	OV	AG	OV	AD	SQ	OV
Birth Weight	7	7	7	7	7	7	7	7	7	7	7	7(2 proxy SNP used)
Childhood Obesity	15	15	15	15	15	15	15	15	14	14	14	15 (4 proxy SNPs used)
Adult BMI	77	77	77	77	72	77	77	77	76 (3 proxy SNPs used)			69*
WHR	14	14	14	14	14	14	14	14	12			13**

*: 10 proxy SNPs used but no good proxy found for the remaining 8 unmatched leading SNPs

**:5 proxy SNPs used but no good proxy for rs1294421

Adult BMI:SNP rs12016871 has been merged into rs9581854 and thus rs9581854 was used for all the analyses instead.
BMI_Lung Cancer Proxy: rs1558902 (sur: rs1421085 ,r2=1); rs12566985 (sur: rs10493544, r2=0.966); rs11165643 (sur: rs10489741, r2=1);
BMI_Colorectal Cancer Proxy: rs11165643 (sur: rs10489741, r2=1); rs1016287 (sur: rs887912, r2=1); rs7599312 (sur:rs13427822, r2=1); rs205262 (sur: rs6457792, r2=0.959); rs1928295 (sur: rs9408902, r2=1); rs11191560 (sur: rs10883832, r2=1); rs1000940 (sur: rs3026101, r2=1); rs11583200 (sur: rs12028252, r2=1); rs7899106 (sur: rs17105752, r2=1); rs1808579 (sur: rs891386, r2=1));
Childhood BMI_Colorectal Cancer Proxy: rs1421085 (sur: rs1558902 ,r2=1); rs7550711(sur: rs17024393, r2=0.85); rs7132908 (sur: rs7138803, r2=0.89); rs987237 (sur: rs2206277, r2=0.89);
WHR_Colorectal Cancer Proxy:rs1011731(sur:rs2301453, r2=1); rs1443512(sur:rs9804784, r2=0.947); rs6795735(sur:rs9311910, r2=1); rs718314(sur:rs7132434, r2=1); rs984222(sur:rs10923712,r2=0.967)

Supplemental Table 1.3: Mendelian randomization odds ratios (ORs) of childhood BMI and adult BMI across five different cancer types obtained using summary data from GAME-ON consortium, ONLY for overlap regions (FTO, MC4R, TMEM18, SEC16B, TNNI3K, TFAP2B).

		Childhood BMI		**Adult BMI (77 SNP)**	
		OR (95%CI)	**p-value**	**OR(95%CI)**	**p-value**
Breast Cancer	*All*	0.63 (0.52,0.76)	2.53x10^{-6} *	0.53 (0.42,0.70)	1.08x10^{-6}*
	ER_negative	0.59 (0.44,0.80)	5.5x10^{-4} *	0.48 (0.32,0.71)	2.61x10^{-4}*
Ovarian Cancer	*All*	1.29 (0.95,1.75)	0.099	1.53 (1.00,2.25)	0.047
	Clear_cell	1.82 (0.76,4.33)	0.17	2.21 (0.70,6.97)	0.18
	Endometrioid	1.88 (1.01, 3.49)	0.045	2.74 (1.21,6.22)	0.016
	Serous	1.07 (0.73,1.55)	0.74	1.1 (0.67,1.8)	0.72
Prostate Cancer	*All*	0.93 (0.74,1.16)	0.52	0.87 (0.65,1.18)	0.38
	Aggressive	1.03 (0.75,1.43)	0.84	0.91 (0.59,1.41)	0.68
Lung Cancer	*All*	1.05 (0.86,1.29)	0.62	1.06 (0.82,1.38)	0.65
	Adenocarcinoma	0.86 (0.63,1.17)	0.33	0.83 (0.56,1.24)	0.37
	Squamous	1.28 (0.93,1.76)	0.13	1.25 (0.83,1.88)	0.28
Colorectal Cancer	*All*	1.34 (0.97,1.86)	0.076	1.44	0.094

Supplemental Table 1.4. Effect estimates from Egger regression for adult BMI, childhood BMI, birth weight, and WHR with various cancer and cancer subtypes.

Adult BMI		**MR**	**Egger regression**					
		OR (95%CI)	**Intercept**	**StdDev**	**p**	**OR_egg**	**StdDev**	**p**
Breast cancer	Overall	0.66 (0.57, 0.77)	0.0035	0.0056	0.53	0.59	0.1949	0.0076
	ER-neg	0.59 (0.46, 0.75)	0.0022	0.0088	0.8	0.55	0.3050	0.049
Ovarian Cancer	Overall	1.35(1.05,1.72)	-0.0093	0.0088	0.29	1.80	0.3082	0.054
	Clearcell	1.68 (0.84, 3.36)	-0.043	0.0251	0.083	6.69	0.8775	0.03
	Endometrioid	1.34 (0.80, 2.26)	-0.033	0.0184	0.078	3.74	0.6403	0.038
	Serous	1.3 (0.97, 1.76)	0.0032	0.0110	0.77	1.17	0.3742	0.68
Prostate Cancer	Overall	1.01 (0.84, 1.21)	0.0096	0.0066	0.15	0.74	0.2324	0.19
	Aggressive	1.11 (0.85, 1.44)	0.018	0.0095	0.062	0.63	0.3317	0.16
Lung Cancer	Overall	1.27 (1.09, 1.49)	0.011	0.0057	0.057	0.90	0.2000	0.59
	Adenocarcinoma	0.93 (0.73, 1.19)	0.0062	0.0088	0.48	0.76	0.3082	0.39
	Squamous	1.54 (1.20, 1.96)	0.013	0.0089	0.14	1.01	0.3130	0.98
Colorectal Cancer	Overall	1.39 (1.06, 1.82)	0.0082	0.0098	0.4	1.08	0.3317	0.82

Supplemental Table 1.4. Effect estimates from Egger regression for adult BMI, childhood BMI, birth weight, and WHR with various cancer and cancer subtypes. (CONTINUED)

Childhood BMI		**MR**	**Egger regression**					
		OR (95%CI)	**Intercept**	**StdDev**	**p**	**OR_egg**	**StdDev**	**p**
Breast cancer	Overall	0.71 (0.60, 0.80)	0.048	0.0214	0.026	0.34	0.3464	0.0017
	ER-neg	0.69 (0.53, 0.98)	0.049	0.0346	0.15	0.32	0.5568	0.039
Ovarian Cancer	Overall	1.07 (0.82, 1.39)	-0.053	0.0332	0.12	2.44	0.5385	0.1
	Clearcell	1.45 (0.68, 3.09)	-0.055	0.0954	0.57	3.42	1.5556	0.43
	Endometrioid	1.47 (0.86, 2.52)	-0.18	0.0678	0.0094	23.81	1.0909	0.0037
	Serous	0.91 (0.65, 1.26)	-0.035	0.0412	0.4	1.57	0.6708	0.5
Prostate Cancer	Overall	1.01 (0.83, 1.22)	-0.02	0.0243	0.42	1.38	0.3873	0.42
	Aggressive	1.1 (0.83, 1.45)	-0.013	0.0346	0.72	1.35	0.5745	0.61
Lung Cancer	Overall	1.01 (0.85, 1.2)	-0.0015	0.0230	0.95	1.04	0.3742	0.92
	Adenocarcinoma	0.9 (0.69, 1.19)	0.0064	0.0361	0.86	0.82	0.5657	0.73
	Squamous	1.08 (0.82, 1.43)	-0.009	0.0361	0.8	1.25	0.5745	0.7
Colorectal Cancer	Overall	1.2 (0.90, 1.59)	-0.02	0.0249	0.41	1.63	0.4000	0.22

Supplemental Table 1.4. Effect estimates from Egger regression for adult BMI, childhood BMI, birth weight, and WHR with various cancer and cancer subtypes. (CONTINUED)

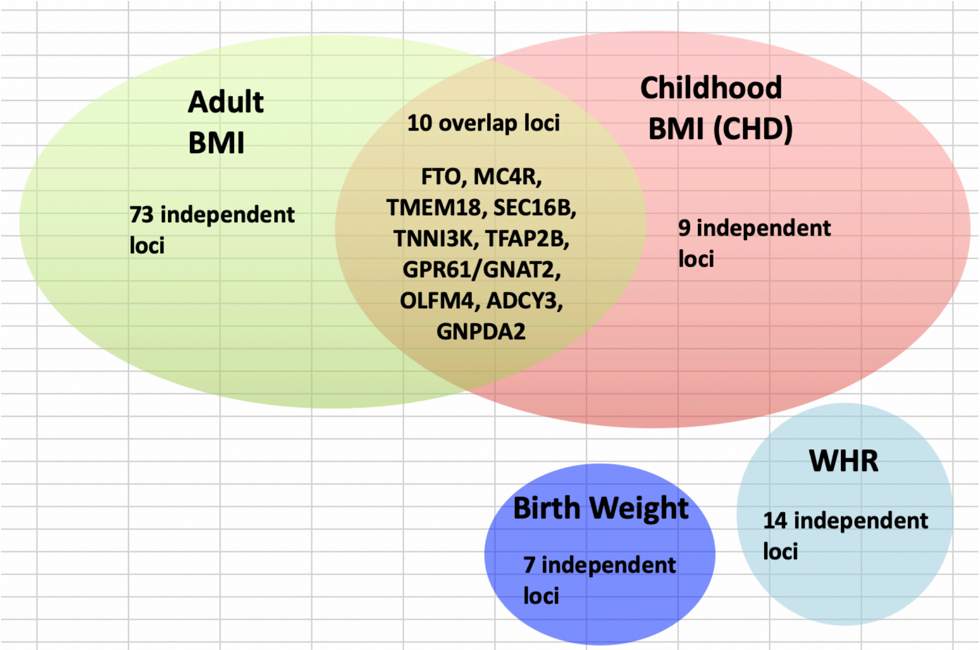
WHR		**MR**	**Egger regression**					
		OR (95%CI)	**Intercept**	**StdDev**	**p**	**OR_egg**	**StdDev**	**p**
Breast cancer	Overall	0.73 (0.53,1.00)	0.0048	0.0263	0.85	0.63	0.8307	0.58
	ER-neg	0.74 (0.45, 1.21)	-0.0021	0.0412	0.96	0.79	1.3000	0.86
Ovarian Cancer	Overall	1.19 (0.73, 1.94)	-0.037	0.0424	0.38	3.67	1.3153	0.32
	Clearcell	1.31 (0.32, 5.30)	-0.027	0.1183	0.82	3.00	3.7683	0.77
	Endometrioid	1.03 (0.38, 2.84)	-0.09	0.0860	0.3	16.61	2.7092	0.3
	Serous	1.34 (0.73, 2.46)	-0.0019	0.0520	0.97	1.42	1.6217	0.83
Prostate Cancer	Overall	1.02 (0.72, 1.46)	0.046	0.0310	0.14	0.25	0.9747	0.15
	Aggressive	1.19 (0.71, 1.98)	-0.0085	0.0447	0.85	1.54	1.4036	0.76
Lung Cancer	Overall	1.15 (0.80, 1.66)	-0.017	0.0316	0.6	1.97	1.0440	0.52
	Adenocarcinoma	0.9 (0.51, 1.58)	-0.076	0.0490	0.12	10.80	1.6125	0.14
	Squamous	1.33 (0.75, 2.36)	-0.0005	0.0500	0.99	1.35	1.6340	0.86
Colorectal Cancer	Overall	1.29 (0.75, 2.22)	-0.068	0.0458	0.14	10.38	1.4318	0.1

Supplemental Table 1.4. Effect estimates from Egger regression for adult BMI, childhood BMI, birth weight, and WHR with various cancer and cancer subtypes. (continued)

Birth Weight		**MR**	**Egger regression**					
		OR (95%CI)	**Intercept**	**StdDev**	**p**	**OR_egg**	**StdDev**	**p**
Breast cancer	Overall	1.22 (0.93, 1.60)	0.04	0.0300	0.18	1.75	1.3231	0.34
	ER-neg	1.01 (0.66, 1.53)	0.078	0.0469	0.1	4.35	2.0855	0.11
Ovarian Cancer	Overall	1.07 (0.69, 1.65)	0.069	0.0469	0.15	3.46	1.8589	0.18
	Clearcell	2.75 (0.82, 9.30)	-0.2	0.1342	0.14	0.01	0.0939	0.07
	Endometrioid	0.79 (0.33, 1.92)	-0.023	0.0980	0.82	0.83	0.9094	0.92
	Serous	0.85 (0.50, 1.45)	0.13	0.0583	0.025	14.01	3.7434	0.021
Prostate Cancer	Overall	1.33 (0.96, 1.82)	0.0043	0.0346	0.9	0.82	0.9048	0.77
	Aggressive	1.63 (1.03, 2.57)	-0.042	0.0500	0.4	0.28	0.5273	0.19
Lung Cancer	Overall	0.93 (0.70, 1.23)	0.0011	0.0307	0.97	1.1	1.0466	0.88
	Adenocarcinoma	0.95 (0.62, 1.46)	-0.0099	0.0469	0.83	0.87	0.9324	0.87
	Squamous	0.99 (0.64, 1.52)	0.01	0.0480	0.83	1.23	1.1107	0.82
Colorectal Cancer	Overall	0.69 (0.44, 1.10)	-0.026	0.0510	0.96	1.38	1.1735	0.75

Supplemental Table 1.5. Association between various genetic score for different adiposity traits were associated with the other traits using summary results from genome-wide association studies for these traits

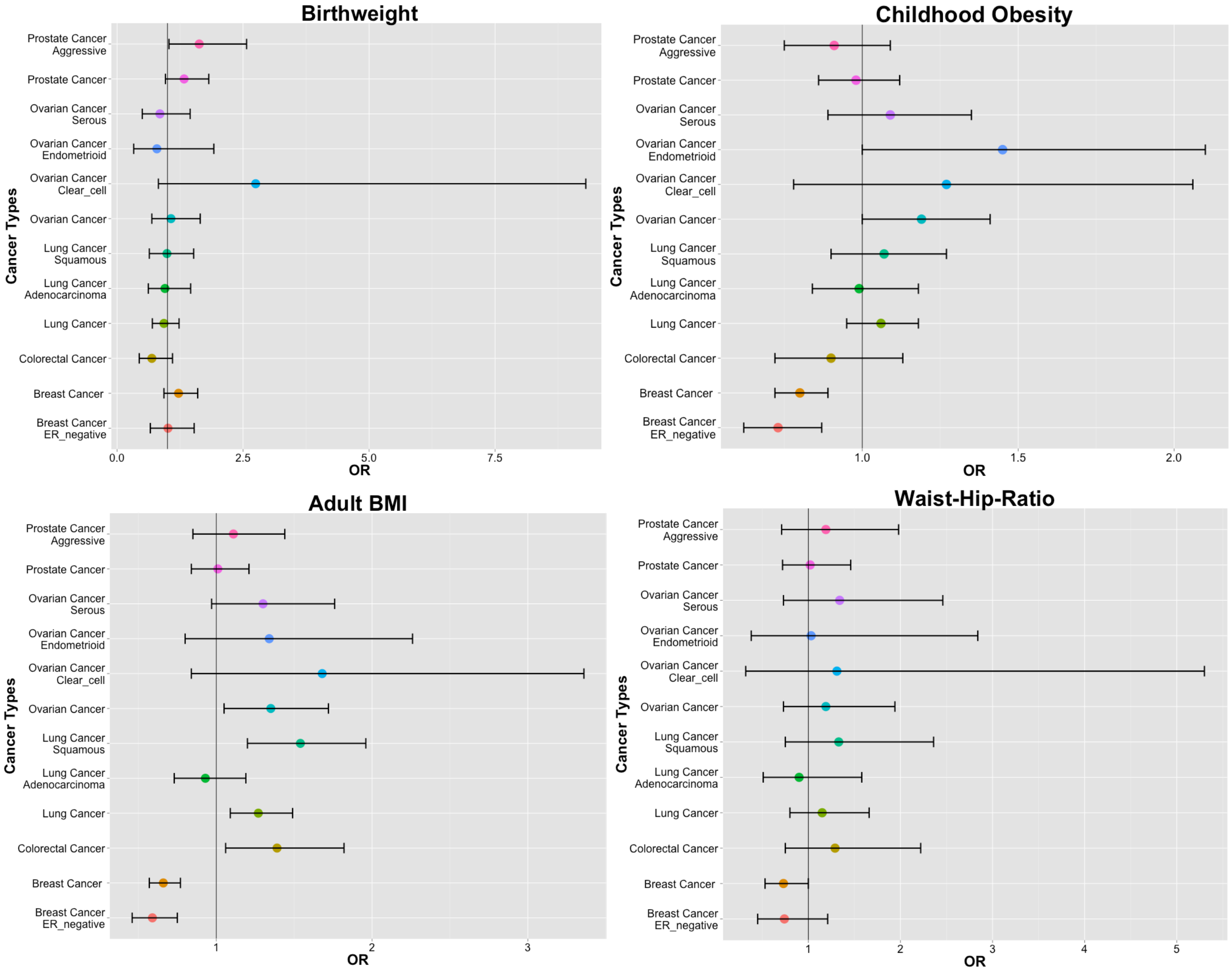
	BMI	**WHR**	**Childhood BMI**
$\mathbf{G}_{\textbf{bmi}}$	---	OR: 0.99 95%CI: 0.96,1.02 p: 0.378	OR: 2.63 95%CI: 2.45, 2.82 p<0.0001
$\mathbf{G}_{\textbf{whr}}$	OR: 0.82 95%CI: 0.77, 0.86 p: $3.9\text{x}10^{-13}$	---	OR: 0.91 95%CI: 0.79, 1.05 p:0.21
$\mathbf{G}_{\textbf{chd bmi}}$	OR: 1.87 95%CI: 1.82, 1.93 p<0.0001	OR: 0.96 95%CI: 0.92, 0.99 p: 0.01	---
$\mathbf{G}_{\textbf{bmi excluding overlap snps}}$	---	---	OR: 2.16 95%CI: 1.98, 2.37 p<0.0001
$\mathbf{G}_{\textbf{chd bmi excluding overlap snps}}$	OR: 1.40 95%CI: 1.31, 1.50 p<0.0001	---	---



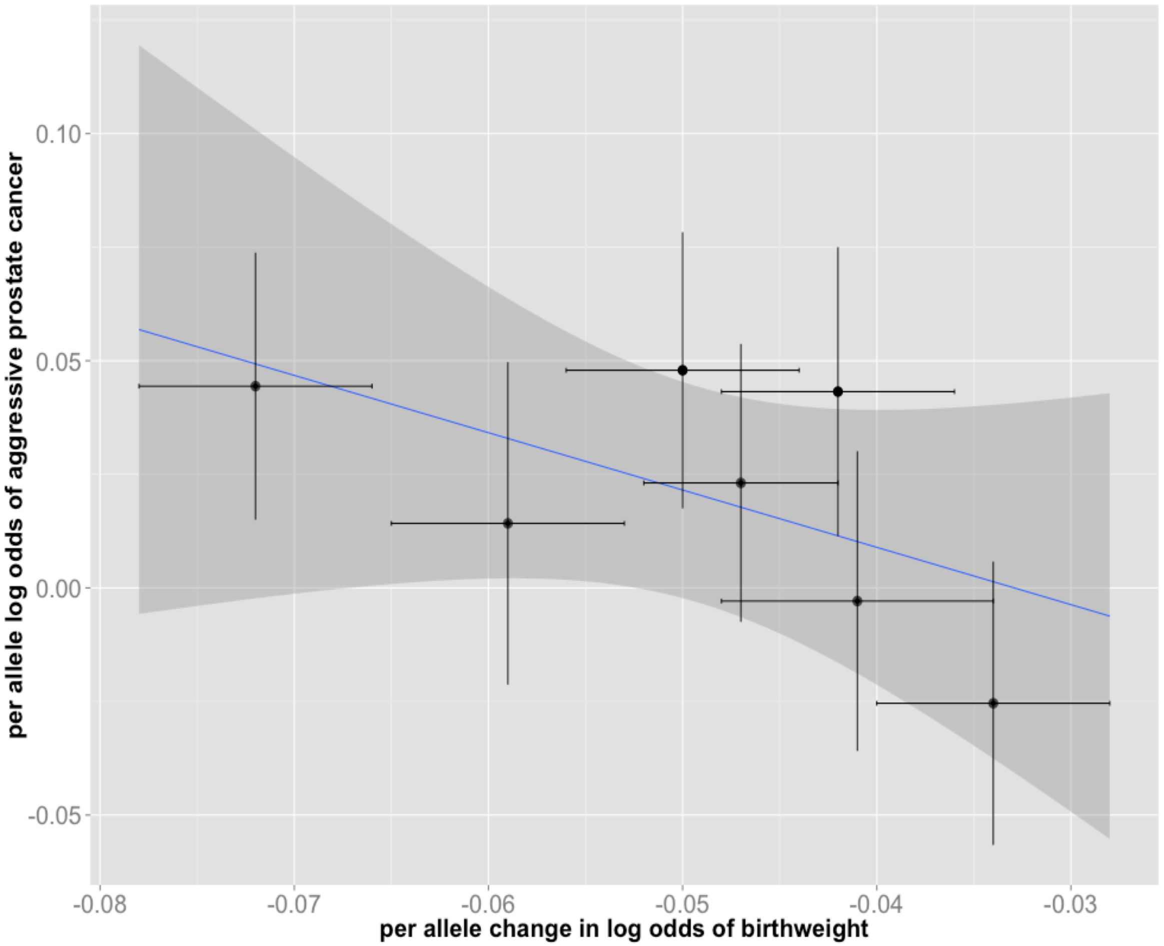
Supplemental Figure 1.1: Illustration of independent and overlap regions between any two traits of WHR, birth weight, childhood BMI and adult BMI

Detailed overlap loci for the 10 genes are shown below:

FTO(rs1421085 for CHD; rs1558902 for adult BMI)
MC4R (rs6567160 for CHD; rs6567160 for adult BMI)
TMEM18 (rs4854349 for CHD; rs13021737 for adult BMI)
SEC16B (rs543874 for both CHD and adult BMI)
TNNI3K (rs12041852 for CHD; and rs12566985 for adult BMI)
TFAP2B (rs987237 for CHD; and rs2207139 for adult BMI)
GPR61/GNAT2 (rs7550711 for CHD; and rs17024393 for adult BMI)
OLFM4 (rs12429545 for CHD; and rs12429545 for adult BMI)
ADCY3 (rs11676272 for CHD; and rs10182181 for adult BMI)
GNPDA2 (rs13130484 for CHD; and rs10938397 for adult BMI)



Supplemental Figure 1.2: Illustrative figure of results from mendelian randomization analysis of birth weight, childhood obesity, adult BMI, and waist-hip-ratio across five different cancer types using summary data from GAME-ON consortium.
*:statistically significance p>0.05



Supplemental Figure 1.3: Scatterplot of SNP-specific effects for the associations with birthweight and aggressive prostate cancer, for all 7 birthweight-associated SNPs. SNP-specific vertical and horizontal bars correspond to standard errors for the aggressive prostate cancer association and BMI association respectively. The shaded region corresponds to 95%CI of the association between BMI and aggressive prostate cancer risk

Supplemental Materials for Chapter 2

Supplemental Table 2.1: Effect sizes of 313 SNPs used to compute the PRS score

chr: pos	Reference_allele	Effect_allele	beta_Overall.Breast.Cancer
1:100880328	A	T	0.0373
1:10566215	A	G	-0.0586
1:110198129	CAAA	C	0.0458
1:114445880	G	A	0.0621
1:118141492	A	C	0.0452
1:120257110	T	C	0.0385
1:121280613	A	G	0.0881
1:121287994	A	G	-0.0673
1:145604302	C	CT	-0.0399
1:149906413	T	C	0.0548
1:155556971	G	A	0.0499
1:168171052	CA	C	-0.068
1:172328767	T	TA	-0.0435
1:18807339	T	C	-0.0564
1:201437832	C	T	0.0917
1:202184600	C	T	-0.0065
1:203770448	T	A	0.0498
1:204502514	T	TTCTGAAACAGGG	-0.0321
1:208076291	G	A	-0.0366
1:217053815	T	G	0.0417
1:217220574	G	A	-0.044
1:220671050	C	T	0.0418
1:242034263	A	G	0.1428
1:41380440	C	T	0.0426
1:41389220	T	C	0.155
1:46670206	TC	T	0.0447
1:51467096	CT	C	0.0374
1:7917076	G	A	-0.0409
1:88156923	G	A	0.0494
1:88428199	C	A	-0.0387
10:114777670	C	T	0.0472
10:115128491	T	C	-0.0592
10:123095209	G	A	-0.0538
10:123340107	A	G	0.1508
10:123340431	GC	G	-0.2408

Supplemental Table 2.1: Effect sizes of 313 SNPs used to compute the PRS score (CONTINUED)

chr: pos	Reference_allele	Effect_allele	beta_Overall.Breast.Cancer
10:123349324	A	T	-0.2609
10:13892298	G	A	0.0371
10:22032942	A	G	-0.058
10:22477776	ACC	A	0.1687
10:22861490	A	C	0.0875
10:38523626	C	A	0.0404
10:5794652	A	G	0.047
10:64299890	A	G	-0.1345
10:64819996	G	T	0.0472
10:71335574	C	T	-0.0404
10:80851257	G	T	-0.0805
10:80886726	A	G	0.0762
10:95292187	CAA	C	-0.0512
11:103614438	T	G	0.0147
11:108267402	C	CA	-0.0022
11:111696440	T	C	-0.0396
11:116727936	A	T	-0.0423
11:122966626	A	G	-0.0383
11:129243417	T	G	-0.0543
11:129461016	A	G	0.0453
11:18664241	T	G	0.0461
11:1895708	C	A	-0.0762
11:42844441	C	T	-0.0336
11:433617	T	C	-0.0437
11:44368892	G	A	0.0374
11:46318032	C	G	-0.0748
11:65553492	C	A	0.0425
11:65572431	G	A	-0.0347
11:69328130	A	T	-0.0423
11:69330983	G	A	0.1022
11:69331418	C	T	0.1782
11:803017	A	G	0.0457
12:103097887	C	T	0.0546
12:111600134	G	T	-0.0442
12:115108136	T	C	0.0465

Supplemental Table 2.1: Effect sizes of 313 SNPs (CONTINUED)			
chr: pos	**Reference_allele**	**Effect_allele**	**beta_Overall.Breast.Cancer**
12:115796577	A	G	-0.0428
12:115835836	T	C	-0.0813
12:120832146	C	T	0.0516
12:14413931	G	C	0.0484
12:28149568	C	T	-0.062
12:28174817	C	T	-0.0856
12:28347382	C	T	-0.0521
12:29140260	G	A	0.0647
12:293626	A	G	0.0401
12:57146069	T	G	-0.0579
12:70798355	A	T	0.0469
12:83064195	G	GA	0.0671
12:85004551	C	T	0.0348
12:96027759	A	G	-0.0867
13:32839990	G	A	0.0424
13:32972626	A	T	0.2687
13:43501356	A	G	0.0517
13:73806982	T	C	0.0345
13:73960952	A	G	0.0399
14:105213978	T	G	0.0399
14:37128564	C	A	-0.0733
14:37228504	C	T	0.039
14:68660428	T	C	-0.0474
14:68979835	T	C	-0.0911
14:91751788	TC	T	0.038
14:91841069	A	G	0.0513
14:93070286	C	T	-0.0577
15:100905819	A	C	-0.0608
15:46680811	C	A	-0.1973
15:50694306	A	G	-0.0417
15:66630569	G	A	-0.0369
15:67457698	A	G	0.0782
15:75750383	T	C	-0.0413
15:91512267	G	T	-0.0589
16:10706580	G	A	-0.074
16:23007047	G	T	0.1218

Supplemental Table 2.1: Effect sizes of 313 SNPs (CONTINUED)			
chr: pos	**Reference_allele**	**Effect_allele**	**beta_Overall.Breast.Cancer**
16:4008542	CAAAAA	C	-0.0329
16:4106788	C	A	-0.03
16:52538825	C	A	0.1147
16:52599188	C	T	0.107
16:53809123	C	T	-0.0704
16:53861139	C	T	-0.0338
16:53861592	G	A	-0.0337
16:54682064	G	A	0.0477
16:6963972	C	G	0.0354
16:80648296	A	G	0.0839
16:85145977	T	C	-0.0211
16:87086492	T	C	-0.0469
17:29168077	G	T	-0.0568
17:39251123	T	C	0.0799
17:40127060	T	C	0.0174
17:40485239	G	T	-0.0571
17:40744470	G	A	0.2017
17:43212339	C	CT	0.0438
17:44283858	G	A	-0.054
17:53209774	A	C	-0.0793
17:77781725	A	G	-0.0401
18:11696613	C	T	-0.0381
18:20634253	C	T	-0.0415
18:24125857	T	C	0.0346
18:24337424	C	G	0.0455
18:24518050	AT	A	-0.0599
18:25407513	C	G	0.0399
18:29981526	G	A	-0.1058
18:42411803	G	C	-0.0877
18:42888797	T	C	-0.0542
19:13249921	G	T	0.0956
19:17393925	C	A	0.0378
19:18569492	C	T	-0.0719
19:19517054	C	CGGGCG	0.0437
19:44283031	T	C	0.0619
19:46166073	T	C	-0.036

Supplemental Table 2.1: Effect sizes of 313 SNPs (CONTINUED)

chr: pos	Reference_allele	Effect_allele	beta_Overall.Breast.Cancer
19:55816678	C	T	-0.0359
2:10138983	T	C	0.0603
2:121058254	A	G	-0.0334
2:121089731	T	C	-0.0427
2:121159205	G	A	-0.044
2:121246568	T	C	0.0992
2:172974566	C	G	-0.0473
2:174212910	A	G	0.0593
2:192381934	C	T	0.0316
2:19315675	T	A	-0.0331
2:202204741	T	C	-0.0492
2:217920769	G	T	-0.1318
2:217955896	GA	G	-0.2016
2:218292158	C	G	-0.0757
2:218714845	G	A	-0.0431
2:241388857	C	A	-0.1232
2:25129473	A	G	-0.0427
2:29179452	G	C	-0.0066
2:29615233	T	C	-0.0427
2:39699510	C	CT	-0.0402
2:70172587	G	A	-0.0412
2:88358825	G	C	0.0473
20:11379842	T	C	0.0844
20:41613706	C	G	0.0315
20:52296849	G	A	0.044
20:5948227	G	A	0.076
21:16364756	T	G	0.0646
21:16566350	A	G	0.0595
21:16574455	C	A	-0.0707
21:47762932	G	A	0.0946
22:19766137	C	T	-0.0367
22:29121087	A	G	0.1839
22:29135543	G	A	0.0654
22:29203724	C	T	0.1405
22:29551872	A	G	-0.1716
22:38583315	AAAAG	AAAAGAAAG	-0.0471

Supplemental Table 2.1: Effect sizes of 313 SNPs (CONTINUED)

chr: pos	Reference_allele	Effect_allele	beta_Overall.Breast.Cancer
22:39343916	T	A	0.0407
22:40904707	CT	C	0.1148
22:43433100	C	T	-0.06
22:45319953	G	A	-0.0134
22:46283297	G	A	0.0736
3:141112859	CTT	C	0.0551
3:172285237	G	A	0.0422
3:189774456	C	T	-0.0478
3:27353716	C	A	0.0748
3:27388664	C	G	0.0502
3:29294845	C	T	-0.1281
3:30684907	C	T	0.0592
3:46888198	T	C	-0.0806
3:4742251	A	G	0.0616
3:49709912	C	CT	-0.0367
3:55970777	A	AT	-0.1195
3:59373745	C	T	-0.0394
3:63887449	T	TTG	0.0648
3:71620370	T	G	-0.0374
3:87037543	A	G	-0.0723
3:99403877	G	A	-0.0376
4:106069013	G	T	0.0471
4:126752992	A	AAT	-0.0377
4:143467195	C	T	-0.0569
4:151218296	CATATTT	C	0.0388
4:175842495	G	A	-0.0898
4:175847436	C	A	0.0348
4:187503758	A	T	0.0357
4:38784633	G	T	0.0489
4:84370124	TAA	TA	-0.0464
4:89240476	G	A	0.0352
4:92594859	TTCTTTC	T	-0.0407
5:104300273	G	T	-0.0487
5:122478676	C	A	-0.0386
5:122705244	C	T	0.0944
5:1279790	C	T	0.0617

Supplemental Table 2.1: Effect sizes of 313 SNPs (CONTINUED)

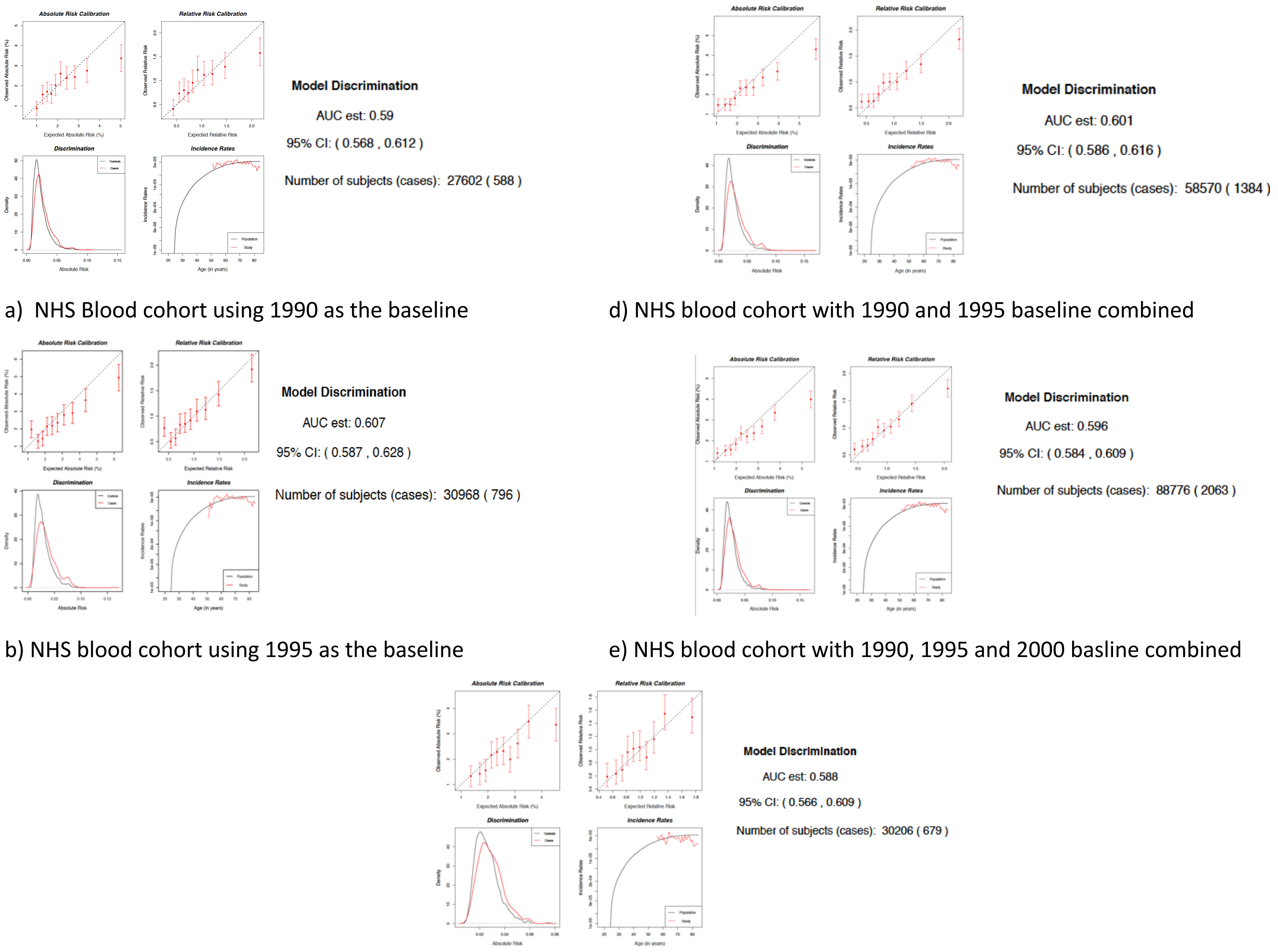
chr: pos	Reference_allele	Effect_allele	beta_Overall.Breast.Cancer
5:1296255	A	AG	-0.0549
5:131640536	A	G	0.0392
5:132407058	C	T	-0.0388
5:1353077	T	C	0.1552
5:158244083	C	T	-0.0677
5:16231194	G	C	-0.0426
5:169591460	T	C	0.0412
5:173358154	G	A	0.0365
5:176134882	T	C	0.0363
5:2777029	G	A	0.0391
5:32579616	TCA	T	0.0363
5:345109	T	C	0.084
5:44508264	G	GT	-0.1177
5:44619502	A	G	-0.1101
5:44649944	C	T	0.0492
5:44706498	A	G	0.0497
5:44853593	G	C	-0.0336
5:52679539	C	CA	0.0571
5:55662540	C	CT	-0.0458
5:55965167	C	T	0.0394
5:56023083	T	G	0.1366
5:56042972	C	T	0.0865
5:56045081	T	C	-0.0564
5:58241712	C	T	-0.0434
5:71965007	G	A	-0.041
5:73234583	T	C	-0.0363
5:77155397	GT	G	-0.0408
5:79180995	G	GA	0.0328
5:81512947	TA	T	-0.0598
5:90789470	G	A	-0.0564
6:130341728	C	CT	0.0472
6:13713366	G	C	-0.0553
6:149595505	T	C	-0.0476
6:151949806	A	C	0.0703
6:151955914	A	G	0.1449
6:152022664	CAAAAAAA	C	0.0137

Supplemental Table 2.1: Effect sizes of 313 SNPs (CONTINUED)

chr: pos	Reference_allele	Effect_allele	beta_Overall.Breast.Cancer
6:152023191	G	A	0.0626
6:152055978	A	T	0.074
6:152432902	C	T	0.0649
6:16399557	C	T	-0.0373
6:169006947	C	G	-0.0308
6:170332621	T	C	0.0373
6:18783140	G	A	0.0326
6:20537845	CA	C	-0.0391
6:21923810	T	C	-0.0321
6:27425644	G	C	-0.0737
6:43227141	G	A	-0.064
6:82263549	AAT	A	0.0477
6:85912194	CAA	C	0.0762
6:87803819	T	C	0.0383
7:101552440	G	A	-0.0568
7:102481842	T	C	0.0418
7:130656911	C	T	-0.0476
7:130674481	G	A	0.0416
7:139943702	CT	C	0.0582
7:144048902	G	T	-0.0563
7:21940960	A	G	-0.0467
7:25569548	C	T	-0.0486
7:28869017	G	A	-0.0572
7:55192256	A	C	-0.0349
7:91459189	A	ATT	0.0452
7:94113799	T	C	0.0449
7:98005235	G	A	-0.0467
7:99948655	T	G	0.042
8:102483100	T	C	0.0593
8:106358620	A	T	-0.0745
8:117209548	A	G	-0.0417
8:120862186	A	G	0.0527
8:124563705	T	C	0.0477
8:124571581	G	A	0.034
8:124739913	T	G	0.0466
8:128213561	C	CA	-0.043

Supplemental Table 2.1: Effect sizes of 313 SNPs (CONTINUED)

chr: pos	Reference_allele	Effect_allele	beta_Overall.Breast.Cancer
8:128370949	C	G	0.0642
8:128372172	A	G	0.0597
8:129199566	G	A	0.0615
8:143669254	A	G	-0.0346
8:170692	T	C	0.0477
8:17787610	CT	C	-0.0377
8:23447496	A	G	-0.0389
8:23663653	C	A	0.0335
8:29509616	A	C	-0.0601
8:36858483	A	G	-0.076
8:76230943	A	G	0.0755
8:76333056	C	T	0.1129
8:76378165	G	T	-0.0391
9:110303808	TAA	T	0.0797
9:110837073	A	G	0.1158
9:110837176	C	T	0.0653
9:110849525	G	T	0.0153
9:110885479	C	T	0.0877
9:119313486	A	G	-0.0462
9:129424719	A	G	-0.0382
9:136146597	C	T	0.04
9:21964882	CAAAA	C	0.055
9:22041998	C	G	0.0289
9:36928288	T	C	0.0249
9:6880263	A	G	0.0348
9:87782211	T	C	0.0361
9:98362587	T	C	0.0576



Model Discrimination

AUC est: 0.59

95% CI: (0.568 , 0.612)

Number of subjects (cases): 27602 (588)

a) NHS Blood cohort using 1990 as the baseline

Model Discrimination

AUC est: 0.601

95% CI: (0.586 , 0.616)

Number of subjects (cases): 58570 (1384)

d) NHS blood cohort with 1990 and 1995 baseline combined

Model Discrimination

AUC est: 0.607

95% CI: (0.587 , 0.628)

Number of subjects (cases): 30968 (796)

b) NHS blood cohort using 1995 as the baseline

Model Discrimination

AUC est: 0.596

95% CI: (0.584 , 0.609)

Number of subjects (cases): 88776 (2063)

e) NHS blood cohort with 1990, 1995 and 2000 basline combined

Model Discrimination

AUC est: 0.588

95% CI: (0.566 , 0.609)

Number of subjects (cases): 30206 (679)

c) NHS blood cohort using 2000 as the baseline

Supplemental Figure 2.1: Validation analysis results of classic risk factor only model using only NHS blood cohort at different baseline time frames.

Supplemental Materials for Chapter 3

Supplemental Table 3.1: effect sizes for 105 SNPs used to construct overall breast cancer and ER negative-specific PRSs

chr_position[a]	rs_number	effect allele	reference allele	Overall BC_beta	ER-negative_beta	ER-positive_beta
1_10566215	rs616488	A	G	0.0604	0.105	0.0604
1_114448389	rs11552449	T	C	0.0543	0.0543	0.0557
1_121280613	rs11249433	G	A	0.0988	0.0101	0.0988
1_145644984	rs12405132	C	T	0.0406	0.0157	0.0406
1_149927034	rs12048493	C	A	0.0496	0.0396	0.0496
1_202187176	rs6678914	G	A	0.0066	0.0823	0.0066
1_204518842	rs4245739	C	A	0.0272	0.127	0.0272
1_242034263	rs72755295	G	A	0.1376	0.1376	0.1481
10_114773927	rs7904519	G	A	0.0456	0.0691	0.0456
10_123093901	rs11199914	C	T	0.0456	0.0045	0.0456
10_123337335	rs2981579	A	G	0.2376	0.0419	0.2376
10_22032942	rs7072776	A	G	0.0618	0.0193	0.0618
10_22315843	rs11814448	C	A	0.1846	0.1188	0.1846
10_5886734	rs2380205	C	T	0.0234	0.0234	0.0261
10_64261198	rs16917302	A	C	0.0421	0.0421	0.0367
10_64278682	rs10995190	G	A	0.129	0.0932	0.129
10_80841148	rs704010	T	C	0.0787	0.0504	0.0787
11_129461171	rs11820646	C	T	0.0482	0.0482	0.0442
11_1941946	rs909116	T	C	0.0676	0.0371	0.0676
11_65583066	rs3903072	G	T	0.0434	0.0253	0.0434
11_69331418	rs78540526	T	C	0.2758	0.0076	0.2758
12_115836522	rs1292011	A	G	0.0822	0.0209	0.0822
12_14413931	rs12422552	C	G	0.0552	0.0552	0.0483

Supplemental Table 3.1: effect sizes for 105 SNPs used to construct overall breast cancer and ER negative-specific PRSs (CONTINUED)

chr_position[a]	**rs_number**	**effect allele**	**reference allele**	**Overall BC_beta**	**ER-negative_beta**	**ER-positive_beta**
12_28124305	rs27633	G	T	0.0054	0.0054	0.0114
12_28155080	rs10771399	A	G	0.1492	0.1641	0.1492
12_96027759	rs17356907	A	G	0.0898	0.0694	0.0898
13_32972626	rs11571833	T	A	0.2727	0.4346	0.2727
13_73957681	rs6562760	G	A	0.0443	0.0826	0.0443
14_37132769	rs2236007	G	A	0.0719	0.0368	0.0719
14_68660428	rs2588809	T	C	0.0628	0.0047	0.0628
14_69034682	rs999737	C	T	0.0967	0.0752	0.0967
14_91841069	rs941764	G	A	0.0463	0.0186	0.0463
14_93104072	rs11627032	T	C	0.0481	0.0481	0.0438
16_52586341	rs3803662	A	G	0.2032	0.1254	0.2032
16_53813367	rs17817449	T	G	0.0599	0.0736	0.0599
16_53855291	rs11075995	A	T	0.0421	0.086	0.0421
16_80650805	rs13329835	G	A	0.0786	0.0426	0.0786
17_48274291	rs2075555	G	T	0.0106	0.0106	0.012
17_53056471	rs6504950	G	A	0.0676	0.0321	0.0676
17_77781725	rs745570	A	G	0.0389	0.0389	0.0349
18_24337424	rs527616	G	C	0.0499	0.0178	0.0499
18_24570667	rs1436904	T	G	0.0489	0.0056	0.0489
18_42399590	rs6507583	A	G	0.087	0.034	0.087
19_17389704	rs8170	A	G	0.0415	0.1479	0.0415
19_18571141	rs4808801	A	G	0.0718	0.0541	0.0718
19_41858921	rs1800470	G	A	0.0012	0.0012	0.007
19_44286513	rs3760982	A	G	0.051	0.051	0.0521

Supplemental Table 3.1: effect sizes for 105 SNPs used to construct overall breast cancer and ER negative-specific PRSs (CONTINUED)						
chr_position[a]	**rs_number**	**effect allele**	**reference allele**	**Overall BC_beta**	**ER-negative_beta**	**ER-positive_beta**
2_121245122	rs4849887	C	T	0.095	0.1135	0.095
2_172972971	rs2016394	G	A	0.0425	0.0084	0.0425
2_174212894	rs1550623	A	G	0.0531	0.0202	0.0531
2_19320803	rs12710696	T	C	0.0365	0.0628	0.0365
2_201717014	rs74943274	A	G	0.0839	0.175	0.0839
2_202149589	rs1045485	G	C	0.0415	0.0415	0.027
2_217905832	rs13387042	A	G	0.1225	0.0484	0.1225
2_218296508	rs16857609	T	C	0.0727	0.0727	0.0721
2_29119585	rs67073037	A	T	0.0052	0.0851	0.0052
2_38377405	rs184577	A	G	0.007	0.0135	0.007
20_32588095	rs2284378	T	C	0.0142	0.0289	0.0142
20_62157646	rs13039229	C	A	0.0052	0.0052	0.0039
20_62217589	rs311499	C	T	0.014	0.0615	0.014
21_16520832	rs2823093	G	A	0.0653	0.0069	0.0653
22_29621477	rs132390	C	T	0.0945	0.0945	0.0824
22_40876234	rs6001930	C	T	0.1201	0.1201	0.1092
3_27416013	rs4973768	T	C	0.0985	0.0413	0.0985
3_30682939	rs12493607	C	G	0.0485	0.0016	0.0485
3_46866866	rs6796502	G	A	0.0828	0.0828	0.0892
3_4742276	rs6762644	G	A	0.055	0.0225	0.055
3_63967900	rs1053338	G	A	0.0588	0.0588	0.0554
4_106084778	rs9790517	T	C	0.0483	0.0125	0.0483
4_175846426	rs6828523	C	A	0.1019	0.0017	0.1019
5_1279790	rs10069690	T	C	0.0599	0.1613	0.0599

Supplemental Table 3.1: effect sizes for 105 SNPs used to construct overall breast cancer and ER negative-specific PRSs (CONTINUED)

chr_position[a]	rs_number	effect allele	reference allele	Overall BC_beta	ER-negative_beta	ER-positive_beta
5_1297488	rs2736108	C	T	0.0622	0.1216	0.0622
5_158244083	rs1432679	C	T	0.0717	0.0717	0.0695
5_16187528	rs13162653	G	T	0.0321	0.0321	0.0287
5_32567732	rs2012709	T	C	0.0358	1.00E-04	0.0358
5_44706498	rs10941679	G	A	0.1278	0.0336	0.1278
5_55995035	rs16886113	G	T	0.1406	0.0264	0.1406
5_56031884	rs889312	C	A	0.1212	0.0594	0.1212
5_58184061	rs10472076	C	T	0.0364	0.0364	0.034
5_58337481	rs1353747	T	G	0.0625	0.0625	0.0629
5_81538046	rs7707921	A	T	0.0513	0.032	0.0513
6_10456706	rs9348512	A	C	0.0017	0.0017	0.0017
6_127606588	rs6569479	T	C	0.008	0.008	0.0121
6_1318878	rs11242675	T	C	0.0249	0.0249	0.02
6_13722523	rs204247	G	A	0.0445	0.016	0.0445
6_149608874	rs9485372	G	A	0.0371	0.0192	0.0371
6_151948366	rs2046210	A	G	0.084	0.1368	0.084
6_151987357	rs9383938	T	G	0.1424	0.2323	0.1424
6_152523550	rs2253407	G	T	0.0055	0.0055	0.0112
6_28926220	rs9257408	C	G	0.034	0.034	0.0339
6_82128386	rs17529111	C	T	0.045	0.0646	0.045
7_130667121	rs4593472	C	T	0.0438	0.0438	0.0455
7_144074929	rs720475	G	A	0.0488	3.00E-04	0.0488
7_91630620	rs6964587	T	G	0.0409	0.0231	0.0409
8_117209548	rs13267382	A	G	0.0437	0.0437	0.0427

Supplemental Table 3.1: effect sizes for 105 SNPs used to construct overall breast cancer and ER negative-specific PRSs (CONTINUED)

chr_position[a]	rs_number	effect allele	reference allele	Overall BC_beta	ER-negative_beta	ER-positive_beta
8_128355618	rs13281615	G	A	0.1001	0.0507	0.1001
8_128694006	rs4733664	C	T	0.0142	0.0142	0.0174
8_129194641	rs11780156	T	C	0.0606	0.0606	0.0621
8_29509616	rs9693444	A	C	0.0626	0.0408	0.0626
8_36858483	rs13365225	A	G	0.0767	0.0963	0.0767
8_76230301	rs6472903	T	G	0.0778	0.0439	0.0778
9_110306115	rs10759243	A	C	0.0595	0.0278	0.0595
9_110888478	rs865686	T	G	0.0984	0.0208	0.0984
9_21854740	rs10965163	C	T	9.00E-04	9.00E-04	0.0003
9_22062134	rs1011970	T	G	0.066	0.066	0.0576

Supplemental Table 3.2: The OR of overall breast cancer for each age group with respect to their PRS (10th percentile, median, 90th percentile) and variant carrier status. Reference group: non carriers with median PRS and no family history

<40 yr old	No Family History			Family History		
OR	**10th% PRS**	**median PRS**	**90th% PRS**	**10th% PRS**	**median PRS**	**90th% PRS**
non-carrier	0.54	1.00	1.85	0.74	1.35	2.52
ATM carrier	1.07	1.95	3.64	1.46	2.66	4.96
CHEK2 carrier	1.28	2.34	4.36	1.74	3.18	5.94
PALB2 carrier	1.89	3.45	6.44	2.57	4.69	8.76
BRCA1 carrier	8.04	14.67	27.38	10.94	19.96	37.26
BRCA2 carrier	9.01	16.44	30.68	12.26	22.37	41.75
BARD1 carrier	0.86	1.57	2.93	1.18	2.15	4.00
BRIP1 carrier	0.80	1.46	2.71	1.09	1.98	3.70
CDH1 carrier	3.17	5.78	10.76	4.31	7.87	14.67
NF1 carrier	1.06	1.94	3.62	1.45	2.64	4.93

40-50 yr old	No Family History			Family History		
OR	**10th% PRS**	**median PRS**	**90th% PRS**	**10th% PRS**	**median PRS**	**90th% PRS**
non-carrier	0.56	1.00	1.79	0.76	1.35	2.44
ATM carrier	1.10	1.95	3.52	1.50	2.66	4.80
CHEK2 carrier	1.33	2.35	4.24	1.81	3.20	5.78
PALB2 carrier	1.96	3.46	6.25	2.67	4.72	8.51
BRCA1 carrier	5.22	9.23	16.66	7.11	12.58	22.70
BRCA2 carrier	6.12	10.82	19.52	8.33	14.74	26.60
BARD1 carrier	0.89	1.57	2.84	1.21	2.15	3.87
BRIP1 carrier	0.82	1.46	2.63	1.12	1.99	3.58
CDH1 carrier	3.27	5.78	10.43	4.45	7.88	14.21
NF1 carrier	1.10	1.94	3.50	1.50	2.65	4.77

Supplemental Table 3.2: The OR of overall breast cancer for each age group with respect to their PRS (10th percentile, median, 90th percentile) and variant carrier status. Reference group: non carriers with median PRS and no family history (CONTINUED)

50-60 yr old	No Family History			Family History		
OR	**10th% PRS**	**median PRS**	**90th% PRS**	**10th% PRS**	**median PRS**	**90th% PRS**
non-carrier	0.58	1.00	1.73	0.79	1.35	2.36
ATM carrier	1.14	1.95	3.41	1.55	2.66	4.65
CHEK2 carrier	1.37	2.35	4.11	1.87	3.21	5.60
PALB2 carrier	2.02	3.46	6.05	2.75	4.72	8.25
BRCA1 carrier	3.30	5.66	9.89	4.50	7.71	13.47
BRCA2 carrier	4.18	7.17	12.52	5.69	9.76	17.06
BARD1 carrier	0.92	1.58	2.75	1.25	2.15	3.75
BRIP1 carrier	0.85	1.46	2.55	1.16	1.99	3.47
CDH1 carrier	3.37	5.78	10.10	4.60	7.88	13.77
NF1 carrier	1.13	1.94	3.39	1.54	2.65	4.62

60-70 yr old	No Family History			Family History		
OR	**10th% PRS**	**median PRS**	**90th% PRS**	**10th% PRS**	**median PRS**	**90th% PRS**
non-carrier	0.60	1.00	1.68	0.81	1.35	2.29
ATM carrier	1.18	1.95	3.30	1.60	2.66	4.50
CHEK2 carrier	1.42	2.35	3.98	1.93	3.21	5.43
PALB2 carrier	2.09	3.47	5.86	2.84	4.72	7.99
BRCA1 carrier	2.09	3.47	5.87	2.84	4.73	8.00
BRCA2 carrier	2.86	4.75	8.03	3.89	6.47	10.94
BARD1 carrier	0.95	1.58	2.67	1.29	2.15	3.63
BRIP1 carrier	0.88	1.46	2.47	1.20	1.99	3.36
CDH1 carrier	3.48	5.78	9.79	4.74	7.88	13.34
NF1 carrier	1.17	1.94	3.29	1.59	2.65	4.48

Supplemental Table 3.2: The OR of overall breast cancer for each age group with respect to their PRS (10th percentile, median, 90th percentile) and variant carrier status. Reference group: non carriers with median PRS and no family history (CONTINUED)

>=70 yr old	No Family History			Family History		
OR	**10th% PRS**	**median PRS**	**90th% PRS**	**10th% PRS**	**median PRS**	**90th% PRS**
non-carrier	0.62	1.00	1.63	0.84	1.35	2.22
ATM carrier	1.21	1.95	3.20	1.65	2.66	4.36
CHEK2 carrier	1.46	2.36	3.86	1.99	3.21	5.26
PALB2 carrier	2.15	3.47	5.68	2.93	4.72	7.74
BRCA1 carrier	1.32	2.13	3.48	1.80	2.90	4.75
BRCA2 carrier	1.95	3.14	5.15	2.66	4.28	7.02
BARD1 carrier	0.98	1.58	2.58	1.33	2.15	3.52
BRIP1 carrier	0.91	1.46	2.39	1.23	1.99	3.26
CDH1 carrier	3.59	5.79	9.48	4.89	7.89	12.92
NF1 carrier	1.21	1.94	3.18	1.64	2.65	4.34

Supplementary Table 3.3: predicted 5-year abolsute risk of developing breast cancer with respect to different PRS, carrier status, and family history. The estimated 5-year risk is displayed with start age of 45, 50 and 55, respectively.

start age: 45	**No Family History**			**Family History**		
OR	**10th% PRS**	**median PRS**	**90th% PRS**	**10th% PRS**	**median PRS**	**90th% PRS**
non-carrier	0.004	0.007	0.012	0.005	0.009	0.016
ATM carrier	0.007	0.013	0.023	0.010	0.018	0.032
CHEK2 carrier	0.009	0.016	0.028	0.012	0.021	0.038
PALB2 carrier	0.013	0.023	0.041	0.018	0.031	0.055
BRCA1 carrier	0.038	0.066	0.118	0.051	0.090	0.159
BRCA2 carrier	0.043	0.076	0.136	0.059	0.104	0.182
start age: 50	**No Family History**			**Family History**		
OR	**10th% PRS**	**median PRS**	**90th% PRS**	**10th% PRS**	**median PRS**	**90th% PRS**
non-carrier	0.007	0.012	0.022	0.010	0.017	0.029
ATM carrier	0.014	0.024	0.042	0.019	0.032	0.056
CHEK2 carrier	0.017	0.029	0.050	0.023	0.039	0.068
PALB2 carrier	0.025	0.042	0.073	0.033	0.057	0.098
BRCA1 carrier	0.044	0.076	0.132	0.061	0.103	0.177
BRCA2 carrier	0.055	0.094	0.162	0.075	0.127	0.217

start age: 55	**No Family History**			**Family History**		
OR	**10th% PRS**	**median PRS**	**90th% PRS**	**10th% PRS**	**median PRS**	**90th% PRS**
non-carrier	0.009	0.014	0.024	0.011	0.019	0.033
ATM carrier	0.016	0.027	0.047	0.022	0.037	0.064
CHEK2 carrier	0.019	0.033	0.057	0.026	0.045	0.077
PALB2 carrier	0.028	0.048	0.083	0.038	0.065	0.111
BRCA1 carrier	0.046	0.078	0.132	0.062	0.105	0.177
BRCA2 carrier	0.058	0.098	0.166	0.079	0.132	0.221

Supplemental Table 3.4:

a) OR of ER- breast cancer for overall BC PRS and ER- PRS across age groups				
	ER- specific PRS		Overall BC PRS	
Age group	OR	95%CI	OR	95%CI
<40	1.467	1.148, 1.874	1.203	1.072, 1.351
40-50	1.465	1.151, 1.865	1.224	1.130, 1.327
50-60	1.463	1.154, 1.856	1.246	1.184, 1.312
60-70	1.462	1.157, 1.847	1.268	1.213, 1.326
>70	1.46	1.161, 1.839	1.292	1.207, 1.381
b) OR of ER+ breast cancer for overall BC PRS and ER+ PRS across age groups				
	ER+ specific PRS		Overall BC PRS	
Age group	OR	95%CI	OR	95%CI
<40	1.928	1.684, 2.209	1.724	1.615, 1.840
40-50	1.922	1.682, 2.196	1.664	1.591, 1.741
50-60	1.916	1.680, 2.184	1.607	1.562, 1.654
60-70	1.909	1.678, 2.173	1.552	1.516, 1.589
>70	1.903	1.676, 2.161	1.499	1.448, 1.552

Supplemental Table 3.5: variant count by gene in the study (restrict to non-Hispanic Europeans)

Gene	Caco	# of variant	non-carriers	% of variant in this case/control group	total# of variant for this gene	% of variant carriers in the total population
BRCA1	control	49	26078	0.2%	300	0.6%
	case	251	26547	0.9%		
BRCA2	control	66	26061	0.3%	441	0.8%
	case	375	26423	1.4%		
BARD1	control	29	26098	0.1%	74	0.1%
	case	45	26753	0.2%		
ATM	control	111	26016	0.4%	339	0.6%
	case	228	26570	0.9%		
BRIP1	control	45	26082	0.2%	109	0.2%
	case	64	26734	0.2%		
CDH1	control	3	26124	0.0%	18	0.0%
	case	15	26783	0.1%		
CHEK2	control	148	25979	0.6%	507	1.0%
	case	359	26439	1.3%		
PALB2	control	36	26091	0.1%	152	0.3%
	case	116	26682	0.4%		
NF1	control	8	26119	0.1%	23	0.0%
	case	15	26783	0.1%		

Supplemental Table 3.6: ER- in the study population by age group

Age Group	ER-	ER+	Ratio ER-/ER+	missing	%missing
<40	86	273	0.315	276	43.5%
40-50	308	1560	0.197	1365	42.2%
50-60	704	3504	0.201	2818	40.1%
60-70	828	5074	0.163	3122	34.6%
>70	711	4711	0.151	1458	21.2%
total	2637	15122	0.1743817	9039	0.50898136

Supplemental Table 3.7: The PRS-by-pathogenic variant interactions for each individual gene

Gene	OR**	95%CI	pvalue
BRCA1	0.63	0.46, 0.88	0.006
BRCA2	0.82	0.62, 1.09	0.16
ATM	1.15	0.89, 1.50	0.29
CHEK2	0.9	0.74, 1.11	0.32
PALB2	0.55	0.36, 0.86	0.0077
BARD1	0.74	0.45, 1.23	0.24
BRIP1	0.82	0.54, 1.27	0.37
CDH1	0.36	0.081, 1.18	0.11
NF1	0.72	0.27, 2.24	0.54
Any genes*	0.81	0.73, 0.91	0.00022

*: if there is a pathogenic variant in any of the nine genes tested

**: this is the effect estimate of the gene x PRS interaction term

Supplemental Table 3.8: Sensitivity analysis from running the final model and including interaction term between carriers of variant in any of the nine genes and PRS for overall Breast Cancer

	<=40	**40-50**	**50-60**	**60-70**	**>70**
PRS in non-carriers	1.63	1.58	1.54	1.51	1.47
	(1.55, 1.71)	(1.53, 1.64)	(1.51, 1.58)	(1.48, 1.54)	(1.42, 1.51)
PRS in any carriers	1.4	1.36	1.33	1.3	1.26
	(1.24, 1.58)	(1.22, 1.53)	(1.19, 1.49)	(1.16, 1.45)	(1.13, 1.42)
BRCA1*	14.18	8.79	5.45	3.38	2.09
	(7.74, 27.5)	(5.79, 13.4)	(3.96, 7.51)	(2.21, 5.17)	(1.10, 3.98)
BRCA2*	15.6	10.4	6.93	4.62	3.08
	(8.69, 29.4)	(6.86, 15.8)	(5.22, 9.20)	(3.41, 6.27)	(1.95, 4.88)
ATM	1.96	1.96	1.96	1.96	1.96
	(1.56, 2.48)	(1.56, 2.48)	(1.56, 2.48)	(1.56, 2.48)	(1.56, 2.48)
CHEK2	2.35	2.35	2.35	2.35	2.35
	(1.93, 2.87)	(1.93, 2.87)	(1.93, 2.87)	(1.93, 2.87)	(1.93, 2.87)
PALB2	3.32	3.32	3.32	3.32	3.32
	(2.28, 4.94)	(2.28, 4.94)	(2.28, 4.94)	(2.28, 4.94)	(2.28, 4.94)
BARD1	1.56	1.56	1.56	1.56	1.56
	(0.98, 2.54)	(0.98, 2.54)	(0.98, 2.54)	(0.98, 2.54)	(0.98, 2.54)
BRIP1	1.47	1.47	1.47	1.47	1.47
	(0.99, 2.21)	(0.99, 2.21)	(0.99, 2.21)	(0.99, 2.21)	(0.99, 2.21)
CDH1	5.46	5.46	5.46	5.46	5.46
	(1.74, 24.0)	(1.74, 24.0)	(1.74, 24.0)	(1.74, 24.0)	(1.74, 24.0)
NF1	1.03	1.03	1.03	1.03	1.03
	(0.58, 1.86)	(0.58, 1.86)	(0.58, 1.86)	(0.58, 1.86)	(0.58, 1.86)

Reference

1. National Cancer Institute. https://wwwcancergov/about-cancer/causes-prevention/risk.
2. Cancer Net. https://wwwcancernet/navigating-cancer-care/cancer-basics/genetics/genetics-cancer.
3. Chen YC, Page JH, Chen R, Giovannucci E. Family history of prostate and breast cancer and the risk of prostate cancer in the PSA era. Prostate. 2008;68(14):1582-91.
4. Lesko SM, Rosenberg L, Shapiro S. Family history and prostate cancer risk. Am J Epidemiol. 1996;144(11):1041-7.
5. Zeegers MP, Jellema A, Ostrer H. Empiric risk of prostate carcinoma for relatives of patients with prostate carcinoma: a meta-analysis. Cancer. 2003;97(8):1894-903.
6. Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. Lancet. 2001;358(9291):1389-99.
7. Fuchs CS, Giovannucci EL, Colditz GA, Hunter DJ, Speizer FE, Willett WC. A prospective study of family history and the risk of colorectal cancer. N Engl J Med. 1994;331(25):1669-74.
8. Mucci LA, Hjelmborg JB, Harris JR, Czene K, Havelick DJ, Scheike T, et al. Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. Jama. 2016;315(1):68-76.
9. Turnbull C, Rahman N. Genetic predisposition to breast cancer: past, present, and future. Annu Rev Genomics Hum Genet. 2008;9:321-45.
10. Bishop DT, Cannon-Albright L, McLellan T, Gardner EJ, Skolnick MH. Segregation and linkage analysis of nine Utah breast cancer pedigrees. Genet Epidemiol. 1988;5(3):151-69.
11. Iselius L, Slack J, Littler M, Morton NE. Genetic epidemiology of breast cancer in Britain. Ann Hum Genet. 1991;55(2):151-9.
12. Williams WR, Anderson DE. Genetic epidemiology of breast cancer: segregation analysis of 200 Danish pedigrees. Genet Epidemiol. 1984;1(1):7-20.
13. Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, et al. Linkage of early-onset familial breast cancer to chromosome 17q21. Science. 1990;250(4988):1684-9.
14. Wooster R, Neuhausen SL, Mangion J, Quirk Y, Ford D, Collins N, et al. Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. Science. 1994;265(5181):2088-90.
15. Garcia-Closas M, Malats N, Silverman D, Dosemeci M, Kogevinas M, Hein DW, et al. NAT2 slow acetylation, GSTM1 null genotype, and risk of bladder cancer: results from the Spanish Bladder Cancer Study and meta-analyses. Lancet. 2005;366(9486):649-59.
16. Lindstrom S, Ma J, Altshuler D, Giovannucci E, Riboli E, Albanes D, et al. A large study of androgen receptor germline variants and their relation to sex hormone levels and prostate cancer risk. Results from the National Cancer Institute Breast and Prostate Cancer Cohort Consortium. J Clin Endocrinol Metab. 2010;95(9):E121-7.
17. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature. 2007;447(7148):1087-93.
18. Michailidou K, Beesley J, Lindstrom S, Canisius S, Dennis J, Lush MJ, et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nat Genet. 2015;47(4):373-80.
19. 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. Nat Genet. 2013;45(4):353-61, 61e1-2.
20. Michailidou K, Lindstrom S, Dennis J, Beesley J, Hui S, Kar S, et al. Association analysis identifies 65 new breast cancer risk loci. Nature. 2017;551(7678):92-4.
21. Al Olama AA, Kote-Jarai Z, Berndt SI, Conti DV, Schumacher F, Han Y, et al. A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. Nat Genet. 2014;46(10):1103-9.
22. Amundadottir LT, Sulem P, Gudmundsson J, Helgason A, Baker A, Agnarsson BA, et al. A common variant associated with prostate cancer in European and African populations. Nat Genet. 2006;38(6):652-8.
23. Eeles RA, Kote-Jarai Z, Giles GG, Olama AA, Guy M, Jugurnauth SK, et al. Multiple newly identified loci associated with prostate cancer susceptibility. Nat Genet. 2008;40(3):316-21.
24. Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat Genet. 2008;40(5):616-22.

25. McKay JD, Hung RJ, Han Y, Zong X, Carreras-Torres R, Christiani DC, et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nat Genet. 2017;49(7):1126-32.
26. Wang Y, Broderick P, Matakidou A, Eisen T, Houlston RS. Chromosome 15q25 (CHRNA3-CHRNA5) variation impacts indirectly on lung cancer risk. PLoS One. 2011;6(4):e19085.
27. Wang Y, McKay JD, Rafnar T, Wang Z, Timofeeva MN, Broderick P, et al. Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. Nat Genet. 2014;46(7):736-41.
28. Houlston RS, Webb E, Broderick P, Pittman AM, Di Bernardo MC, Lubbe S, et al. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat Genet. 2008;40(12):1426-35.
29. Orlando G, Law PJ, Palin K, Tuupanen S, Gylfe A, Hanninen UA, et al. Variation at 2q35 (PNKD and TMBIM1) influences colorectal cancer risk and identifies a pleiotropic effect with inflammatory bowel disease. Hum Mol Genet. 2016;25(11):2349-59.
30. Schumacher FR, Schmit SL, Jiao S, Edlund CK, Wang H, Zhang B, et al. Genome-wide association study of colorectal cancer identifies six new susceptibility loci. Nat Commun. 2015;6:7138.
31. Tomlinson IP, Webb E, Carvajal-Carmona L, Broderick P, Howarth K, Pittman AM, et al. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. Nat Genet. 2008;40(5):623-30.
32. Abnet CC, Freedman ND, Hu N, Wang Z, Yu K, Shu XO, et al. A shared susceptibility locus in PLCE1 at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. Nat Genet. 2010;42(9):764-7.
33. Helgason H, Rafnar T, Olafsdottir HS, Jonasson JG, Sigurdsson A, Stacey SN, et al. Loss-of-function variants in ATM confer risk of gastric cancer. Nat Genet. 2015;47(8):906-10.
34. Pharoah PD, Tsai YY, Ramus SJ, Phelan CM, Goode EL, Lawrenson K, et al. GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. Nat Genet. 2013;45(4):362-70, 70e1-2.
35. Childs EJ, Mocci E, Campa D, Bracci PM, Gallinger S, Goggins M, et al. Common variation at 2p13.3, 3q29, 7p13 and 17q25.1 associated with susceptibility to pancreatic cancer. Nat Genet. 2015;47(8):911-6.
36. Petersen GM, Amundadottir L, Fuchs CS, Kraft P, Stolzenberg-Solomon RZ, Jacobs KB, et al. A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. Nat Genet. 2010;42(3):224-8.
37. Wolpin BM, Rizzato C, Kraft P, Kooperberg C, Petersen GM, Wang Z, et al. Genome-wide association study identifies multiple susceptibility loci for pancreatic cancer. Nat Genet. 2014;46(9):994-1000.
38. Lilyquist J, Ruddy KJ, Vachon CM, Couch FJ. Common Genetic Variation and Breast Cancer Risk-Past, Present, and Future. Cancer Epidemiol Biomarkers Prev. 2018;27(4):380-94.
39. Mavaddat N, Pharoah PD, Michailidou K, Tyrer J, Brook MN, Bolla MK, et al. Prediction of breast cancer risk based on profiling with common genetic variants. J Natl Cancer Inst. 2015;107(5).
40. 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. Int J Epidemiol. 2018;47(2):526-36.
41. Vachon CM, Pankratz VS, Scott CG, Haeberle L, Ziv E, Jensen MR, et al. The contributions of breast density and common genetic variation to breast cancer risk. J Natl Cancer Inst. 2015;107(5).
42. Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. Am J Hum Genet. 2019;104(1):21-34.
43. Kuchenbaecker KB, McGuffog L, Barrowdale D, Lee A, Soucy P, Dennis J, et al. Evaluation of Polygenic Risk Scores for Breast and Ovarian Cancer Risk Prediction in BRCA1 and BRCA2 Mutation Carriers. J Natl Cancer Inst. 2017;109(7).
44. Noto H, Goto A, Tsujimoto T, Osame K, Noda M. Latest insights into the risk of cancer in diabetes. J Diabetes Investig. 2013;4(3):225-32.
45. Carpenter CL, Ross RK, Paganini-Hill A, Bernstein L. Effect of family history, obesity and exercise on breast cancer risk among postmenopausal women. Int J Cancer. 2003;106(1):96-102.
46. Song X, Pukkala E, Dyba T, Tuomilehto J, Moltchanov V, Mannisto S, et al. Body mass index and cancer incidence: the FINRISK study. Eur J Epidemiol. 2014;29(7):477-87.
47. Lu L, Risch H, Irwin ML, Mayne ST, Cartmel B, Schwartz P, et al. Long-term overweight and weight gain in early adulthood in association with risk of endometrial cancer. Int J Cancer. 2011;129(5):1237-43.

48. Arslan AA HK, Kooperberg C, Patel AV, et al. Anthropometric measures, body mass index, and pancreatic cancer: a pooled
analysis from the Pancreatic Cancer Cohort Consortium (PanScan). Arch Intern Med. 2010.
49. Michaud DS GE, Willett WC, Colditz GA, Stampfer MJ, Fuchs CS. Physical activity, obesity, height, and the risk of pancreatic cancer. JAMA. 2001.
50. Bhaskaran K, Douglas I, Forbes H, dos-Santos-Silva I, Leon DA, Smeeth L. Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5·24 million UK adults. The Lancet. 2014;384(9945):755-65.
51. Smith L, Brinton LA, Spitz MR, Lam TK, Park Y, Hollenbeck AR, et al. Body mass index and risk of lung cancer among never, former, and current smokers. J Natl Cancer Inst. 2012;104(10):778-89.
52. Baer HJ, Tworoger SS, Hankinson SE, Willett WC. Body fatness at young ages and risk of breast cancer throughout life. Am J Epidemiol. 2010;171(11):1183-94.
53. Morimoto LM WE, Chen Z, Chlebowski RT, Hays J, Kuller L, Lopez AM,, Manson J MK, Muti PC, Stefanick ML, McTiernan A. Obesity, body size, and risk of postmenopausal breast cancer: the Women's Health Initiative (United States). Cancer Causes Control. 2002.
54. White AJ, Nichols HB, Bradshaw PT, Sandler DP. Overall and central adiposity and breast cancer risk in the sister study. Cancer. 2015.
55. Lubin F. Body Mass Index at Age 18 Years and during Adult Life and Ovarian Cancer Risk. American Journal of Epidemiology. 2003;157(2):113-20.
56. Stoll BA. Obesity and breast cancer. Int J Obes Relat Metab Disord. 1996;20(5):389-92.
57. Hollmann M, Runnebaum B, Gerhard I. Impact of waist-hip-ratio and body-mass-index on hormonal and metabolic parameters in young, obese women. Int J Obes Relat Metab Disord. 1997;21(6):476-83.
58. Borugian MJ. Waist-to-Hip Ratio and Breast Cancer Mortality. American Journal of Epidemiology. 2003;158(10):963-8.
59. Huang Z WW, Colditz GA, Hunter DJ, Manson JE, Rosner B, Speizer FE,, SE H. Waist circumference, waist:hip ratio, and risk of breast cancer in
the Nurses' Health Study. Am J Epidemiol. 1999.
60. Moore LL, Bradlee ML, Singer MR, Splansky GL, Proctor MH, Ellison RC, et al. BMI and waist circumference as predictors of lifetime colon cancer risk in Framingham Study adults. Int J Obes Relat Metab Disord. 2004;28(4):559-67.
61. Wang Y, Jacobs EJ, Patel AV, Rodriguez C, McCullough ML, Thun MJ, et al. A prospective study of waist circumference and body mass index in relation to colorectal cancer incidence. Cancer Causes Control. 2008;19(7):783-92.
62. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum Mol Genet. 2014;23(R1):R89-98.
63. Palmer TM, Sterne JA, Harbord RM, Lawlor DA, Sheehan NA, Meng S, et al. Instrumental variable estimation of causal risk ratios and causal odds ratios in Mendelian randomization analyses. Am J Epidemiol. 2011;173(12):1392-403.
64. Verduijn M, Siegerink B, Jager KJ, Zoccali C, Dekker FW. Mendelian randomization: use of genetics to enable causal inference in observational studies. Nephrol Dial Transplant. 2010;25(5):1394-8.
65. Gordon Fehringer PK, C.A. Haiman, et. al. Rayjean J. Hung. Cross-Cancer genomewide association analysis of cancers in the lung, ovary, breast, prostate and colon among 61,851 cases and 61,820 controls using GAME-ON Network and GECCO data. 2015.
66. Zhang C, Doherty JA, Burgess S, Hung RJ, Lindstrom S, Kraft P, et al. Genetic determinants of telomere length and risk of common cancers: a Mendelian randomization study. Hum Mol Genet. 2015.
67. Horikoshi M, Yaghootkar H, Mook-Kanamori DO, Sovio U, Taal HR, Hennig BJ, et al. New loci associated with birth weight identify genetic links between intrauterine growth and adult height and metabolism. Nat Genet. 2013;45(1):76-82.
68. Janine F. Felix JPB, Claire Monnereau, Ralf J.P. van der Valk, Evie Stergiakouli, Alessandra Chesi, Romy Gaillard, Bjarke Feenstra, Elisabeth Thiering, Eskil Kreiner-Møller, Anubha Mahajan,Fernando Rivadeneira, Hakon Hakonarson, Susan M. Ring, George Davey Smith, Thorkild I.A. Sørensen, Nicholas J. Timpson, Struan F.A. Grant, Vincent W.V. Jaddoe, et. al. for the Early Growth Genetics (EGG) Consortium. Genome-wide association analysis identifies three new susceptibility loci for childhood body mass index. Hum Mol Genet.

69. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015;518(7538):197-206.
70. Heid IM, Jackson AU, Randall JC, Winkler TW, Qi L, Steinthorsdottir V, et al. Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. Nat Genet. 2010;42(11):949-60.
71. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. Genet Epidemiol. 2013;37(7):658-65.
72. VanderWeele TJ, Tchetgen Tchetgen EJ, Cornelis M, Kraft P. Methodological challenges in mendelian randomization. Epidemiology. 2014;25(3):427-35.
73. Al Olama AA, Kote-Jarai Z, Berndt SI, Conti DV, Schumacher F, Han Y, et al. A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. Nat Genet. 2014;46(10):1103-9.
74. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int J Epidemiol. 2015;44(2):512-25.
75. Baer HJ, Colditz GA, Willett WC, Dorgan JF. Adiposity and sex hormones in girls. Cancer Epidemiol Biomarkers Prev. 2007;16(9):1880-8.
76. Baer HJ, Colditz GA, Rosner B, Michels KB, Rich-Edwards JW, Hunter DJ, et al. Body fatness during childhood and adolescence and incidence of breast cancer in premenopausal women: a prospective cohort study. Breast Cancer Res. 2005;7(3):R314-25.
77. Magnusson C BJ, Persson I, Wolk A, Bergström R, Trichopoulos D, Adami HO. Body size in different periods of life and breast cancer risk in post-menopausal women. International Journal of Cancer. 1998.
78. van den Brandt PA SD, Yaun SS, Adami HO, Beeson L, Folsom AR, Fraser G, Goldbohm RA, Graham S, Kushi L, Marshall JR, Miller AB, Rohan T, Smith-Warner SA, Speizer FE, Willett WC, Wolk A, Hunter DJ. Pooled analysis of prospective cohort studies on height, weight, and breast cancer risk. Am J Epidemiol. 2000.
79. Davey Smith G, Sterne JA, Fraser A, Tynelius P, Lawlor DA, Rasmussen F. The association between BMI and mortality using offspring BMI as an indicator of own BMI: large intergenerational mortality study. BMJ. 2009;339:b5043.
80. Yan Guo SWA, Xiao-Ou Shu, Kyriaki Michailidou, Manjeet K. Bolla, Qin Wang, Montserrat Garcia-Closas, et al. Peter Kraft, Paul Pharoah, David Hunter, Douglas F. Easton, Wei Zheng. Body Mass Index and Breast Cancer Risk: Mendelian Randomization Analyses of Data from 145 000 women of European descent [submitted]. Plos Medicine. 2015.
81. Emaus MJ, van Gils CH, Bakker MF, Bisschop CN, Monninkhof EM, Bueno-de-Mesquita HB, et al. Weight change in middle adulthood and breast cancer risk in the EPIC-PANACEA study. Int J Cancer. 2014;135(12):2887-99.
82. Sandholt CH, Allin KH, Toft U, Borglykke A, Ribel-Madsen R, Sparso T, et al. The effect of GWAS identified BMI loci on changes in body weight among middle-aged Danes during a five-year period. Obesity (Silver Spring). 2014;22(3):901-8.
83. B. Rosner AHE, A.T. Toriola, W. Chen, S.E. Hankinson, W.C. Willett, C.S. Berkey, G. Colditz. Weight change across the life-course and breast cancer risk according to receptor classification among pre and postmenopausal women[submitted]. JAMA. 2016.
84. Bardou M, Barkun AN, Martel M. Obesity and colorectal cancer. Gut. 2013;62(6):933-47.
85. Petridou ET, Sergentanis TN, Antonopoulos CN, Dessypris N, Matsoukis IL, Aronis K, et al. Insulin resistance: an independent risk factor for lung cancer? Metabolism. 2011;60(8):1100-6.
86. Kabat GC, Miller AB, Rohan TE. Body Mass Index and Lung Cancer Risk in Women. Epidemiology. 2007;18(5):607-12.
87. Tarnaud C, Guida F, Papadopoulos A, Cenee S, Cyr D, Schmaus A, et al. Body mass index and lung cancer risk: results from the ICARE study, a large, population-based case-control study. Cancer Causes Control. 2012;23(7):1113-26.
88. El-Zein M, Parent ME, Nicolau B, Koushik A, Siemiatycki J, Rousseau MC. Body mass index, lifetime smoking intensity and lung cancer risk. Int J Cancer. 2013;133(7):1721-31.
89. OncoArray Network. http://epigrantscancergov/oncoarray/.
90. McKee SA, Sinha R, Weinberger AH, Sofuoglu M, Harrison EL, Lavery M, et al. Stress decreases the ability to resist smoking and potentiates smoking intensity and reward. J Psychopharmacol. 2011;25(4):490-502.

91. Childs E, de Wit H. Effects of acute psychosocial stress on cigarette craving and smoking. Nicotine Tob Res. 2010;12(4):449-53.
92. Aune D, Navarro Rosenblatt DA, Chan DS, Abar L, Vingeliene S, Vieira AR, et al. Anthropometric factors and ovarian cancer risk: a systematic review and nonlinear dose-response meta-analysis of prospective studies. Int J Cancer. 2015;136(8):1888-98.
93. Baer HJ, Hankinson SE, Tworoger SS. Body size in early life and risk of epithelial ovarian cancer: results from the Nurses' Health Studies. Br J Cancer. 2008;99(11):1916-22.
94. Tworoger SS, Fairfield KM, Colditz GA, Rosner BA, Hankinson SE. Association of oral contraceptive use, other contraceptive methods, and infertility with ovarian cancer risk. Am J Epidemiol. 2007;166(8):894-901.
95. Olsen CM, Green AC, Whiteman DC, Sadeghi S, Kolahdooz F, Webb PM. Obesity and the risk of epithelial ovarian cancer: a systematic review and meta-analysis. Eur J Cancer. 2007;43(4):690-709.
96. HA. R. Hormonal etiology of epithelial ovarian cancer, with a hypothesis
concerning the role of androgens and progesterone. J Natl Cancer Inst. 1998.
97. Keimling M, Renehan AG, Behrens G, Fischer B, Hollenbeck AR, Cross AJ, et al. Comparison of associations of body mass index, abdominal adiposity, and risk of colorectal cancer in a large prospective cohort study. Cancer Epidemiol Biomarkers Prev. 2013;22(8):1383-94.
98. Thrift AP, Gong J, Peters U, Chang-Claude J, Rudolph A, Slattery ML, et al. Mendelian Randomization Study of Body Mass Index and Colorectal Cancer Risk. Cancer Epidemiol Biomarkers Prev. 2015;24(7):1024-31.
99. Harriss DJ, Atkinson G, George K, Cable NT, Reilly T, Haboubi N, et al. Lifestyle factors and colorectal cancer risk (1): systematic review and meta-analysis of associations with body mass index. Colorectal Dis. 2009;11(6):547-63.
100. Burgess S, Davies NM, Thompson SG, Consortium EP-I. Instrumental variable analysis with a nonlinear exposure-outcome relationship. Epidemiology. 2014;25(6):877-85.
101. Burgess S. Sample size and power calculations in Mendelian randomization with a single instrumental variable and a binary outcome. Int J Epidemiol. 2014;43(3):922-9.
102. Thomas DC, Conti DV. Commentary: the concept of 'Mendelian Randomization'. Int J Epidemiol. 2004;33(1):21-5.
103. Michailidou K, Beesley J, Lindstrom S, Canisius S, Dennis J, Lush MJ, et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nat Genet. 2015;47(4):373-80.
104. Spencer EA, Appleby PN, Davey GK, Key TJ. Validity of self-reported height and weight in 4808 EPIC-Oxford participants. Public Health Nutr. 2002;5(4):561-5.
105. Incident Cases of Breast Cancer in 2014. National Cancer
Institute.http://www.cancer.gov/cancertopics/types/breast/. Accessed October 14,
2014.
106. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2017). National Cancer Institute, DCCPS, Surveillance Research Program, released April 2017.
107. Hall P, Easton D. Breast cancer screening: time to target women at risk. Br J Cancer. 2013;108(11):2202-4.
108. Pashayan N, Duffy SW, Chowdhury S, Dent T, Burton H, Neal DE, et al. Polygenic susceptibility to prostate and breast cancer: implications for personalised screening. Br J Cancer. 2011;104(10):1656-63.
109. Burton H, Chowdhury S, Dent T, Hall A, Pashayan N, Pharoah P. Public health implications from COGS and potential for risk stratification and screening. Nat Genet. 2013;45(4):349-51.
110. Parkin DM, Boyd L, Walker LC. 16. The fraction of cancer attributable to lifestyle and environmental factors in the UK in 2010. Br J Cancer. 2011;105 Suppl 2:S77-81.
111. Madigan MP, Ziegler RG, Benichou J, Byrne C, Hoover RN. Proportion of breast cancer cases in the United States explained by well-established risk factors. J Natl Cancer Inst. 1995;87(22):1681-5.
112. Darabi H, Czene K, Zhao W, Liu J, Hall P, Humphreys K. Breast cancer risk prediction and individualised screening based on common genetic variation and breast density measurement. Breast Cancer Res. 2012;14(1):R25.
113. Gail MH. Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. J Natl Cancer Inst. 2008;100(14):1037-41.

114. Husing A, Canzian F, Beckmann L, Garcia-Closas M, Diver WR, Thun MJ, et al. Prediction of breast cancer risk by genetic risk factors, overall and by hormone receptor status. J Med Genet. 2012;49(9):601-8.
115. Wacholder S, Hartge P, Prentice R, Garcia-Closas M, Feigelson HS, Diver WR, et al. Performance of common genetic variants in breast-cancer risk models. N Engl J Med. 2010;362(11):986-93.
116. Maas P, Barrdahl M, Joshi AD, Auer PL, Gaudet MM, Milne RL, et al. Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States. JAMA Oncol. 2016;2(10):1295-302.
117. Garcia-Closas M, Gunsoy NB, Chatterjee N. Combined associations of genetic and environmental risk factors: implications for prevention of breast cancer. J Natl Cancer Inst. 2014;106(11).
118. Colditz GA, Hankinson SE. The Nurses' Health Study: lifestyle and health among women. Nat Rev Cancer. 2005;5(5):388-96.
119. Lindstrom S, Loomis S, Turman C, Huang H, Huang J, Aschard H, et al. A comprehensive survey of genetic variation in 20,691 subjects from four large cohorts. PLoS One. 2017;12(3):e0173997.
120. 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.
121. Cox DR: Regression Models and Life-Tables. Journal of the Royal Statistical Society Series B (Methodological).34:187-220.
122. Parsons VL, Moriarity C, Jonas K, Moore TF, Davis KE, Tompkins L. Design and estimation for the national health interview survey, 2006-2015. Vital Health Stat 2. 2014(165):1-53.
123. Amber W PP, Chi G et al. Prospective Evaluation of Breast Cancer Risk Model Integrating Classical Risk Factors and Polygenic Risk in 15 Cohorts from Six Countries. [Submitted to The Journal of Clinical Oncology].
124. Banegas MP, John EM, Slattery ML, Gomez SL, Yu M, LaCroix AZ, et al. Projecting Individualized Absolute Invasive Breast Cancer Risk in US Hispanic Women. J Natl Cancer Inst. 2017;109(2).
125. Matsuno RK, Costantino JP, Ziegler RG, Anderson GL, Li H, Pee D, et al. Projecting individualized absolute invasive breast cancer risk in Asian and Pacific Islander American women. J Natl Cancer Inst. 2011;103(12):951-61.
126. McCarthy AM, Armstrong K, Handorf E, Boghossian L, Jones M, Chen J, et al. Incremental impact of breast cancer SNP panel on risk classification in a screening population of white and African American women. Breast Cancer Res Treat. 2013;138(3):889-98.
127. Yon Ho J CG, Jihye K et al. . Validating breast-cancer risk prediction models in the Korean Cancer Prevention Study-II Biobank. [submitted to Cancer Epidemiology, Biomarkers & Prevention]. 2020.
128. Lee A, Mavaddat N, Wilcox AN, Cunningham AP, Carver T, Hartley S, et al. BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. Genet Med. 2019;21(8):1708-18.
129. Brouckaert O, Rudolph A, Laenen A, Keeman R, Bolla MK, Wang Q, et al. Reproductive profiles and risk of breast cancer subtypes: a multi-center case-only study. Breast Cancer Res. 2017;19(1):119.
130. Holm J, Eriksson L, Ploner A, Eriksson M, Rantalainen M, Li J, et al. Assessment of Breast Cancer Risk Factors Reveals Subtype Heterogeneity. Cancer Res. 2017;77(13):3708-17.
131. Yang XR, Chang-Claude J, Goode EL, Couch FJ, Nevanlinna H, Milne RL, et al. Associations of breast cancer risk factors with tumor subtypes: a pooled analysis from the Breast Cancer Association Consortium studies. J Natl Cancer Inst. 2011;103(3):250-63.
132. Surveillance, Epidemiology, and End Results (SEER) Program Research Data
133. Bleyer A, Welch HG. Effect of three decades of screening mammography on breast-cancer incidence. N Engl J Med. 2012;367(21):1998-2005.
134. The benefits and harms of breast cancer screening: an independent review. Lancet. 2012;380(9855):1778-86.
135. Yadav S, Couch FJ. Germline Genetic Testing for Breast Cancer Risk: The Past, Present, and Future. Am Soc Clin Oncol Educ Book. 2019;39:61-74.
136. Antoniou AC, Beesley J, McGuffog L, Sinilnikova OM, Healey S, Neuhausen SL, et al. Common breast cancer susceptibility alleles and the risk of breast cancer for BRCA1 and BRCA2 mutation carriers: implications for risk prediction. Cancer Res. 2010;70(23):9742-54.

137. Kuchenbaecker KB, Neuhausen SL, Robson M, Barrowdale D, McGuffog L, Mulligan AM, et al. Associations of common breast cancer susceptibility alleles with risk of breast cancer subtypes in BRCA1 and BRCA2 mutation carriers. Breast Cancer Res. 2014;16(6):3416.
138. Muranen TA, Greco D, Blomqvist C, Aittomaki K, Khan S, Hogervorst F, et al. Genetic modifiers of CHEK2*1100delC-associated breast cancer risk. Genet Med. 2017;19(5):599-603.
139. Calle EE, Rodriguez C, Jacobs EJ, Almon ML, Chao A, McCullough ML, et al. The American Cancer Society Cancer Prevention Study II Nutrition Cohort: rationale, study design, and baseline characteristics. Cancer. 2002;94(9):2490-501.
140. Patel AV, Jacobs EJ, Dudas DM, Briggs PJ, Lichtman CJ, Bain EB, et al. The American Cancer Society's Cancer Prevention Study 3 (CPS-3): Recruitment, study design, and baseline characteristics. Cancer. 2017;123(11):2014-24.
141. Bernstein L, Allen M, Anton-Culver H, Deapen D, Horn-Ross PL, Peel D, et al. High breast cancer incidence rates among California teachers: results from the California Teachers Study (United States). Cancer Causes Control. 2002;13(7):625-35.
142. Kolonel LN, Henderson BE, Hankin JH, Nomura AM, Wilkens LR, Pike MC, et al. A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. Am J Epidemiol. 2000;151(4):346-57.
143. MV C. Mayo Mammography Health Study
144. Rice MS, Eliassen AH, Hankinson SE, Lenart EB, Willett WC, Tamimi RM. Breast Cancer Research in the Nurses' Health Studies: Exposures Across the Life Course. Am J Public Health. 2016;106(9):1592-8.
145. Bao Y, Bertoia ML, Lenart EB, Stampfer MJ, Willett WC, Speizer FE, et al. Origin, Methods, and Evolution of the Three Nurses' Health Studies. Am J Public Health. 2016;106(9):1573-81.
146. Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. Control Clin Trials. 1998;19(1):61-109.
147. Sandler DP, Hodgson ME, Deming-Halverson SL, Juras PS, D'Aloisio AA, Suarez LM, et al. The Sister Study Cohort: Baseline Methods and Participant Characteristics. Environ Health Perspect. 2017;125(12):127003.
148. Ambrosone CB, Ciupak GL, Bandera EV, Jandorf L, Bovbjerg DH, Zirpoli G, et al. Conducting Molecular Epidemiological Research in the Age of HIPAA: A Multi-Institutional Case-Control Study of Breast Cancer in African-American and European-American Women. J Oncol. 2009;2009:871250.
149. Kelemen LE, Couch FJ, Ahmed S, Dunning AM, Pharoah PD, Easton DF, et al. Genetic variation in stromal proteins decorin and lumican with breast cancer: investigations in two case-control studies. Breast Cancer Res. 2008;10(6):R98.
150. Trentham-Dietz A, Sprague BL, Hampton JM, Miglioretti DL, Nelson HD, Titus LJ, et al. Modification of breast cancer risk according to age and menopausal status: a combined analysis of five population-based case-control studies. Breast Cancer Res Treat. 2014;145(1):165-75.
151. Lange V, Bohme I, Hofmann J, Lang K, Sauter J, Schone B, et al. Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. BMC Genomics. 2014;15:63.
152. Hu C, Hart SN, Polley EC, Gnanaolivu R, Shimelis H, Lee KY, et al. Association Between Inherited Germline Mutations in Cancer Predisposition Genes and Risk of Pancreatic Cancer. Jama. 2018;319(23):2401-9.
153. Buys SS, Sandbach JF, Gammon A, Patel G, Kidd J, Brown KL, et al. A study of over 35,000 women with breast cancer tested with a 25-gene panel of hereditary cancer genes. Cancer. 2017;123(10):1721-30.
154. Couch FJ, Shimelis H, Hu C, Hart SN, Polley EC, Na J, et al. Associations Between Cancer Predisposition Testing Panel Genes and Breast Cancer. JAMA Oncol. 2017;3(9):1190-6.
155. Kurian AW, Li Y, Hamilton AS, Ward KC, Hawley ST, Morrow M, et al. Gaps in Incorporating Germline Genetic Testing Into Treatment Decision-Making for Early-Stage Breast Cancer. J Clin Oncol. 2017;35(20):2232-9.
156. Susswein LR, Marshall ML, Nusbaum R, Vogel Postula KJ, Weissman SM, Yackowski L, et al. Pathogenic and likely pathogenic variant prevalence among the first 10,000 patients referred for next-generation cancer panel testing. Genet Med. 2016;18(8):823-32.
157. van Buuren S G-OK. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software.45(3):1-67.
158. Tibshirani R. Regression Shrinkage and Selection via the Lasso. JSTOR. 1996;58:267-88.

159. Dupont WD. Converting relative risks to absolute risks: a graphical approach. Stat Med. 1989;8(6):641-51.
160. American Cancer Society recommendations for early breast cancer detection in women without breast symptoms. American Cancer Society.http://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/american-cancer-society-recommendations-for-the-early-detection-of-breast-cancer.html.
161. Bevers TB, Helvie M, Bonaccio E, Calhoun KE, Daly MB, Farrar WB, et al. Breast Cancer Screening and Diagnosis, Version 3.2018, NCCN Clinical Practice Guidelines in Oncology. J Natl Compr Canc Netw. 2018;16(11):1362-89.
162. Tung N, Domchek SM, Stadler Z, Nathanson KL, Couch F, Garber JE, et al. Counselling framework for moderate-penetrance cancer-susceptibility mutations. Nat Rev Clin Oncol. 2016;13(9):581-8.
163. Feuer EJ, Wun LM, Boring CC, Flanders WD, Timmel MJ, Tong T. The lifetime risk of developing breast cancer. J Natl Cancer Inst. 1993;85(11):892-7.
164. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019;51(4):584-91.