# Risk Factors for Breast Cancer: An In-depth Analysis of UK Biobank Data

CID: 01913993, 01814471, 01808753, 01333689, 01377486, 01902920
Imperial College London

## ABSTRACT

**Introduction**: Breast cancer is one of the foremost cancers in women and the second most common cancer in the world. As the incidence of breast cancer in women has kept on increment in the past two decades, exploring risk factors for breast cancer has been progressively recognised by authorities and researchers. Risk factors of breast cancer have been shown to be multi-factorial. In this study, joint effects and co-occurring patterns between predictors and breast cancer will be explored.

**Methods**: The UK Biobank dataset was used to study 43 variables of 8 categories in our study cohort, including demographic, socioeconomic, lifestyle, health risk, medical history, female-specific, biomarkers and environmental exposure. We used principal component analysis to recognize major dietary patterns of diet items. Univariate and multivariate logistic regression models were used to measure the risk for each covariate to the development of breast cancer and its subtypes. We additionally accounted for correlation across covariates using logistic LASSO regression. To evaluate joint effects on breast cancer and its subtypes, we also performed a series of PLS-DA to select the most relevant set of covariates with relation to breast cancer status. Furthermore, to validate for our findings, sensitivity analysis was conducted by family history of breast cancer, stillbirth situation, menopause status, oral contraceptive use and age at menarche.

**Results**: Consistent associations have been found across different methods. Among all models, variables recruitment age, family history of breast cancer, and hypertension were positively associated with the future risk of female breast cancer and its subtypes (inner quadrant and outer quadrant), while Black and Other ethnicities showed a protective effect. In sgPLS-DA, demographic and female-specific variables were associated in pooled and outer subtype; demographic and health risk factors were selected for the inner subtype. Lastly, sensitivity analysis demonstrated no substantial mismatch between the stratified study population and outcome.

**Conclusion**: The study found that post-menopausal, older, hypertensive, first-degree family history contributes to an increased risk of breast cancer development. Our research provides a theoretical basis for risk factors for breast cancer and contributes to the understanding of the pathogenesis of breast cancer.

## 1 INTRODUCTION

Breast cancer is one of the most common types of cancer amongst women and the second most commonly occurring cancer worldwide. In total, there were approximately 2.1 million diagnoses recorded in 2018, contributing to over 11.6% incidence burden of all the cancer types, according to the latest data from the International Agency for Research on Cancer [1]. The breast cancer incidence is relatively higher in developed countries such as Europe and Northern America [2]. Female breast cancer contributes to around 627,000 deaths in 2018 globally, which ranks as the fifth leading cause of death (6.6%)[3]. More than 80% of the female breast cancer incidence was diagnosed at middle-aged in the UK (aged 50+). With the consistent growing trend of breast cancer incidence in women in recent years [4], identifying risk factors of breast cancer becomes increasingly crucial.

Contribution of risk factors on breast cancer have been shown to be multi-factorial. Some factors identified in the previous studies were related to population structure (age, ethnicity), environmental exposures, lifestyle, reproductive factors (menopause status, exogenous hormone intake) and the accumulated effect of anthropometric factors (body mass index, waist circumference) [5]. However, existing research were reported on relatively limited sample sizes, revealing few definite effects using linear mixed models, and left a broad scope of uncertainty. Our study will estimate associations of risk factors and biological markers using penalised methods and partial least squares models along with calibration procedures to identify sets of variables associated with future risk of breast cancer and its subtypes (inner quadrant and outer quadrant) in the UK Biobank cohort.

## 2 METHODS

### 2.1 Study population and data sources

The UK Biobank is a large prospective cohort consisting of 502,536 participants aged 37-73 (recruited across England, Wales and Scotland between 2006 and 2010). Comprehensive data were provided by each participant on a broad range of demographic, clinical, lifestyle, environmental, genetic and social outcomes, including clinical measurements using standardised protocols and face-to-face interviews at 22 assessment centres.[6] According to Public Health England statistics in 2017, women aged between 50 and 69 contributed to 50% of the total incidence of breast cancer cases. Although males could also develop breast cancer, the total incidence is considerably lower than in females (1:99 ratio) [3]. Thus, we put our focus solely on the female population in this study.

Of the 502,537 participants, we excluded (i) 31 non-consenting participants, (ii) 229,122 males, (iii) 74,105 participants diagnosed with another type of cancer and benign neoplasms, (iv) 33,208 individuals with missing menopause status and hysterectomy status, (v) 3,349 prevalent cases of diagnosed breast cancer before recruitment and within one year of entry, and (vi) 6,266 participants with missing three or four of these variables: average total household income, exposed to breastfeeding as a baby, exposed to maternal smoking around birth and family breast cancer history, leaving 156,456 participants for the analysis. The missing proportion of all variables was less than 15%, thereby justifying the use of imputation on the remaining missing values.

## 2.2 Outcome variable

With linkage to the Hospital Episode Statistics data (HES), we identified 4,228 cases of malignant neoplasms of the breast using C50 from ICD-10 and 179 from ICD-9. Further stratification was conducted by categorising inner quadrant subtype (C50.2 upper-inner quadrant and C50.3 lower-inner quadrant, n = 624) and outer quadrant subtype (C50.4 upper-outer quadrant and C50.5 lower-outer quadrant, n = 1225).

## 2.3 Predictors

A total of 43 variables have been studied in our study cohort and could be categorised into 8 groups: demographic, socioeconomic, lifestyle, health risk, medical history, female-specific, biomarkers and environmental exposure. Demographic variables included age recorded at the time of study recruitment and ethnicity defined by the following categories: White, Black, Asian and other ethnic groups.

Socioeconomic variables included Townsend deprivation index, education levels measured by three categories in descending order: High (College or University degree), Intermediate (A/AS levels, O levels/GCSE, other equivalent professional qualifications) and Low (none of the above). Housing information included types of accommodation (house, flat, other facilities), number of individuals living in a household ($< 3$, $\geq 3$), whether accommodation was owned or rented, average household income (<£18,000; £18,000–30,999; £31,000–51,999; >£52,000) and current employment status (unemployed, employed, retired).

Lifestyle factors consisted of smoking status (current, previous, never), alcohol drinking habits (current, previous, never), average sleep duration (<7 hours, 7-8 hours, >8 hours), frequency of insomnia (rarely, sometimes, usually), physical activity status (days of moderate activity, days of vigorous activity), exposed to breastfeeding as baby, exposed to maternal smoking around birth and dietary information summarised by three principal components.

Health-risk factors included physical measurements of waist circumference in cm and BMI in $kg/m^2$(<25, 25-30, 30-40, >40).

Disease status of comorbidities diagnosed before recruitment was derived from the linked HES database for cardiovascular diseases, hypertension, diabetes and respiratory diseases.

Female-specific variables included a series of reproductive and breast cancer related factors, consisting of menarche age, menopause status, family history of breast cancer, number of live births (<1, 1-3 births, >3), history of stillbirth or birth terminations, history of oral contraceptive use and history of hormone replacement therapy.

Biomarker levels were collected from triglycerides, glycated haemoglobin (HbA1c), glucose, high density lipoprotein (HDL), sex hormone-binding globulin (SHBG) and insulin-like growth factor (IGF-1). Environmental exposures were assessed using air quality and noise pollution, including levels of Nitrogen oxides, PM10, PM2.5 and average sound level of 24 hours.

## 2.4 Statistical analysis

We provided means (± standard deviation) and proportions for descriptive analysis. Categorically variables were assessed using Chi-squared test and T-test was used to compare continuous variables. Categorical variables in the dataset had undergone one-hot encoding with the baseline column for each dummy variable removed to act as a reference category for subsequent analysis.

The principal component analysis (PCA) was used to recognise major dietary patterns with regards to 18 food items. PCA is often regarded as an exploratory method to derive dietary patterns from habitual diets, which can characterise food intake with different levels of food consumption [7]. Principal components were chosen by the factor loadings and the scree plot. Foods with loadings above $|\geq 0.3|$ on a component were considered to contribute significantly to the dietary pattern [8]. Dietary patterns referred to the factors with eigenvalues >1.0 and they were determined by the break in the scree plot and the interpretability of each identified pattern [9]. Factor scores were calculated for each of the derived patterns by summing the products of the observed consumption frequency and the factor loading for each of the significant food groups [10].

We used univariate and multivariate logistic regression models to measure the risk for each covariate to the development of breast cancer (breast cancer cases vs controls) and its subtypes (inner breast cancer and outer breast cancer). In order to account for potential confounding, we further adjusted on recruitment age and BMI. Continuous variables were standardised to ensure comparability of Odds Ratios (ORs), per standard deviation increase.

We additionally accounted for correlation across covariates using logistic LASSO regression, calibrated using Area under the ROC Curve (AUC) with cross-validation. This method identified a parsimonious set of variables and estimated joint effects explaining risks of breast cancer and its subtypes. In order to measure prediction performance using ROC Curve, we used 80% subsamples as training sets and calculated the AUC in the remaining 20% test sets.

To evaluate the joint effect on breast cancer and its histological subtypes, we also performed a series of Partial Least Squares – Discriminant Analysis (PLS-DA) to select the most relevant set of covariates with relation to breast cancer status. The first component was used for PLS-DA and penalty calibration was conducted on the variables to induce sparsity on the covariates in each subsequent model. Calibrations for each model were conducted using 5-folds cross validation over 100 iterations, selecting the optimal parameters that result in the lowest misclassification rate in sPLS-DA with an additional penalty parameter in sgPLS-DA. Due to major class imbalance, downsampling algorithm from the ROSE package [11] was applied to the majority class to aid the process of sparsity calibration for each sPLS-DA model. To explore the consistency of variable selection in each sPLS-DA analysis, we further conducted stability analyses by sub-sampling 80% of the participants over 100 iterations, the threshold of stability selection would be equivalent to the number of variables calibrated in each PLS-DA model and the variables that were considered stably selected would have a selection proportion of over 80% at the cut-off. In addition, in order to capture the effect of functional groupings in the model, group PLS-DA (gPLS-DA) and sparse group PLS-DA (sgPLS-DA) were conducted to implement prior knowledge of group membership on the covariates. Variables from the medical history category had been combined into the health risk group in sgPLS analysis to provide a more informative grouping structure, hence a total of 7 categories were applied.

To account for multiple confounding, we sequentially adjusted for (i) demographics, (ii) social factors, (iii) health risk factors,

(iv)female-specific factors, (v) lifestyle factors, (vi) medical risk factors, (vii) environmental factors, and (viii) biomarkers. Sensitivity analysis was conducted by separately considering models with stratification on family history of breast cancer, history of stillbirth , menopause status, history of oral contraceptives use and menarche age (early, normal, late). All analyses were performed in R, version 4.0.3.

# 3 RESULTS

## 3.1 Dietary pattern

With the use of PCA analysis, three principal components (PCs) were generated from 18 food groups as shown in Table 1. Table 1 also presents the factor-loading matrices. PC1 presented meat dietary pattern, which was identified with high consumption of processed meat intake, poultry intake, beef intake, lamb intake and pork intake. PC2 represented a healthier dietary pattern, which was identified with high intakes of cooked vegetable, raw vegetable, fresh fruit, dried fruit, oily fish, non-oily fish and water. PC3 illustrated an relatively unhealthy dietary pattern with high intake on beverages, which was associated with more coffee and alcohol intake and less tea and cereal intake. In total, 32% of the variances were explained by these dietary structures[supplementary figure 12], which was in line with other studies [12]. We replaced 18 diet items with three PCs in the following analysis.

## 3.2 Participants characteristics

Descriptive statistics comparing those diagnosed breast cancer after 1 year of study entry (n = 4228) and healthy controls (n = 152228) from our study population (N = 156456) were shown in Table 2. For each variable, the differences between the case and control populations were evaluated using a student's t-test for continuous variables, and chi-squared test for categorical variables.

**Table 1: Factor loadings of major dietary patterns**

**Standardized loadings (pattern matrix) based upon correlation matrix**

| | PC1 | PC2 | PC3 |
|---|---|---|---|
| cooked_vegetable_intake | -0.18 | **0.51** | 0.10 |
| raw_vegetable_intake | -0.24 | **0.54** | 0.17 |
| fresh_fruit_intake | -0.26 | **0.51** | -0.06 |
| dried_fruit_intake | -0.21 | **0.36** | -0.08 |
| oily_fish_intake | 0.03 | **0.59** | -0.03 |
| non.oily_fish_intake | 0.17 | **0.50** | -0.06 |
| processed_meat_intake | **0.64** | -0.05 | -0.03 |
| poultry_intake | **0.56** | 0.26 | -0.06 |
| beef_intake | **0.70** | 0.12 | 0.00 |
| lamb_intake | **0.63** | 0.24 | -0.02 |
| pork_intake | **0.67** | 0.16 | 0.01 |
| cheese_intake | 0.04 | -0.16 | 0.25 |
| bread_intake | 0.14 | -0.23 | -0.12 |
| cereal_intake | -0.06 | 0.10 | **-0.33** |
| tea_intake | 0.07 | -0.04 | **-0.70** |
| coffee_intake | 0.10 | -0.05 | **0.69** |
| water_intake | -0.23 | **0.40** | 0.07 |
| alcohol_intake_frequency | 0.20 | 0.04 | **0.38** |

## 3.3 Univariate logistic regression analysis

The univariate logistic regression analysis indicated significant associations between breast cancer and SHBG levels (OR=0.94, 95% CI [0.92, 0.97]), Black ethnicity (OR=0.53, 95% CI [0.39, 0.71]; Black vs. White), menopause status (OR=1.41, 95% CI [1.31, 1.51]; yes vs. no), diet PC1 (OR=1.06, 95% CI [1.03, 1.09], cardiovascular (OR=1.15, 95% CI [1.07, 1.24]; yes vs. no), family history (OR=1.53, 95% CI [1.38, 1.68]; yes vs. no), hypertension (OR=1.33, 95% CI [1.24, 1.42]; yes vs. no), retirement (OR=1.29, 95%CI [1.14, 1.45]; retired vs. unemployed), and number in household (OR=0.79, 95% CI [0.74, 0.84]; ≥3 vs. 1 2) at the significance level with Bonferroni correction. Model with further adjustment on age and BMI revealed that Black ethnicity (OR=0.58, 95% CI [0.43, 0.79]; Black vs. White), number of live births (OR=0.84, 95% CI [0.77, 0.90]; 1 3 vs. 0 and OR=0.74, 95% CI [0.64, 0.85]; >3 vs. 0), current alcohol drinker (OR=1.29, 95% CI [1.12, 1.49]; current vs. never), family history of breast cancer (OR=1.49, 95% CI [1.35, 1.64]; yes vs. no), hypertension (OR=1.22, 95% CI [1.14, 1.31]; yes vs. no) and average total household income between 31,000 51,999 (OR=1.18, 95% CI [1.08, 1.29]; 31,000 51,999 vs. <18,000) were significantly associated with breast cancer. (figure 1 and Supplementary figure 13) For inner and outer subtypes of breast cancer, family history (OR=1.62, 95% CI [1.27, 2.06] and OR=1.43, 95% CI [1.20, 1.72] respectively; yes vs. no) and hypertension (OR=1.48, 95% CI [1.23, 1.77] and OR=1.35, 95% CI [1.19, 1.53] respectively; yes vs. no) had significant associations with the subtypes. (figure 2-3 and Supplementary figure 14-15)

## 3.4 Multiple logistic regression analysis

In the fully adjusted model [Figure 4], variables independently associated with the breast cancer with significance with odds ratio (OR) ≥ 1.05 or ≤ 0.9 were [supplementary table 3] :Black or other vs White ethnicity (OR ≤ 0.9, p < 10-2); sleep duration <7 vs 7-8 (OR = 0.88, 95% CI [0.81–0.95], p = $2.5 \times 10^{-3}$); average total household income 31,000 51,999 vs <18,000 (OR = 1.11, 95% CI [1.00–1.21], p = $4.410^{-2}$); live births number 1 3 or >3 vs 0 (OR ≤ 0.9, p < $10^{-2}$); BMI 30-40 or >40 vs <25 (OR ≤ 0.9, p < $1.02 \times 10^{-2}$); family history (yes vs no) (OR = 1.47, 95% CI [1.34–1.62], p < $10^{-13}$); hypertension (yes vs no) (OR = 1.26, 95% CI [1.16–1.36], p < $10^{-8}$); respiratory (yes vs no) (OR = 1.09, 95% CI [1.01–1.17], p =$2.71 \times 10^{-2}$ ). When the model was sequentially adjusted, the strengths of the associations between the risk of breast cancer with the biomarkers and the environmental exposures were attenuated. Of the associations identified in the univariate analyses for risk of breast cancer, "Black or other vs White ethnicity", "family history (yes vs no)", and "hypertension (yes vs no)" were consistently identified significant in the fully adjusted model.

## 3.5 Logistic lasso analysis

We used logistic LASSO models to account for correlation and joint contribution across covariates on overall breast cancer and its subtypes. These models with lambda.1se selected 16 variables for overall breast cancer (Figure 5 and Supplementary 16), 22 variables for inner quadrant subtype (Figure 6 and Supplementary 17) and 16 variables for outer quadrant subtype (Figure 7 and Supplementary 18). Among selected variables, family history of breast cancer ($\beta$ =0.362), Black and others ethnicity ($\beta_{Black}$ = -0.202, $\beta_{other}$ = -0.224), hypertension ($\beta$ =0.171) and age ($\beta$ =0.116) contributed largely to overall breast cancer with AUC of 0.561(Figure 8); Black and other ethnicity ($\beta_{Black}$ = -0.559, $\beta_{other}$ = -0.375), family history of breast

**Table 2: Characteristics of the control and case populations for breast cancer in the UK Biobank study**

| | Control(N=152228) | Case(N=4228) | P-value | Overall(N=156456) |
|---|---|---|---|---|
| **Demographic** | | | | |
| **Age at recruitment (years)** | | | | |
| Mean (SD) | 55.8 (8.26) | 57.2 (7.65) | **<0.001** | 55.8 (8.24) |
| Median [Min, Max] | 57.0 [40.0, 71.0] | 59.0 [40.0, 70.0] | | 57.0 [40.0, 71.0] |
| **Ethnicity** | | | | |
| White | 142744 (93.8%) | 4058 (96.0%) | **<0.001** | 146802 (93.8%) |
| Asian | 3079 (2.0%) | 69 (1.6%) | | 3148 (2.0%) |
| Black | 2864 (1.9%) | 43 (1.0%) | | 2907 (1.9%) |
| Others | 3541 (2.3%) | 58 (1.4%) | | 3599 (2.3%) |
| **Waist circumference (cm)** | | | | |
| Mean (SD) | 84.5 (12.5) | 85.0 (12.0) | 0.515 | 84.5 (12.5) |
| Median [Min, Max] | 83.0 [20.0, 171] | 84.0 [56.0, 137] | | 83.0 [20.0, 171] |
| **Environmental** | | | | |
| **Nitrogen oxides (g/m^3)** | | | | |
| Mean (SD) | 44.2 (15.3) | 43.8 (15.5) | 0.988 | 44.2 (15.3) |
| Median [Min, Max] | 42.7 [19.7, 266] | 42.0 [19.7, 242] | | 42.6 [19.7, 266] |
| **PM10 (g/m^3)** | | | | |
| Mean (SD) | 16.3 (1.88) | 16.2 (1.91) | 0.555 | 16.3 (1.88) |
| Median [Min, Max] | 16.1 [11.8, 30.7] | 16.0 [11.8, 25.5] | | 16.1 [11.8, 30.7] |
| **PM2.5 (g/m^3)** | | | | |
| Mean (SD) | 10.0 (1.04) | 9.98 (1.05) | 0.793 | 10.0 (1.04) |
| Median [Min, Max] | 9.96 [8.17, 21.3] | 9.94 [8.17, 18.3] | | 9.96 [8.17, 21.3] |
| **24-hour noise level** | | | | |
| Mean (SD) | 56.1 (4.26) | 56.0 (4.18) | 0.938 | 56.1 (4.25) |
| Median [Min, Max] | 54.9 [51.5, 86.1] | 55.0 [51.6, 85.9] | | 54.9 [51.5, 86.1] |
| **Female-specific** | | | | |
| **Age when periods started (years)** | | | | |
| Mean (SD) | 13.0 (1.62) | 13.0 (1.60) | 0.049 | 13.0 (1.62) |
| Median [Min, Max] | 13.0 [5.00, 25.0] | 13.0 [6.00, 22.0] | | 13.0 [5.00, 25.0] |
| **Had menopause** | | | | |
| No | 46579 (30.6%) | 1008 (23.8%) | **<0.001** | 47587 (30.4%) |
| Yes | 105649 (69.4%) | 3220 (76.2%) | | 108869 (69.6%) |
| **Number of live births** | | | | |
| 0 | 27614 (18.1%) | 847 (20.0%) | 0.003 | 28461 (18.2%) |
| 1~3 | 114823 (75.4%) | 3137 (74.2%) | | 117960 (75.4%) |
| >3 | 9791 (6.4%) | 244 (5.8%) | | 10035 (6.4%) |
| **Ever had stillbirth** | | | | |
| No | 100869 (66.3%) | 2867 (67.8%) | 0.037 | 103736 (66.3%) |
| Yes | 51359 (33.7%) | 1361 (32.2%) | | 52720 (33.7%) |
| **Family breast cancer history** | | | | |
| No | 140822 (92.5%) | 3763 (89.0%) | **<0.001** | 144585 (92.4%) |
| Yes | 11406 (7.5%) | 465 (11.0%) | | 11871 (7.6%) |
| **Ever taken oral contraceptive pill** | | | | |
| No | 28225 (18.5%) | 788 (18.6%) | 0.889 | 29013 (18.5%) |
| Yes | 124003 (81.5%) | 3440 (81.4%) | | 127443 (81.5%) |
| **Ever used HRT** | | | | |
| No | 99402 (65.3%) | 2657 (62.8%) | 0.001 | 102059 (65.2%) |
| Yes | 52826 (34.7%) | 1571 (37.2%) | | 54397 (34.8%) |
| **Health risks** | | | | |
| **Breastfed as a baby** | | | | |
| No | 46209 (30.4%) | 1190 (28.1%) | 0.002 | 47399 (30.3%) |
| Yes | 106019 (69.6%) | 3038 (71.9%) | | 109057 (69.7%) |
| **Maternal smoking** | | | | |
| No | 108671 (71.4%) | 3064 (72.5%) | 0.129 | 111735 (71.4%) |
| Yes | 43557 (28.6%) | 1164 (27.5%) | | 44721 (28.6%) |
| **BMI** | | | | |
| <25 | 61348 (40.3%) | 1652 (39.1%) | 0.045 | 63000 (40.3%) |
| 25-30 | 55773 (36.6%) | 1609 (38.1%) | | 57382 (36.7%) |
| 30-40 | 31448 (20.7%) | 886 (21.0%) | | 32334 (20.7%) |
| >40 | 3659 (2.4%) | 81 (1.9%) | | 3740 (2.4%) |
| **Lifestyles** | | | | |
| **Days of moderate activity (days/Week)** | | | | |
| <3 | 53252 (35.0%) | 1541 (36.4%) | 0.051 | 54793 (35.0%) |
| >=3 | 98976 (65.0%) | 2687 (63.6%) | | 101663 (65.0%) |
| **Days of vigorous activity (days/week)** | | | | |
| <3 | 106893 (70.2%) | 3032 (71.7%) | 0.038 | 109925 (70.3%) |
| >=3 | 45335 (29.8%) | 1196 (28.3%) | | 46531 (29.7%) |
| **Sleep (hours/day)** | | | | |
| 7~8 | 103397 (67.9%) | 2946 (69.7%) | 0.026 | 106343 (68.0%) |
| <7 | 36778 (24.2%) | 946 (22.4%) | | 37724 (24.1%) |
| >8 | 12053 (7.9%) | 336 (7.9%) | | 12389 (7.9%) |
| **Insomnia** | | | | |
| Rarely | 29678 (19.5%) | 760 (18.0%) | 0.039 | 30438 (19.5%) |
| Sometimes | 74784 (49.1%) | 2098 (49.6%) | | 76882 (49.1%) |
| Usually | 47766 (31.4%) | 1370 (32.4%) | | 49136 (31.4%) |
| **Alcohol drinker status** | | | | |
| Never | 9079 (6.0%) | 203 (4.8%) | 0.004 | 9282 (5.9%) |
| Previous | 5428 (3.6%) | 140 (3.3%) | | 5568 (3.6%) |
| Current | 137721 (90.5%) | 3885 (91.9%) | | 141606 (90.5%) |
| **Smoking status** | | | | |
| Never | 91273 (60.0%) | 2454 (58.0%) | 0.04 | 93727 (59.9%) |
| Previous | 47652 (31.3%) | 1394 (33.0%) | | 49046 (31.3%) |
| Current | 13303 (8.7%) | 380 (9.0%) | | 13683 (8.7%) |
| **PC1** | | | | |
| Mean (SD) | -0.00149 (1.00) | 0.0538 (0.973) | 0.498 | <0.00001 (1.00) |
| Median [Min, Max] | 0.103 [-6.77, 6.00] | 0.130 [-4.44, 3.21] | | 0.103 [-6.77, 6.00] |
| **PC2** | | | | |
| Mean (SD) | -0.000323 (1.00) | 0.0116 (0.958) | 0.498 | <-0.00001 (1.00) |
| Median [Min, Max] | -0.0410 [-4.47, 15.8] | -0.0187 [-3.05, 7.23] | | -0.0403 [-4.47, 15.8] |
| **PC3** | | | | |
| Mean (SD) | -0.000756 (1.00) | 0.0272 (0.977) | 0.498 | <0.000001 (1.00) |
| Median [Min, Max] | -0.0326[-12.9, 12.6] | 0.00730 [-7.13, 4.10] | | -0.0314 [-12.9, 12.6] |
| **Medical risk** | | | | |
| **Diabetes** | | | | |
| No | 143723 (94.4%) | 3983 (94.2%) | 0.585 | 147706 (94.4%) |
| Yes | 8505 (5.6%) | 245 (5.8%) | | 8750 (5.6%) |
| **Cardiovascular** | | | | |
| No | 125720 (82.6%) | 3402 (80.5%) | **<0.001** | 129122 (82.5%) |
| Yes | 26508 (17.4%) | 826 (19.5%) | | 27334 (17.5%) |
| **Hypertension** | | | | |
| No | 114839 (75.4%) | 2952 (69.8%) | **<0.001** | 117791 (75.3%) |
| Yes | 37389 (24.6%) | 1276 (30.2%) | | 38665 (24.7%) |
| **Respiratory** | | | | |
| No | 124339 (81.7%) | 3369 (79.7%) | 0.001 | 127708 (81.6%) |
| Yes | 27889 (18.3%) | 859 (20.3%) | | 28748 (18.4%) |
| **Family breast cancer history** | | | | |
| No | 140822 (92.5%) | 3763 (89.0%) | **<0.001** | 144585 (92.4%) |
| Yes | 11406 (7.5%) | 465 (11.0%) | | 11871 (7.6%) |
| **Social** | | | | |
| **Townsend Deprivation Index** | | | | |
| Mean (SD) | -1.30 (3.05) | -1.43 (2.97) | 0.996 | -1.31 (3.05) |
| Median [Min, Max] | -2.11 [-6.26, 10.6] | -2.21 [-6.26, 8.58] | | -2.11 [-6.26, 10.6] |
| **Own or rent accommodation lived in** | | | | |
| own | 137377 (90.2%) | 3871 (91.6%) | 0.005 | 141248 (90.3%) |
| rent | 14851 (9.8%) | 357 (8.4%) | | 15208 (9.7%) |
| **Type of accommodation lived in** | | | | |
| House | 137275 (90.2%) | 3815 (90.2%) | 0.98 | 141090 (90.2%) |
| Flat | 14498 (9.5%) | 401 (9.5%) | | 14899 (9.5%) |
| Care home/temporary or shel | 455 (0.3%) | 12 (0.3%) | | 467 (0.3%) |
| **Number in household** | | | | |
| 1~2 | 98896 (65.0%) | 2965 (70.1%) | **<0.001** | 101861 (65.1%) |
| >=3 | 53332 (35.0%) | 1263 (29.9%) | | 54595 (34.9%) |
| **Average total household income** | | | | |
| <18000 | 39770 (26.1%) | 1081 (25.6%) | 0.156 | 40851 (26.1%) |
| 18000~30999 | 39974 (26.3%) | 1155 (27.3%) | | 41129 (26.3%) |
| 31000~51999 | 37926 (24.9%) | 1080 (25.5%) | | 39006 (24.9%) |
| >52000 | 34558 (22.7%) | 912 (21.6%) | | 35470 (22.7%) |
| **Employment** | | | | |
| Unemployed | 13627 (9.0%) | 331 (7.8%) | **<0.001** | 13958 (8.9%) |
| Employed | 87630 (57.6%) | 2303 (54.5%) | | 89933 (57.5%) |
| Retired | 50971 (33.5%) | 1594 (37.7%) | | 52565 (33.6%) |
| **Education** | | | | |
| Low | 24614 (16.2%) | 696 (16.5%) | 0.653 | 25310 (16.2%) |
| Intermediate | 78354 (51.5%) | 2146 (50.8%) | | 80500 (51.5%) |
| High | 49260 (32.4%) | 1386 (32.8%) | | 50646 (32.4%) |
| **Biomarker** | | | | |
| **Triglyceride (mmol/L)** | | | | |
| Mean (SD) | 1.52 (0.837) | 1.55 (0.853) | 0.004 | 1.52 (0.838) |
| Median [Min, Max] | 1.31 [0.238, 11.2] | 1.34 [0.327, 8.64] | | 1.31 [0.238, 11.2] |
| **Glycated haemoglobin (mmol/L)** | | | | |
| Mean (SD) | 35.7 (6.01) | 35.8 (5.88) | 0.031 | 35.7 (6.01) |
| Median [Min, Max] | 35.1 [15.3, 296] | 35.2 [15.6, 122] | | 35.1 [15.3, 296] |
| **Glucose (mmol/L)** | | | | |
| Mean (SD) | 5.05 (1.06) | 5.06 (0.946) | 0.884 | 5.05 (1.06) |
| Median [Min, Max] | 4.90 [1.41, 34.5] | 4.91 [2.62, 21.4] | | 4.90 [1.41, 34.5] |
| **HDL cholesterol (mmol/L)** | | | | |
| Mean (SD) | 1.59 (0.374) | 1.60 (0.381) | 0.002 | 1.59 (0.374) |
| Median [Min, Max] | 1.55 [0.293, 4.40] | 1.56 [0.644, 3.71] | | 1.55 [0.293, 4.40] |
| **IGF-1 (nmol/L)** | | | | |
| Mean (SD) | 21.3 (5.77) | 21.2 (5.68) | 1 | 21.3 (5.76) |
| Median [Min, Max] | 21.0 [2.16, 125] | 21.0 [4.16, 81.3] | | 21.0 [2.16, 125] |
| **SHBG (nmol/L)** | | | | |
| Mean (SD) | 61.8 (30.2) | 60.1 (30.1) | 0.836 | 61.7 (30.2) |
| Median [Min, Max] | 56.6 [0.390, 240] | 54.5 [5.91, 228] | | 56.5 [0.390, 240] |

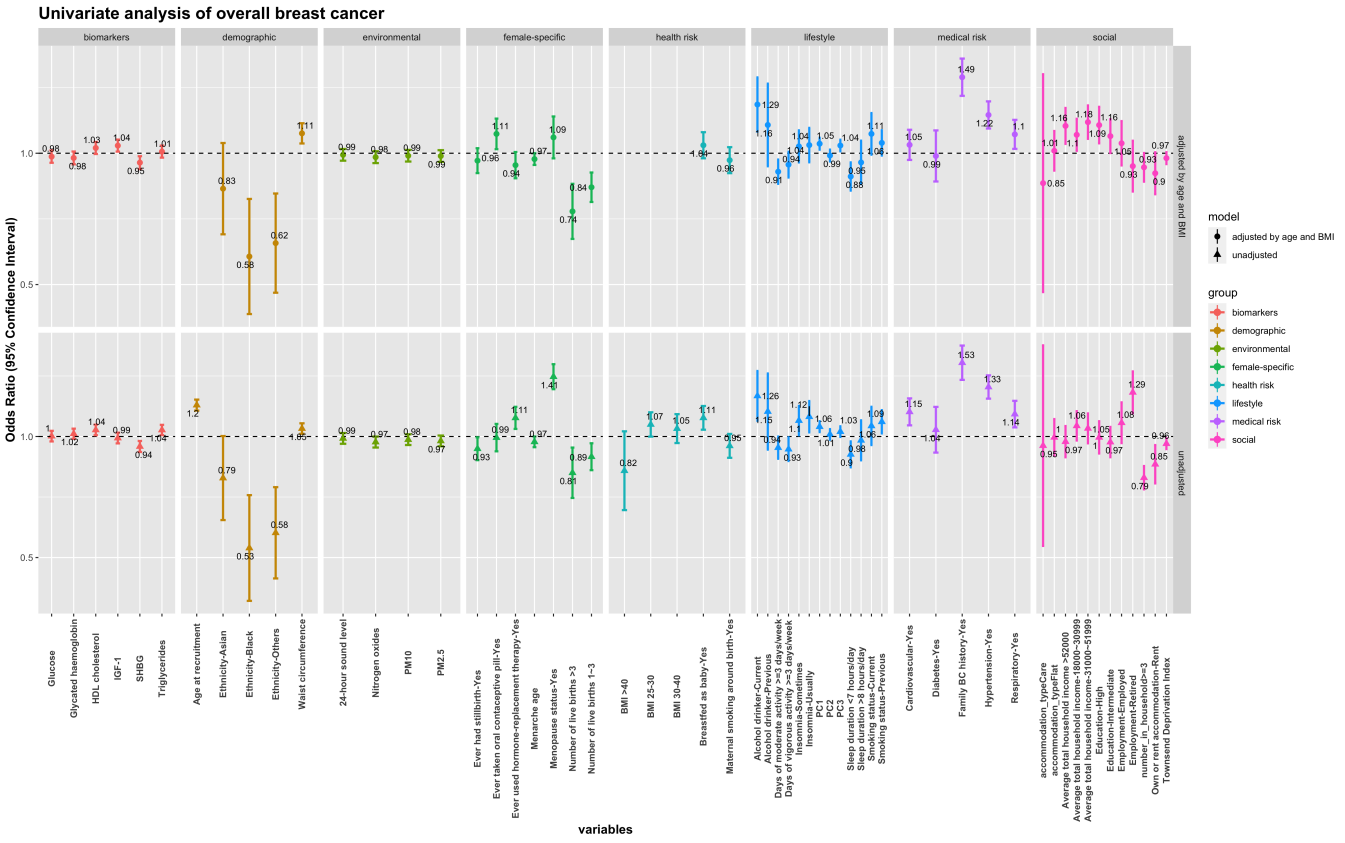


**Figure 1: Forest plot for overall breast cancer**

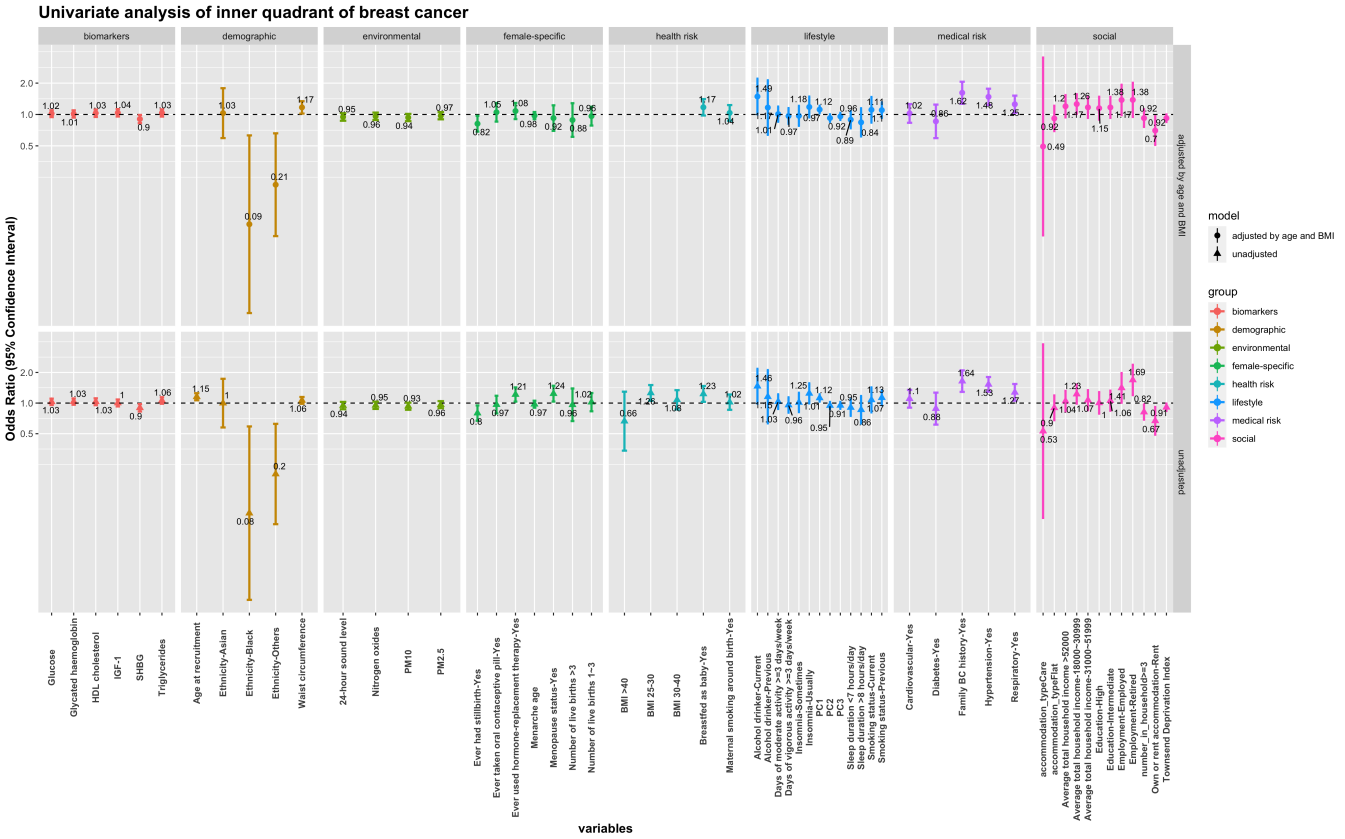


**Figure 2: Forest plot for inner quadrant breast cancer**

cancer ($\beta$ =0.377), hypertension ($\beta$ = 0.273), BMI more than 40 ($\beta$ =-0.170), diabetes ($\beta$ =-0.139) and living in rented accommodation ($\beta$ =-0.109) contributed largely to inner quadrant subtype with AUC of 0.594 (Figure 9); hypertension ($\beta$ =0.249), family history of breast cancer ($\beta$ =0.204), Black and other ethnicity ($\beta_{Black}$ = -0.169, $\beta_{other}$ = -0.154), current smoking status ($\beta$ =0.122), and post-menopause ($\beta$ =0.106) contributed largely to outer quadrant subtype with AUC of 0.574 (Figure 10).

## 3.6 Partial Least Squares

With relation to the loading direction of breast cancer status in PLS-DA, negative loadings in covariates were interpreted to have a deleterious effect on breast cancer development while positive loadings were indicative of a protective effect. The calibration of the sPLS-DA analysis for pooled cases selected 7 variables (Figure 11:A). Recruitment age, retirement, post-menopausal, hypertension and family history of breast cancer presented strong negative loadings. More than three people living in a household and Black ethnic group generated moderate to low positive loadings. In the stability analysis of size 7, all variables listed were stably selected except Black ethnicity. Calibration of sgPLS-DA selected a total of 2 groups with a sparsity parameter of 0.7, which were the demographic and the female-specific groups.

sPLS-DA calibration on the stratification of breast cancer cases by inner quadrant breast cancer selected 15 variables ((Figure 11:B). In the stability analysis of size 15, 7 were stably selected, these variables were recruitment age, retirement, diet PC1, hypertension and family history of breast cancer with negative loadings; black ethnicity, other ethnic group with positive loadings. Calibration of the sgPLS selected two groups: demographic and health risk groups. Negative loadings were consistently observed in recruitment age, normal BMI range, hypertension and respiratory diseases. Higher positive loadings were observed in black ethnicity and other ethnic group while lower positive loading was observed with obesity (BMI>40 kg/m$^2$).

sPLS-DA calibration on outer quadrant subtype selected 53 variables in total (Figure 11:C). The highest negative loadings were consistent with the findings from the base model of overall breast cancer. SgPLS-DA had chosen 2 groups (demographic and female-specific groups), with strong negative loadings in recruitment age, family history of breast cancer, post-menopause status and history of hormone replacement therapy. Positive loadings were observed in Asian ethnicity, Black ethnicity and other ethnic group. In the stability analysis for outer quadrant breast cancer, recruitment age, hypertension, more than three people living in a household and post-menopausal were highly selected, showing consistency with the base model. Results of stability analyses and calibration plots

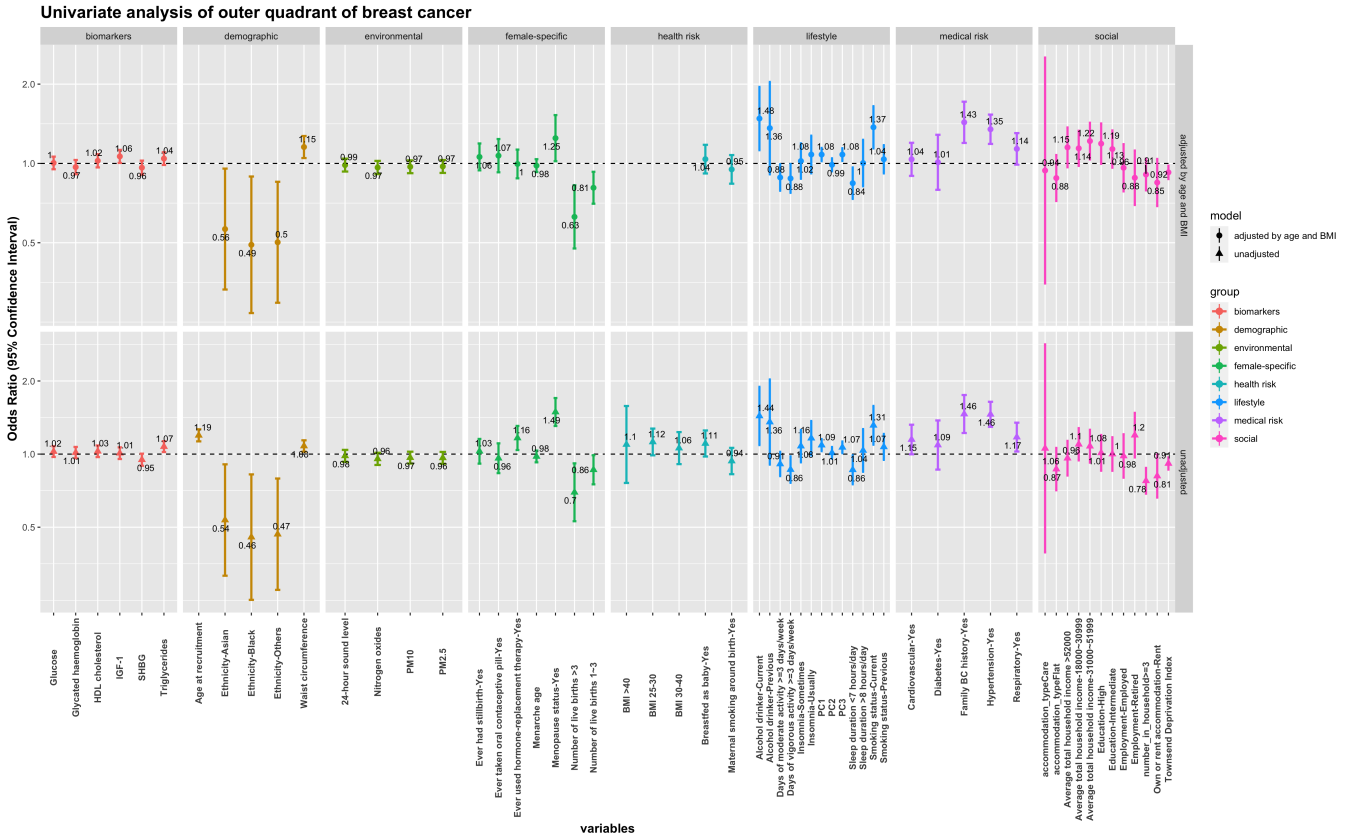


**Figure 3: Forest plot for outer quadrant breast cancer**

could be found in the supplementary materials. (Figure 20 - 22). (Figure 19).

## 3.7 Sensitivity analysis

In order to examine whether the uncertainty of models exists under different constraints, which would yield different results and reduce the statistical power of our analysis, the adjusted univariate model was re-analysed and stratified by five assumptions: (1)Population with/without breast cancer family history(Figure 23), (2)population has/has not used oral contraceptives(Figure 24) , (3)population with early(<12 years old)/normal(12-14 years old)/late(>14 years old) menarche(Figure 25), (4)population with pre-/post-menopause (Figure 26), and (5) population without/without history of stillbirth (Figure 27). Although the effects of several covariates slightly fluctuated in these models, there was no strong evidence of mismatch found.

# 4 DISCUSSION

In this study, we consistently found that the future risk of female breast cancer and its subtypes (inner quadrant and outer quadrant) were positively associated with recruitment age, family history of breast cancer, hypertension in all models, whereas being Black and Other ethnicity showed a protective effect.

In the fully adjusted model, we found that Black or other vs White ethnicity; sleep duration <7 vs 7-8; average total household income 31,000 51,999 vs <18,000; live births number 1 3 or >3 vs 0; BMI 30-40 or >40 vs <25; family history (yes vs no); hypertension (yes vs no); respiratory (yes vs no) were all independently associated with the risk of breast cancer. Protective effects of higher BMI were detected in the previous studies, but the biological mechanisms of the effects required further investigations. We found that the associations linking alcohol drinker Current vs Never and the risk of breast cancer were strongly attenuated while adjusting for the other covariates.

We used logistic Lasso models to select variables and measured their joint contributions to the risk of breast cancer and its subtypes, capturing complex and combined effects. From overall and subtype analysis in breast cancer, logistic Lasso models consistently found that family history of breast cancer and hypertension were positively associated, while Black and others ethnicity was negatively associated. Our results were in line with other studies. Engmann et al. found that the odds ratio of family history on breast cancer was 1.71 (95% CI 1.59–1.84) [13]. A cohort with over 113,000 women in the UK population also showed that women having two or more relatives with breast cancer had a 2.5-fold (95% CI 1.83–3.47) risk of developing breast cancer [14], implying an underlying influence of genetics. A familial breast cancer research found that 50–85% of

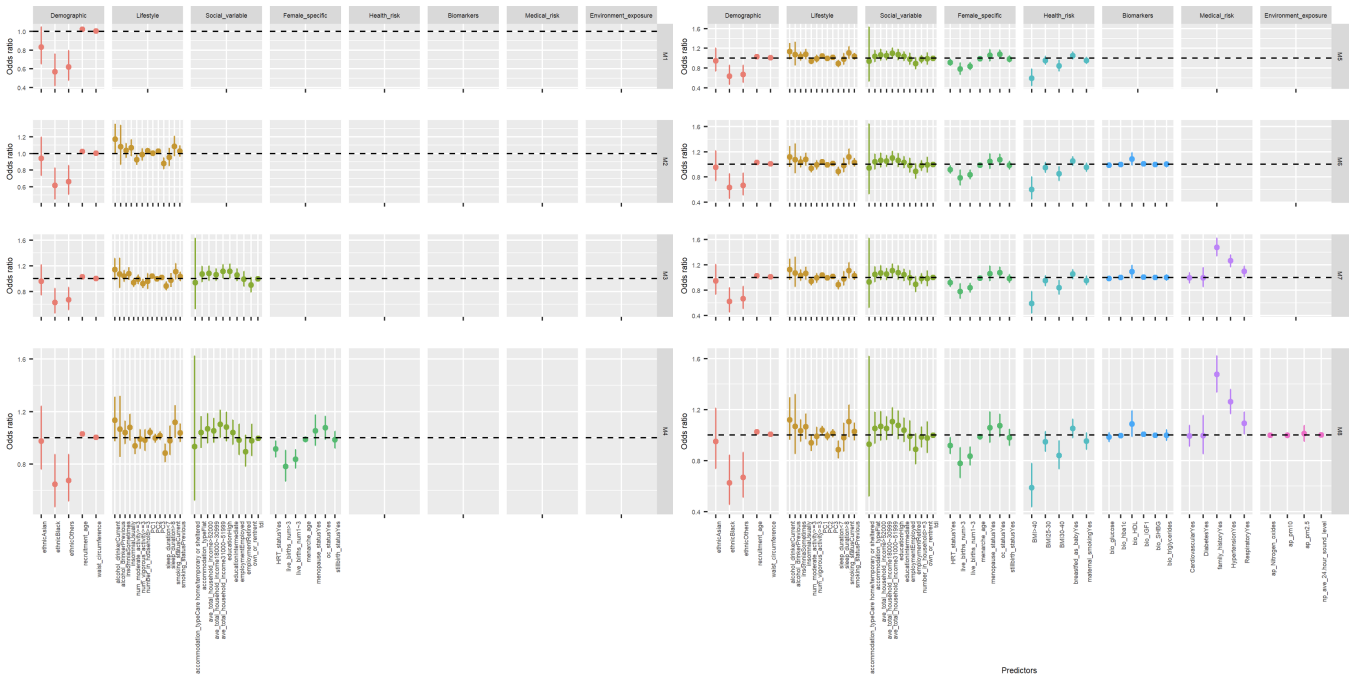


**Figure 4: Odds ratio with 95% confidence intervals and p-values from the multivariate models for the risk of breast cancer. Results are represented for chronologically adjusted models. Baseline predictors are defined as demographic variables (Model 1), and models are additionally adjusted for lifestyle (Model 2), social variable (Model 3), female specific (Model 4), health risk (Model 5), biomarkers (Model 6), medical risk (Model 7) and environment exposure (Model 8).**

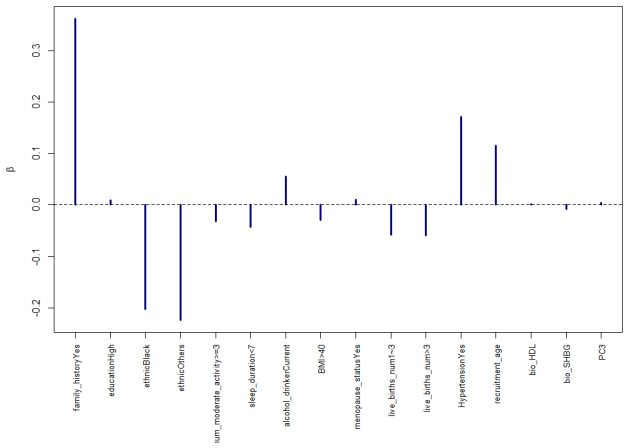


**Figure 5: Logistic lasso for overall breast cancer**

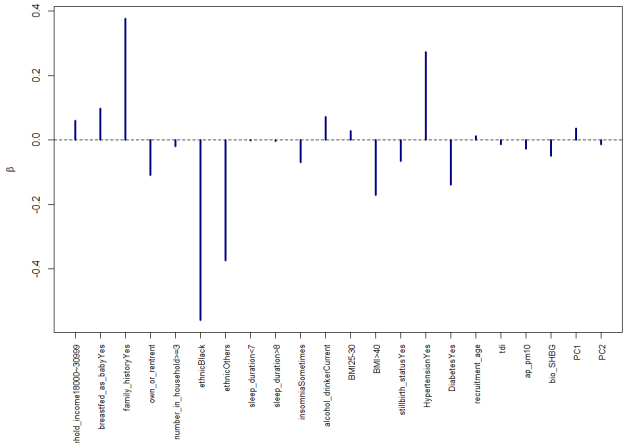


**Figure 6: Logistic lasso for inner quadrant subtype**

women with mutations in BRCA1 and BRCA2 would develop breast cancer during their lifetime and these mutations could be inherited and enriched in families [15]. Li et al. and Choi et al. reported that uncommon mutations, like PTEN and ATM, could also lead to breast and ovarian cancer [16,17]. Moreover, several systematic reviews and meta-analysis reported statistically significant association between hypertension and increased breast cancer risk (RR: 1.15; 95% CI: 1.08, 1.22) [18] and possible mechanisms had been proposed to explain this relationship. There were some evidence showing that breast cancer and hypertension share a common pathophysiological pathway mediated by adipose tissue, which could initiate chronic inflammation and increase risk of both breast cancer and hypertension [19,20,21]. Secondly, hypertension may increase breast cancer risk by blocking and subsequently modifying mechanism of apoptosis, thereby affecting the regulation of cell turnover [22,23]. The relationships between Black and other ethnicity and breast cancer

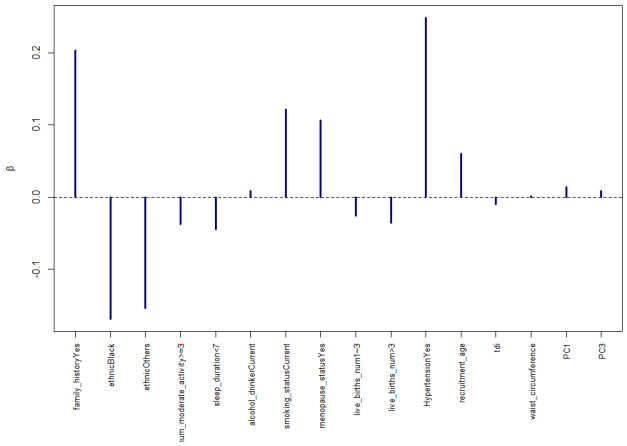


**Figure 7: Logistic lasso for outer quadrant subtype**

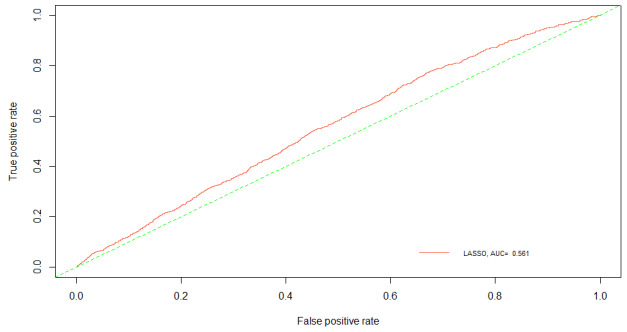


**Figure 8: AUC for overall breast cancer**

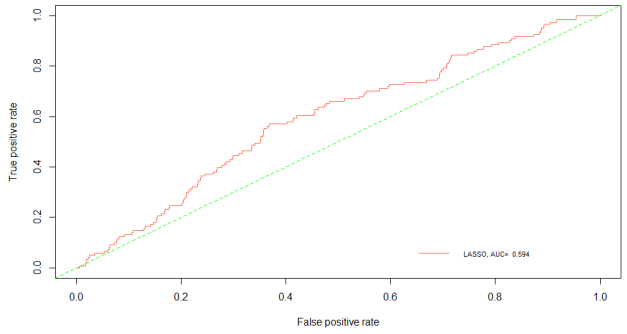


**Figure 9: AUC for inner quadrant subtype**

observed in our study were not fully supported from other literature evidence [24]. However, Black and others ethnicity only accounted for 1.9% and 2.3% respectively in our study population which might imposed some uncertainty in our findings. Further work should be

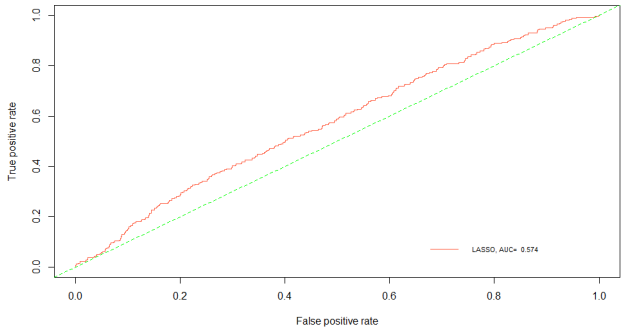


**Figure 10: AUC for outer quadrant subtype**

conducted focusing on investigating of such associations in more racially balanced cohort.

Results from PLS-DA analysis on pooled cases and subtypes presented consistent and similar findings to logistic Lasso models. Moreover, we identified slight differential pattern in loadings and group assignments of inner quadrant subtype compared to that of other subtypes. In particular, negative loading of diet PC1 was identified in sPLS-DA analysis and stably selected in the inner subtype model, indicating a negative impact on cancer risk. Nutrition has been previously described to play a role in 35% of cancer aetiology [25]. A possible mechanism linking the consumption of red meat to breast cancer development has been proposed by Rose et al, where the accumulation of adipose tissue from pursuing a high fat diet might indirectly increase the amount of free-circulating oestrogen, leading to the activation of breast growth by hormonal stimulations [26]. Further investigation is needed to confirm this association and examine how dietary choice contributes to the formation of breast cancer subtypes. Low positive loading was observed in obesity of the inner subtype model demonstrating protective effect to breast cancer development. This association had been detailed in some studies with controversial findings amongst pre- and post-menopausal women of hormone-receptor positive breast cancer, which required further research to validate the robustness of such findings. Studies had suggested that obesity generated a 20-40% increased risk of developing hormone-receptor positive breast cancer in post-menopausal women whereas associated with a reduced risk in pre-menopausal obese individuals [27]. In future work, ROC and AUC with additive effect on variable groups would be beneficial to identify how the inclusion of each category impacts the overall model prediction and could be compared to results from sgPLS models.

Our study has the advantages of sampling participants from the UK Biobank, consisting of a large population size and availability on a wide range of risk factors detailing systematic differences between individual characteristics. We identified robust and consistent associations of risk factors with relation to breast cancer using a series of statistical analysis to study the individual and joint effects of covariates using univariate to multivariate techniques. There are several limitations to our findings that might reduce generalisability

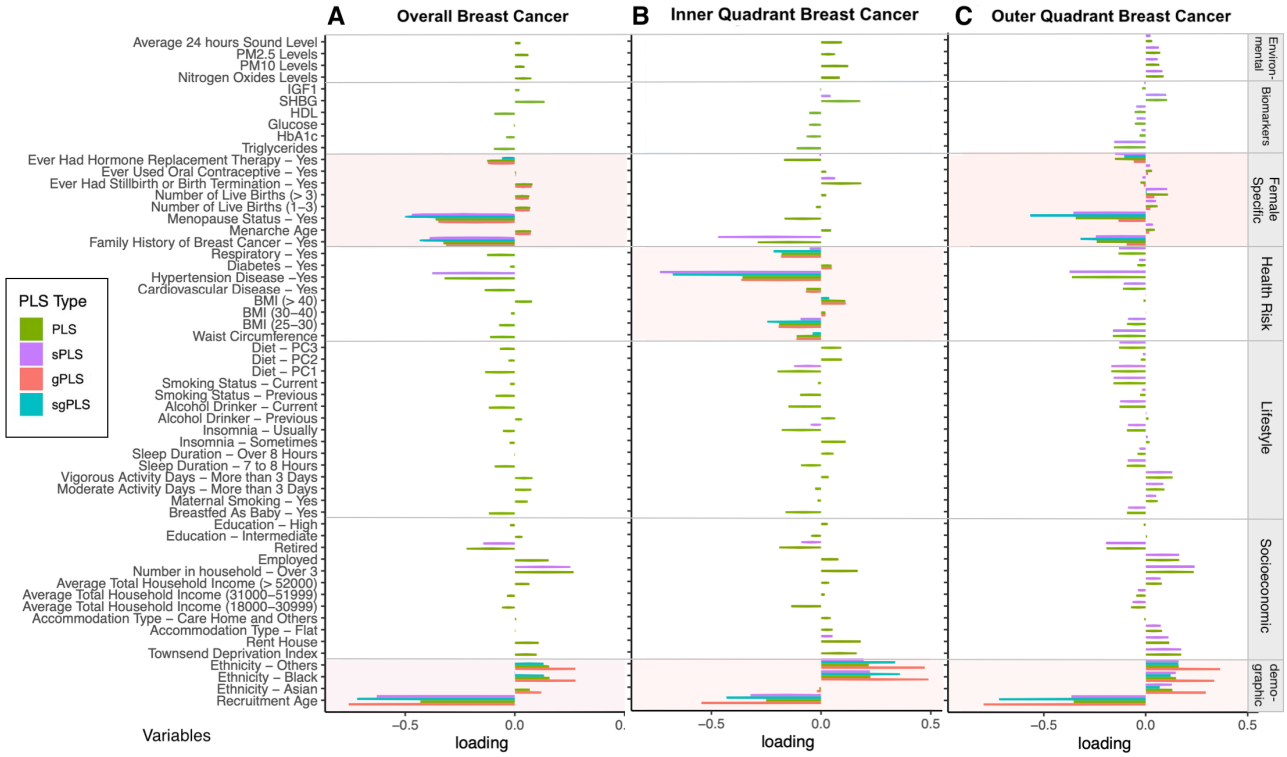


**Figure 11: Results of PLS-DA analyses for A) overall breast cancer B) inner quadrant subtype C) outer quadrant subtype. Loadings coefficients for sPLS-DA and sgPLS-DA in purple and blue respectively. sgPLS selected variable groups highlighted in pink for each model.**

our study results. Firstly, the UK biobank data employed participants of older age with higher socio-economic status [28]. Due to the selection bias of healthy controls and the over-representation of White ethnicity in our study population, our dataset contained a limited number of other racial groups which might not be representative of the general population in the country [29]. In addition, some of the lifestyle related factors included in our dataset such as exposure to maternal smoking at birth, were self-reported and therefore some uncertainty might have been introduced in our models though possible response bias. Finally, the UK Biobank did not contain information on molecular signatures of breast cancer (ER, PR and HER2 status), thus we stratified our breast cancer subtypes by the position of breast tumour, which is less common in literature research of breast cancer and more difficult to compare our findings on inner and outer quadrant breast cancer to similar published literature.

## 5 CONCLUSION

Large prospective cohort like the UK biobank enabled us to identify independent and joint associations of covariates with the future risk of breast cancer. We found robust associations that postmenopausal, older, hypertensive women with first degree family history contributed to an increased risk of breast cancer development. Our study provided a theoretical basis of risk factors involved in breast cancer and facilitated further research on the understanding breast cancer etiology with relation to the associated findings. Further analysis could be considered to integrate genetic information from the UK Biobank to explore polygenic risk score as a predictor of breast cancer and applying time-to-event models such as Cox regression to incorporate the length of time to breast cancer diagnosis. As our study focused solely on the female population, it would also be valuable to validate our findings on the male cohort as an additional sensitivity analysis to our study.

## 6 AUTHOR CONTRIBUTIONS

All authors discussed and finalized the analytical plan. 01814471, 01333689, 01377486 and 01913993 extracted variables from biobank and 01808753 did data processing. For statistical analysis, 01902920 did descriptive analysis, 01333689 for univariate analysis, 01808753 for Lasso regression, 01377486 for PLS analysis, 01913993 for fully adjusted model and 01814471 for sensitivity analysis. For writing manuscript, each author finished their own part in results and discussions. 01902920, 01814471, 01913993 were responsible for abstract and introduction and 01377486, 01333689, 01808753 for method and conclusion. All authors approved the final draft.

# REFERENCES


[1]: Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394-424.

[2]: Who.int. 2021. Latest global cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018. [online] Available at: <https://www.who.int/cancer/PRGlobocanFinal.pdf> [Accessed 5 May 2021].

[3]: Cancer Research UK. 2021. Breast cancer incidence (invasive) statistics. [online] Available at: <https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer/incidence-invasive#heading-Three> [Accessed 5 May 2021].

[4] Bray F, McCarron P, Parkin DM. The changing global patterns of female breast cancer incidence and mortality. Breast Cancer Res. 2004;6(6):229-239.

[5] Al-Ajmi K, Lophatananon A, Ollier W, et al. Risk of breast cancer in the UK biobank female cohort and its relationship to anthropometric and reproductive factors. PLoS One. 2018;13(7):e0201097. [6]: UK-Biobank (2010) UK Biobank. Protocol for a large-scale prospective epidemiological resources [Online] http://www.ukbiobank.ac.uk/wp-content/uploads/2011/11/UK-Biobank-Protocol.pdf Accessed 10 July 2016.

[7] Schwedhelm C, Iqbal K, Knüppel S, et al. Contribution to the understanding of how principal component analysis-derived dietary patterns emerge from habitual data on food consumption. Am J Clin Nutr. 2018;107(2):227-235.

[8] Castelló A, Lope V, Vioque J, et al. Reproducibility of data-driven dietary patterns in two groups of adult Spanish women from different studies. Br J Nutr. 2016;116(4):734-742

[9] Fransen HP, May AM, Stricker MD, et al. A posteriori dietary patterns: how many patterns to retain? J Nutr. 2014;144(8):1274-1282.

[10] Kim JO, Mueller CW. Factor analysis: statistical methods and practical issues. Thousand Oaks: Sage Publications; 1978.

[11] Lunardon N, Menardi G, Torelli N. ROSE: A package for binary imbalanced learning. R Journal, 2014;6(1):79–89.

[12] Jordan I, Hebestreit A, Swai B, et al. Dietary patterns and breast cancer risk among women in northern Tanzania: a case-control study. Eur J Nutr. 2013;52(3):905-915.

[13] Engmann NJ, Golmakani MK, Miglioretti DL, Sprague BL, Kerlikowske K; Breast Cancer Surveillance Consortium. Population-Attributable Risk Proportion of Clinical Risk Factors for Breast Cancer. JAMA Oncol. 2017;3(9):1228-1236.

[14] Brewer HR, Jones ME, Schoemaker MJ, et al. Family history and risk of breast cancer: an analysis accounting for family structure. Breast Cancer Res Treat. 2017;165(1):193-200.

[15] Thompson ER, Rowley SM, Li N, et al. Panel Testing for Familial Breast Cancer: Calibrating the Tension Between Research and Clinical Care. J Clin Oncol. 2016;34(13):1455-1459.

[16] Li S, Shen Y, Wang M, et al. Loss of PTEN expression in breast cancer: association with clinicopathological characteristics and prognosis. Oncotarget. 2017;8(19):32043-32054.

[17] Choi M, Kipps T, Kurzrock R. ATM Mutations in Cancer: Therapeutic Implications. Mol Cancer Ther. 2016;15(8):1781-1791.

[18] Han H, Guo W, Shi W, et al. Hypertension and breast cancer risk: a systematic review and meta-analysis. Sci Rep. 2017;7:44877.

[19] Largent JA, McEligot AJ, Ziogas A, et al. Hypertension, diuretics and breast cancer risk. J Hum Hypertens. 2006;20(10):727-732.

[20] Balkwill F, Charles KA, Mantovani A. Smoldering and polarized inflammation in the initiation and promotion of malignant disease. Cancer Cell. 2005;7(3):211-217.

[21] Li JJ, Fang CH, Hui RT. Is hypertension an inflammatory disease? Med Hypotheses. 2005;64(2):236-240.

[22] Hamet P. Cancer and hypertension. An unresolved issue. Hypertension. 1996;28(3):321-324.

[23] Hamet P. Cancer and hypertension: a potential for crosstalk? J Hypertens. 1997;15(12 Pt 2):1573-1577.

[24] Brennan M. Breast cancer in ethnic minority groups in developed nations: Case studies of the United Kingdom and Australia. Maturitas. 2017 May;99:16-19.

[25] Doll R, Peto R. The causes of cancer: quantitative estimates of avoidable risks of cancer in the United States today. J Natl Cancer Inst. 1981;66(6):1191-308.

[26] Rose DP. Effects of dietary fatty acids on breast and prostate cancers: evidence from in vitro experiments and animal studies. Am J Clin Nutr. 1997;66(6 Suppl):1513S-1522S.

[27] Munsell MF, Sprague BL, Berry DA, et al. Body mass index and breast cancer risk according to postmenopausal estrogen-progestin use and hormone receptor status. Epidemiol Rev. 2014;36(1):114-36.

[28] Khanji MY, Aung N, Chahal CAA, et al. COVID-19 and the UK Biobank-Opportunities and Challenges for Research and Collaboration With Other Large Population Studies. Front Cardiovasc Med. 2020;27;7:156.

[29] Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. Am J Epidemiol. 2017;186(9):1026-1034.


# SUPPLEMENTARY MATERIALS

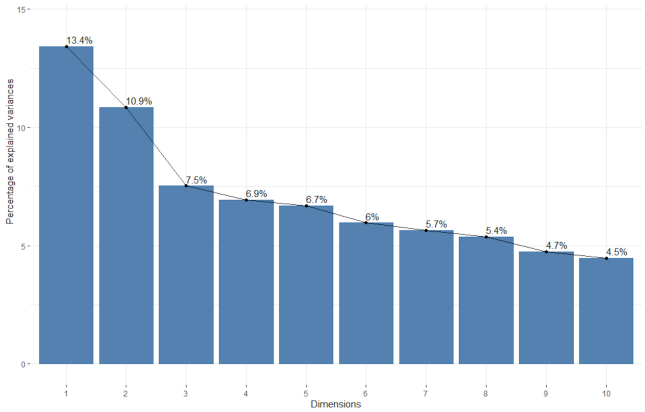


**Figure 12: Scree plot for dietary pattern**

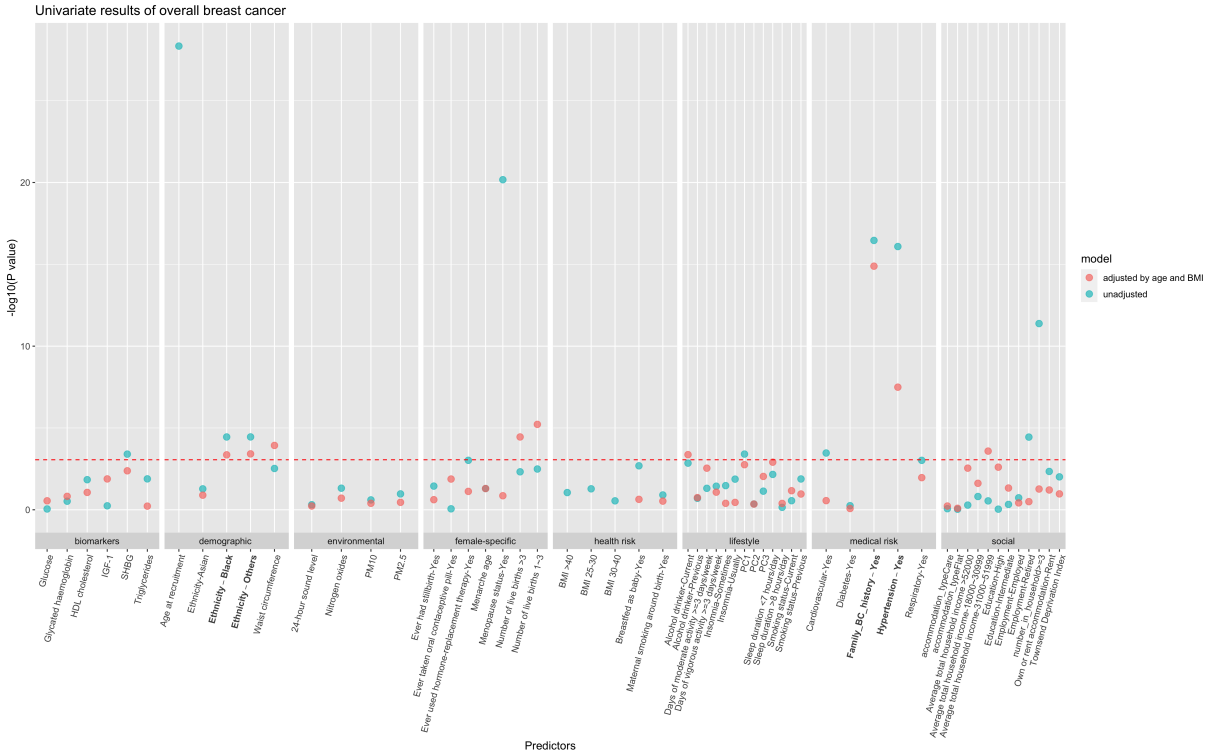


**Figure 13: Manhattan plot for overall breast cancer**

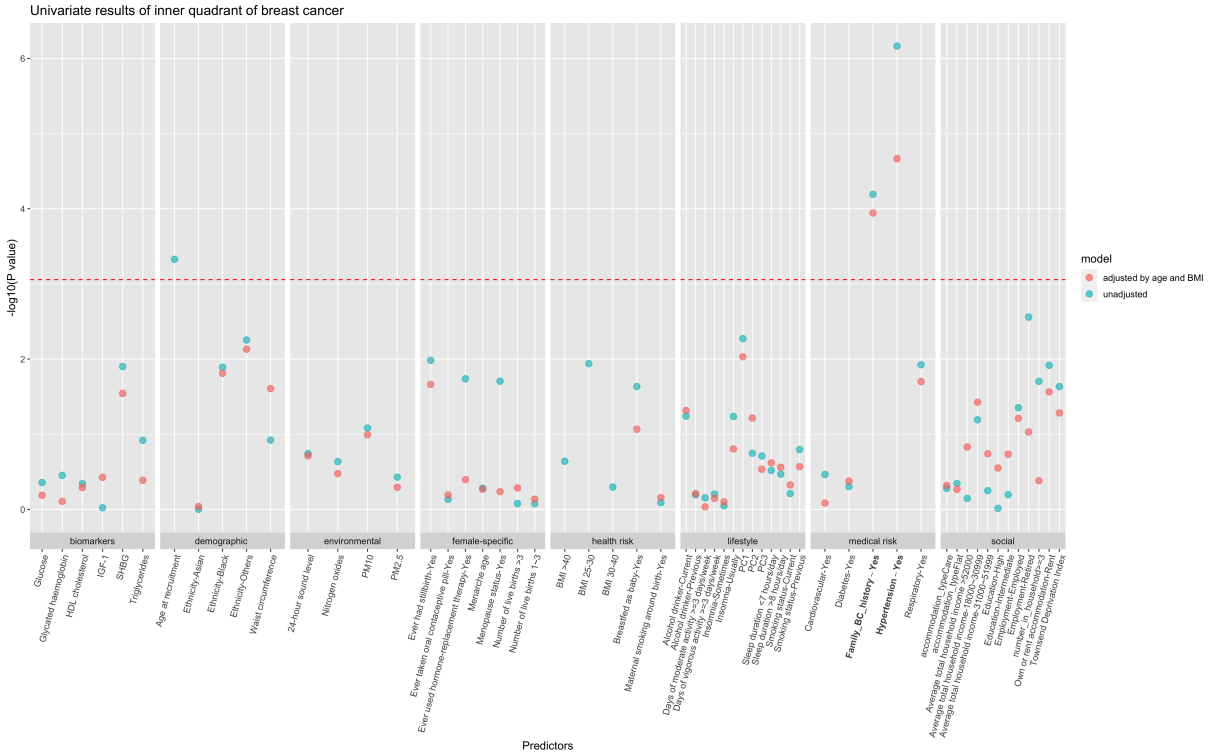


**Figure 14: Manhattan plot for inner quadrant breast cancer**

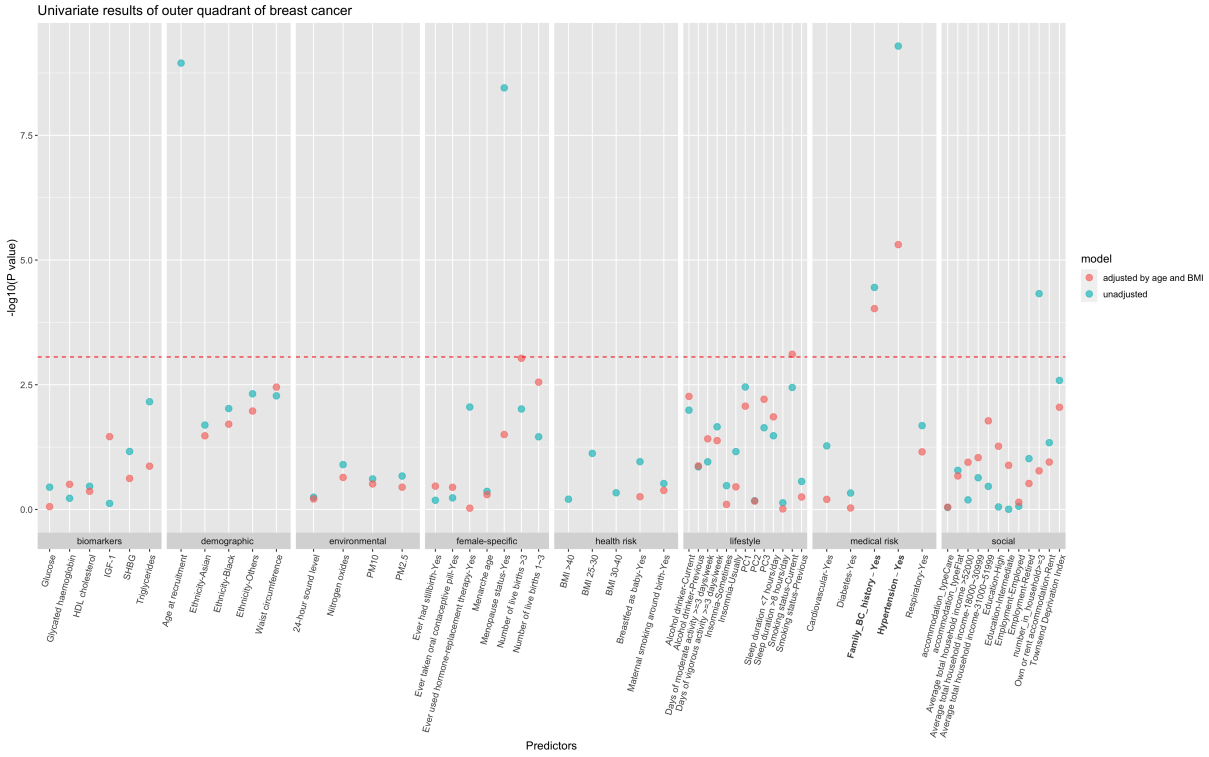


**Figure 15: Manhattan plot for outer quadrant breast cancer**

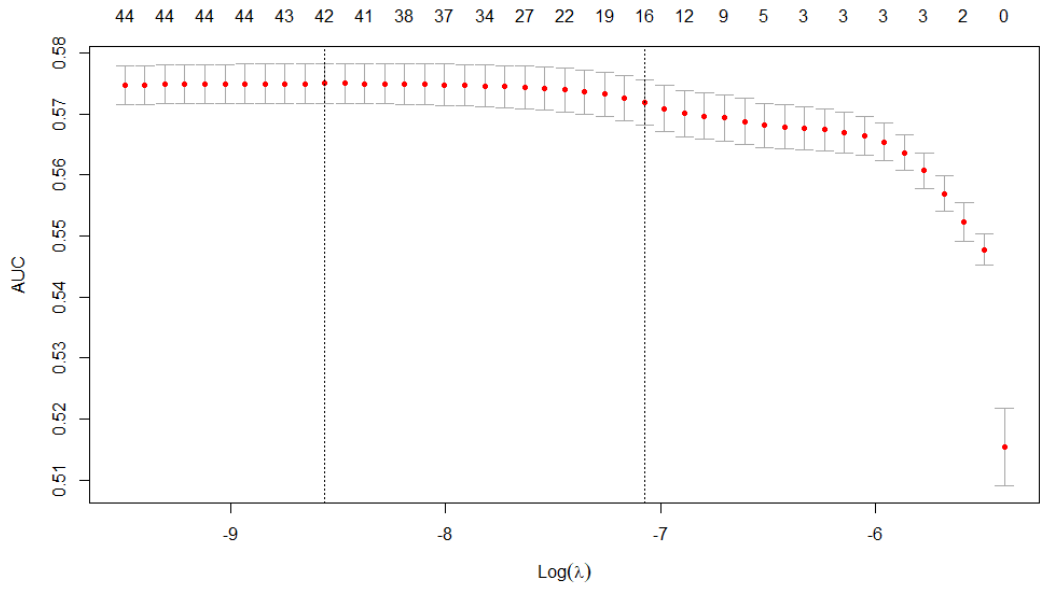


**Figure 16: Calibration of logistic lasso for outer breast cancer**

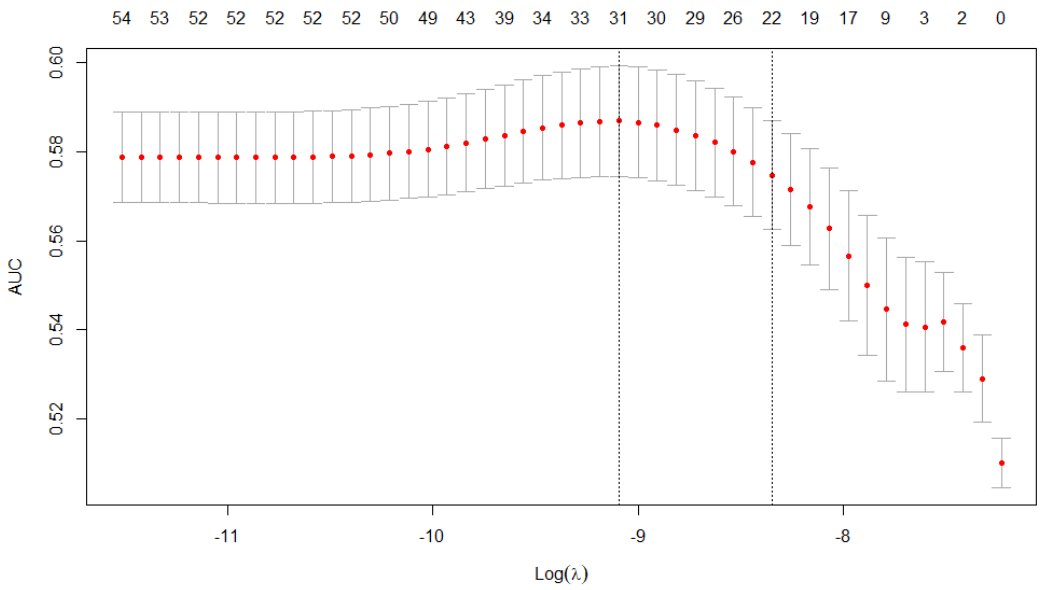


**Figure 17: Calibration of logistic lasso for inner breast cancer**

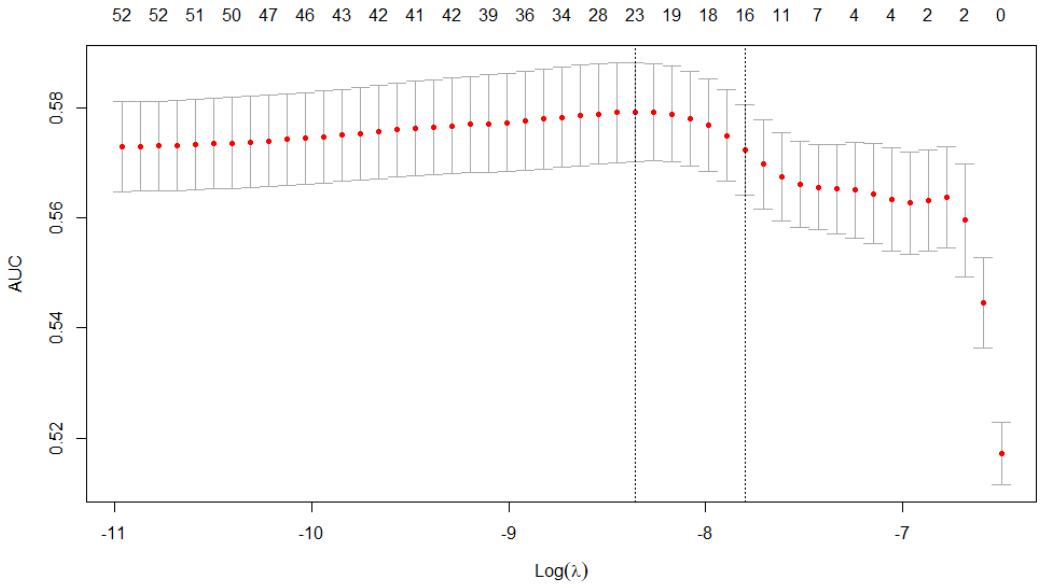


**Figure 18: Calibration of logistic lasso for outer breast cancer**

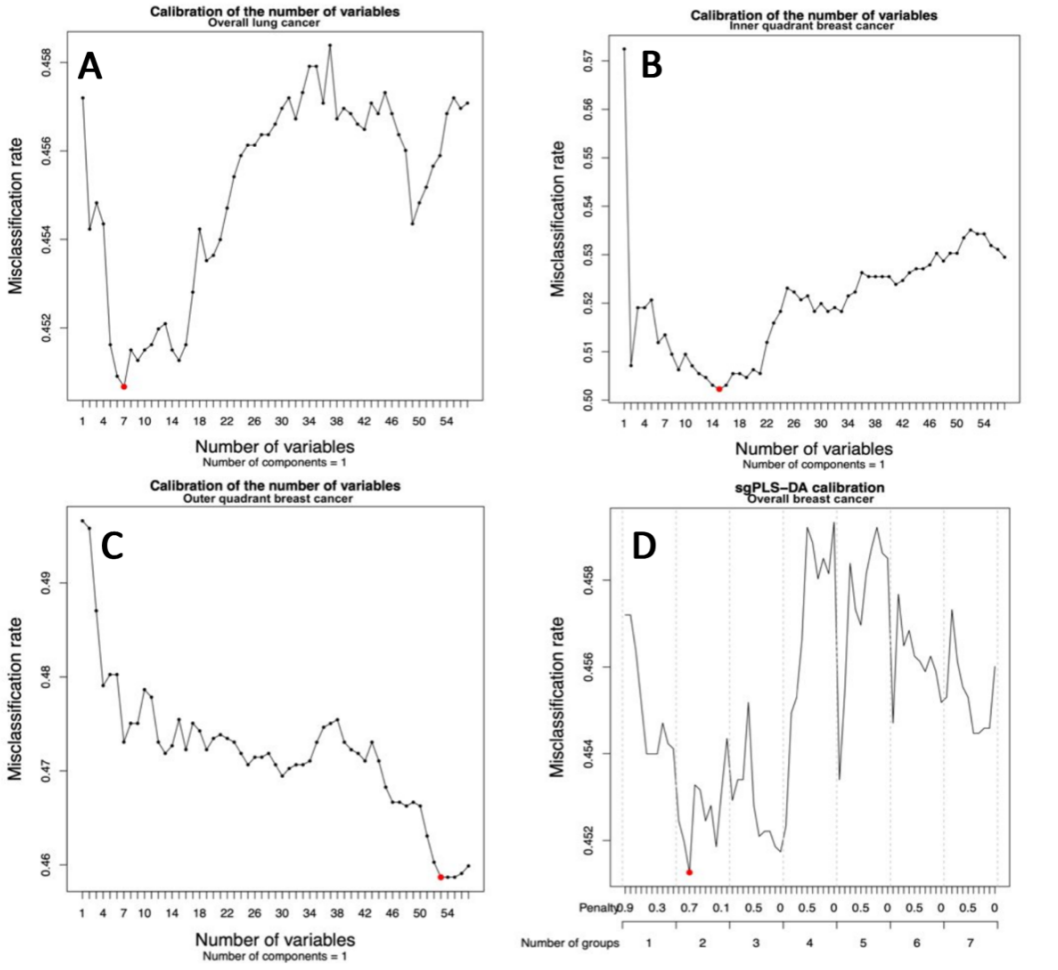


**Figure 19: Calibration results of sPLS on one component for A) overall breast cancer B) inner quadrant subtype C) outer quadrant subtype and D) sgPLS calibration results for overall breast cancer. In each calibration plot, X-axis indicated number of parameters and Y-axis represented misclassification rate. The optimal number of parameters yielding the lowest misclassification rate for each model calibration was highlighted in red.**

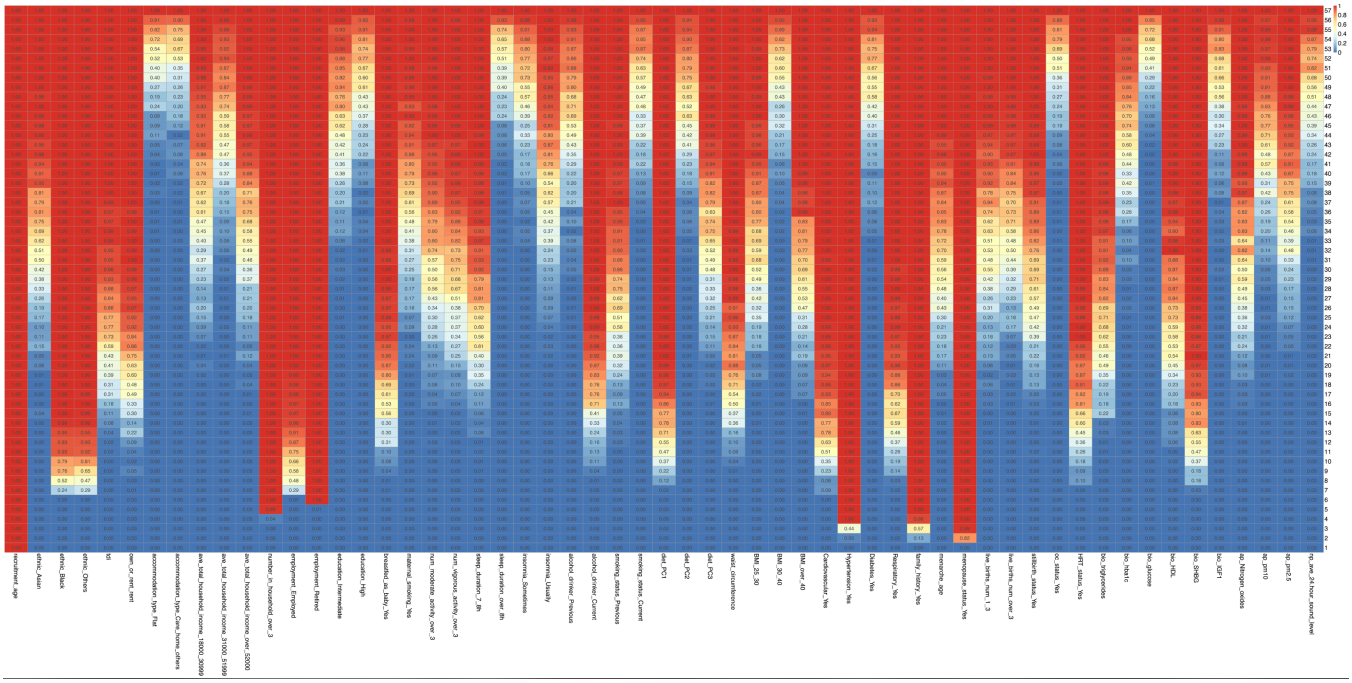

**Figure 20: Sensitivity analysis and selection proportion of sPLS findings on model of overall breast cancer**

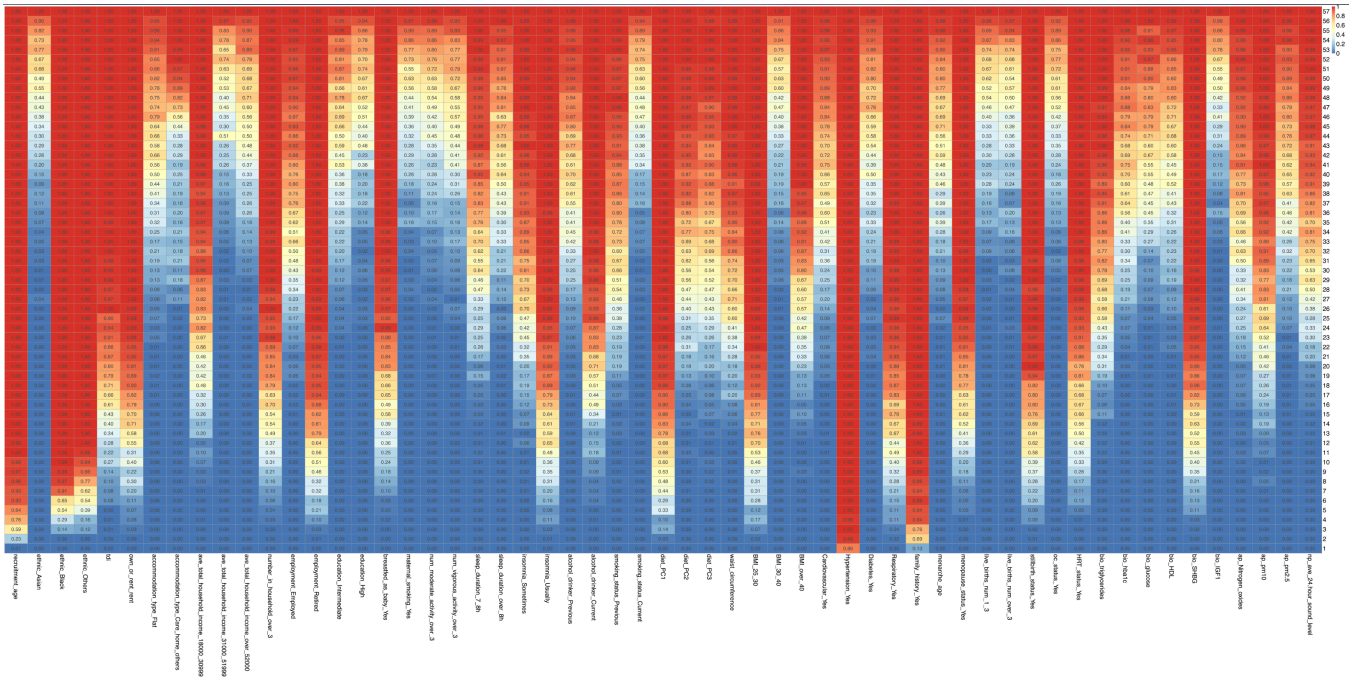

**Figure 21: Sensitivity analysis and selection proportion of sPLS findings on model of inner breast cancer**

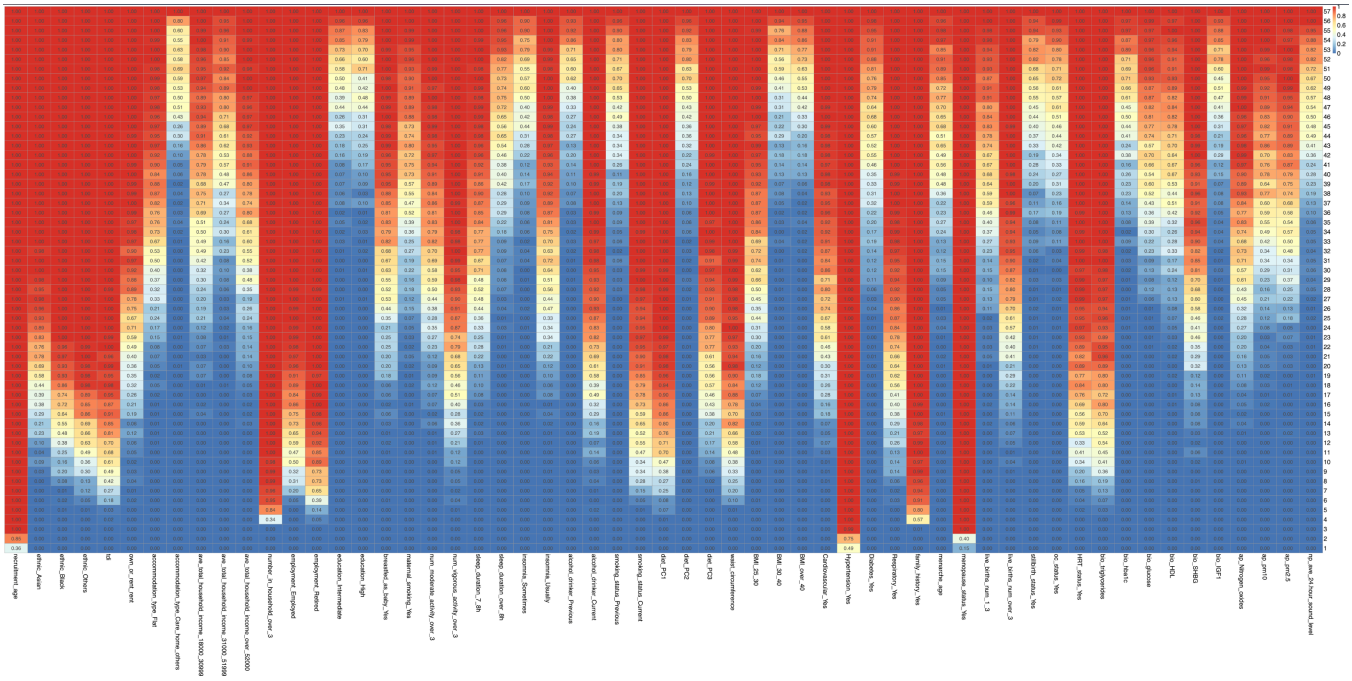

**Figure 22: Sensitivity analysis and selection proportion of sPLS findings on model of outer breast cancer**

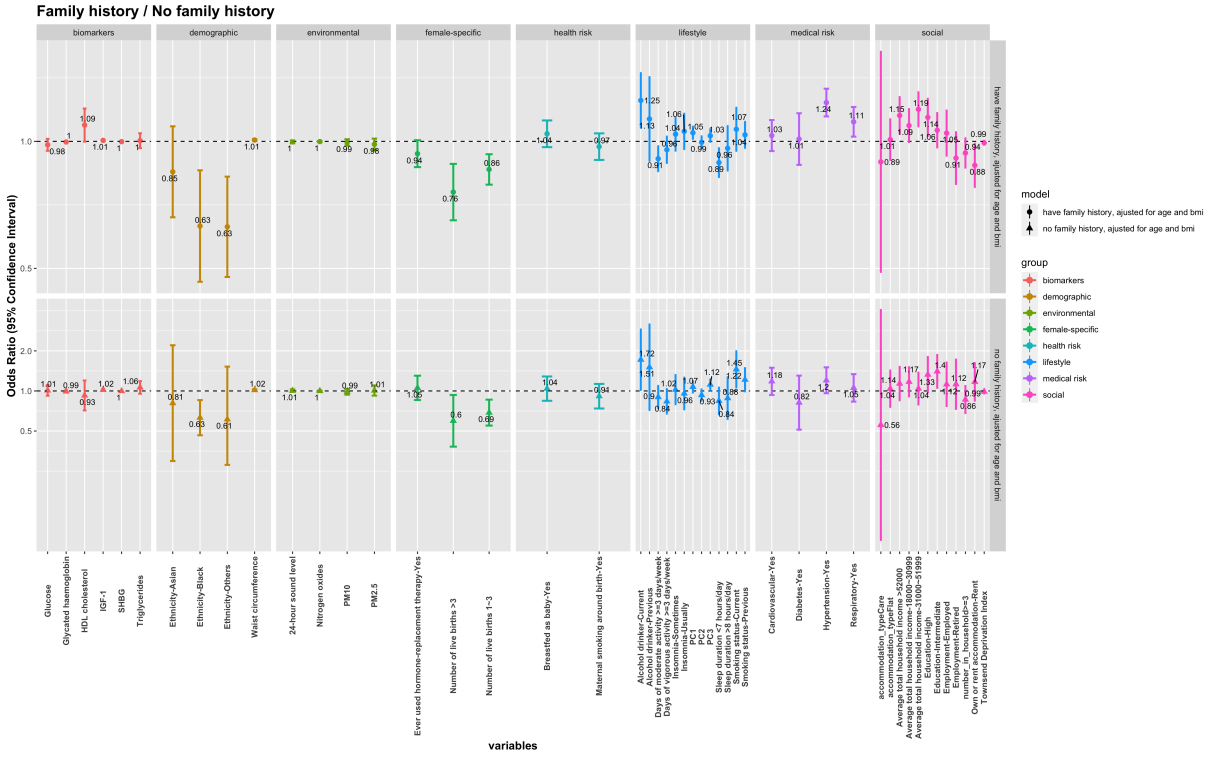


**Figure 23: Sensitivity analysis on breast cancer family history**

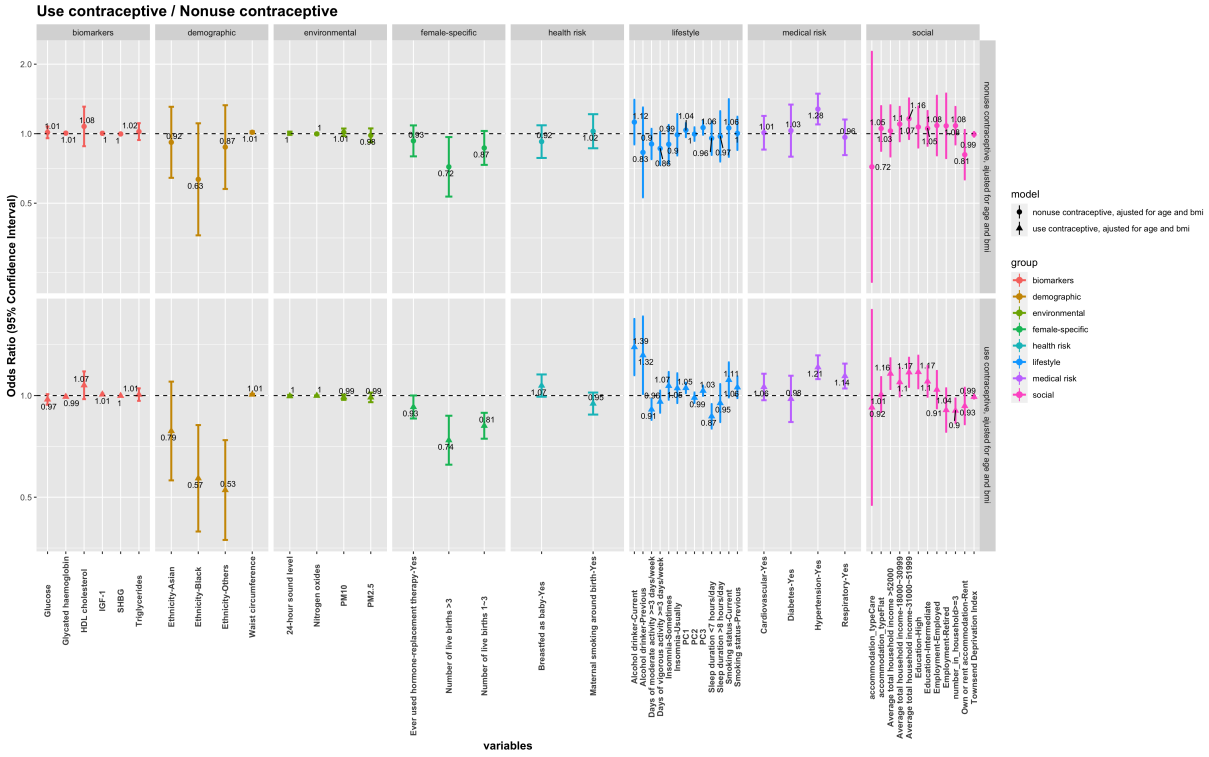


**Figure 24: Sensitivity analysis on oral contraceptives use**

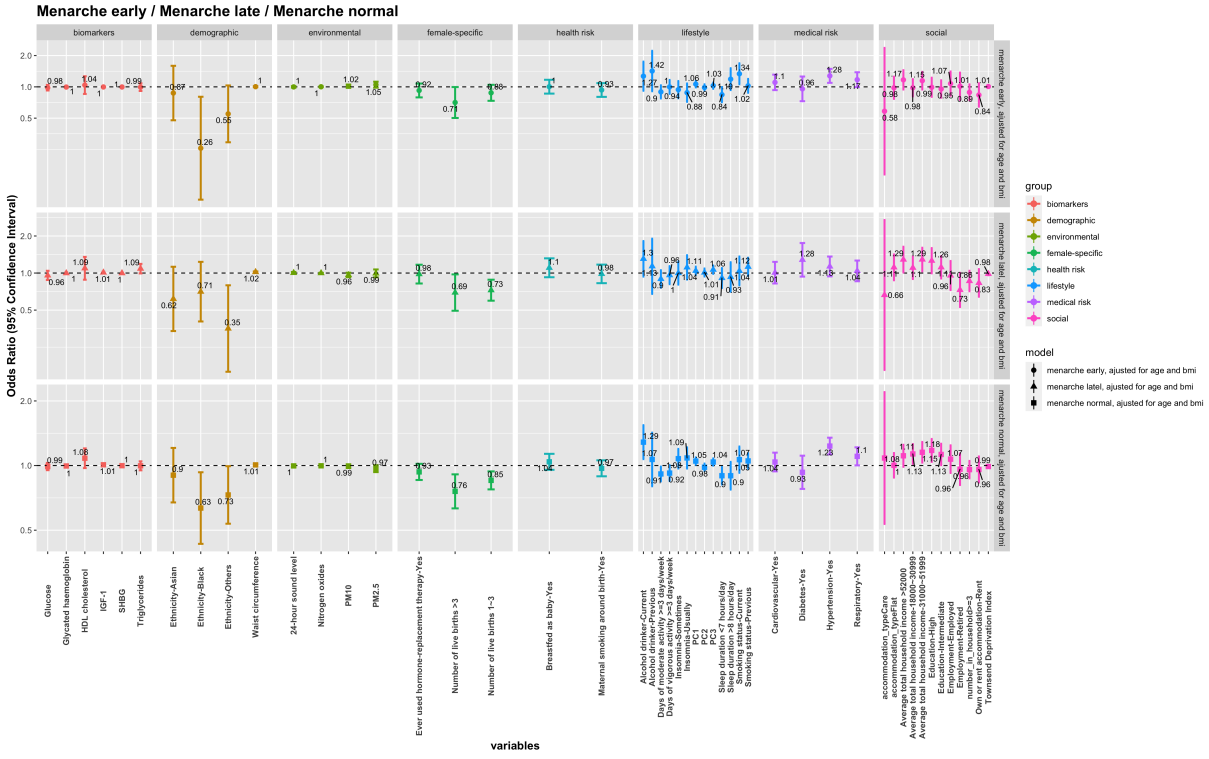


**Figure 25: Sensitivity analysis on menarche age**

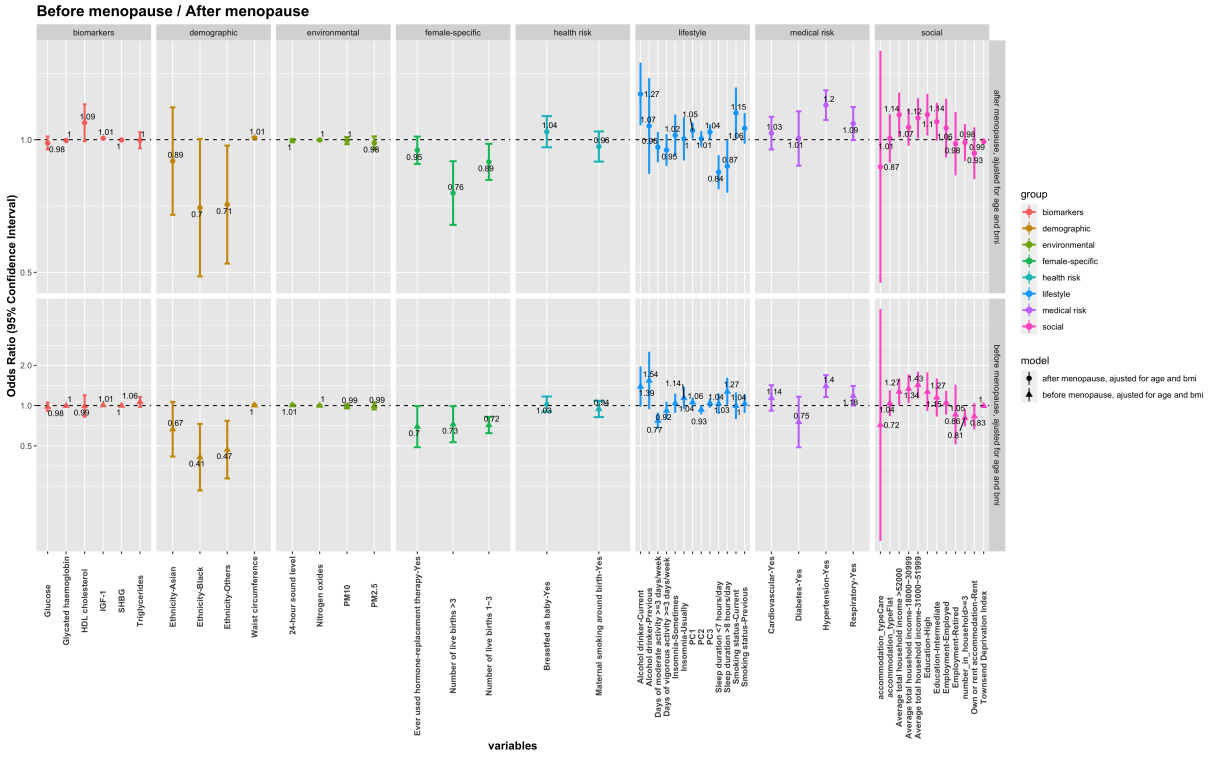


**Figure 26: Sensitivity analysis on menopause**

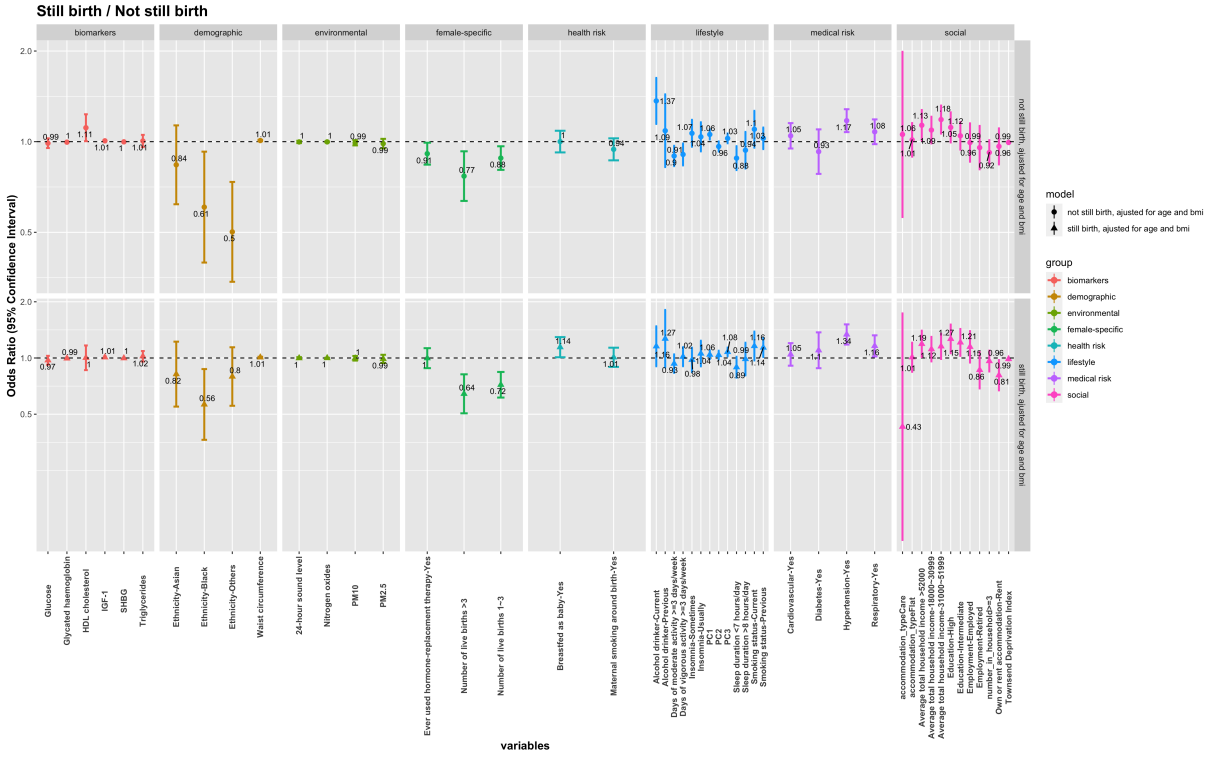


**Figure 27: Sensitivity analysis on history of stillbirth**

**Table 3: Odds ratio with 95% confidence intervals and p-values from the multivariate models for the risk of breast cancer. Results are represented for chronologically adjusted models. Baseline predictors are defined as demographic variables (Model 1), and models are additionally adjusted for lifestyle (Model 2), social variable (Model 3), female specific (Model 4), health risk (Model 5), biomarkers (Model 6), medical risk (Model 7) and environment exposure (Model 8).**

| | | Model 1 | | Model 2 | | Model 3 | | Model 4 | | Model 5 | | Model 6 | | Model 7 | | Model 8 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Category Name | Coefficient Name | OR | p-value | OR | p-value | OR | p-value | OR | p-value | OR | p-value | OR | p-value | OR | p-value | OR | p-value |
| Demographic | recruitment_age | 1.02 [ 1.02 , 1.02 ] | 8.23E-25 | 1.02 [ 1.02 , 1.03 ] | 1.97E-25 | 1.03 [ 1.02 , 1.03 ] | 2.50E-18 | 1.03 [ 1.02 , 1.04 ] | 1.70E-12 | 1.03 [ 1.02 , 1.04 ] | 2.81E-11 | 1.03 [ 1.02 , 1.04 ] | 4.57E-12 | 1.03 [ 1.02 , 1.03 ] | 2.34E-09 | 1.03 [ 1.02 , 1.03 ] | 2.20E-09 |
| Demographic | ethnicAsian | 0.83 [ 0.65 , 1.05 ] | 1.33E-01 | 0.94 [ 0.73 , 1.19 ] | 6.31E-01 | 0.96 [ 0.75 , 1.22 ] | 7.33E-01 | 0.98 [ 0.76 , 1.24 ] | 8.49E-01 | 0.95 [ 0.74 , 1.20 ] | 6.59E-01 | 0.95 [ 0.74 , 1.21 ] | 7.11E-01 | 0.94 [ 0.73 , 1.20 ] | 6.58E-01 | 0.95 [ 0.74 , 1.21 ] | 6.85E-01 |
| Demographic | ethnicBlack | 0.57 [ 0.42 , 0.76 ] | 2.50E-04 | 0.61 [ 0.45 , 0.82 ] | 1.84E-03 | 0.63 [ 0.46 , 0.85 ] | 3.28E-03 | 0.65 [ 0.47 , 0.87 ] | 6.01E-03 | 0.64 [ 0.47 , 0.86 ] | 4.30E-03 | 0.63 [ 0.46 , 0.85 ] | 3.82E-03 | 0.62 [ 0.45 , 0.84 ] | 2.88E-03 | 0.62 [ 0.46 , 0.84 ] | 3.23E-03 |
| Demographic | ethnicOthers | 0.62 [ 0.48 , 0.79 ] | 3.18E-04 | 0.66 [ 0.51 , 0.85 ] | 2.20E-03 | 0.67 [ 0.51 , 0.86 ] | 2.88E-03 | 0.68 [ 0.52 , 0.87 ] | 3.82E-03 | 0.67 [ 0.51 , 0.86 ] | 2.69E-03 | 0.67 [ 0.51 , 0.86 ] | 2.97E-03 | 0.67 [ 0.51 , 0.86 ] | 2.73E-03 | 0.67 [ 0.51 , 0.86 ] | 3.03E-03 |
| Demographic | waist_circumference | 1.00 [ 1.00 , 1.01 ] | 3.45E-02 | 1.00 [ 1.00 , 1.00 ] | 7.86E-02 | 1.00 [ 1.00 , 1.01 ] | 2.78E-02 | 1.00 [ 1.00 , 1.01 ] | 4.13E-02 | 1.01 [ 1.00 , 1.01 ] | 1.35E-04 | 1.01 [ 1.00 , 1.01 ] | 1.41E-04 | 1.01 [ 1.00 , 1.01 ] | 1.54E-03 | 1.01 [ 1.00 , 1.01 ] | 1.52E-03 |
| Lifestyle | num_moderate_activity>=3 | | | 0.92 [ 0.86 , 0.99 ] | 2.45E-02 | 0.94 [ 0.87 , 1.00 ] | 5.91E-02 | 0.94 [ 0.88 , 1.00 ] | 7.13E-02 | 0.94 [ 0.87 , 1.00 ] | 6.11E-02 | 0.94 [ 0.87 , 1.00 ] | 5.90E-02 | 0.94 [ 0.88 , 1.00 ] | 7.63E-02 | 0.94 [ 0.88 , 1.00 ] | 7.61E-02 |
| Lifestyle | num_vigorous_activity>=3 | | | 0.99 [ 0.92 , 1.06 ] | 7.32E-01 | 0.99 [ 0.92 , 1.06 ] | 7.57E-01 | 0.99 [ 0.92 , 1.06 ] | 7.75E-01 | 0.99 [ 0.92 , 1.06 ] | 8.01E-01 | 0.99 [ 0.92 , 1.06 ] | 7.90E-01 | 0.99 [ 0.92 , 1.06 ] | 8.08E-01 | 0.99 [ 0.92 , 1.06 ] | 8.06E-01 |
| Lifestyle | sleep_duration<7 | | | 0.88 [ 0.81 , 0.95 ] | 1.53E-03 | 0.88 [ 0.82 , 0.95 ] | 2.29E-03 | 0.88 [ 0.82 , 0.95 ] | 2.17E-03 | 0.89 [ 0.82 , 0.96 ] | 2.98E-03 | 0.89 [ 0.82 , 0.96 ] | 3.36E-03 | 0.88 [ 0.82 , 0.95 ] | 2.46E-03 | 0.89 [ 0.82 , 0.95 ] | 2.50E-03 |
| Lifestyle | sleep_duration>8 | | | 0.95 [ 0.85 , 1.06 ] | 3.97E-01 | 0.97 [ 0.87 , 1.09 ] | 6.59E-01 | 0.98 [ 0.87 , 1.09 ] | 6.99E-01 | 0.98 [ 0.87 , 1.10 ] | 7.37E-01 | 0.98 [ 0.88 , 1.10 ] | 7.74E-01 | 0.98 [ 0.87 , 1.10 ] | 7.30E-01 | 0.98 [ 0.87 , 1.10 ] | 7.28E-01 |
| Lifestyle | insomniaSometimes | | | 1.03 [ 0.95 , 1.12 ] | 4.34E-01 | 1.04 [ 0.95 , 1.13 ] | 3.81E-01 | 1.04 [ 0.95 , 1.13 ] | 3.77E-01 | 1.04 [ 0.95 , 1.13 ] | 3.72E-01 | 1.04 [ 0.95 , 1.13 ] | 4.01E-01 | 1.03 [ 0.95 , 1.12 ] | 4.50E-01 | 1.03 [ 0.95 , 1.12 ] | 4.51E-01 |
| Lifestyle | insomniaUsually | | | 1.07 [ 0.97 , 1.17 ] | 1.83E-01 | 1.08 [ 0.98 , 1.18 ] | 1.32E-01 | 1.08 [ 0.98 , 1.18 ] | 1.23E-01 | 1.08 [ 0.98 , 1.18 ] | 1.07E-01 | 1.08 [ 0.98 , 1.18 ] | 1.24E-01 | 1.07 [ 0.97 , 1.17 ] | 1.92E-01 | 1.07 [ 0.97 , 1.17 ] | 1.93E-01 |
| Lifestyle | alcohol_drinkerPrevious | | | 1.08 [ 0.87 , 1.34 ] | 4.85E-01 | 1.07 [ 0.86 , 1.32 ] | 5.55E-01 | 1.07 [ 0.85 , 1.32 ] | 5.72E-01 | 1.07 [ 0.86 , 1.33 ] | 5.44E-01 | 1.07 [ 0.86 , 1.33 ] | 5.38E-01 | 1.07 [ 0.86 , 1.32 ] | 5.64E-01 | 1.07 [ 0.85 , 1.32 ] | 5.67E-01 |
| Lifestyle | alcohol_drinkerCurrent | | | 1.17 [ 1.01 , 1.35 ] | 3.89E-02 | 1.14 [ 0.98 , 1.32 ] | 8.96E-02 | 1.13 [ 0.98 , 1.31 ] | 1.02E-01 | 1.13 [ 0.97 , 1.31 ] | 1.13E-01 | 1.11 [ 0.96 , 1.29 ] | 1.62E-01 | 1.12 [ 0.96 , 1.30 ] | 1.41E-01 | 1.12 [ 0.96 , 1.30 ] | 1.43E-01 |
| Lifestyle | smoking_statusPrevious | | | 1.03 [ 0.96 , 1.10 ] | 4.31E-01 | 1.03 [ 0.96 , 1.10 ] | 3.64E-01 | 1.04 [ 0.97 , 1.11 ] | 2.92E-01 | 1.03 [ 0.96 , 1.10 ] | 3.47E-01 | 1.03 [ 0.96 , 1.10 ] | 3.72E-01 | 1.03 [ 0.96 , 1.10 ] | 4.49E-01 | 1.03 [ 0.96 , 1.10 ] | 4.41E-01 |
| Lifestyle | smoking_statusCurrent | | | 1.09 [ 0.97 , 1.21 ] | 1.49E-01 | 1.11 [ 0.99 , 1.24 ] | 7.49E-02 | 1.12 [ 1.00 , 1.25 ] | 5.57E-02 | 1.11 [ 0.99 , 1.24 ] | 7.99E-02 | 1.12 [ 1.00 , 1.25 ] | 5.59E-02 | 1.11 [ 0.99 , 1.24 ] | 8.70E-02 | 1.11 [ 0.99 , 1.24 ] | 8.50E-02 |
| Lifestyle | PC1 | | | 1.03 [ 1.00 , 1.06 ] | 6.11E-02 | 1.04 [ 1.00 , 1.07 ] | 2.79E-02 | 1.04 [ 1.01 , 1.08 ] | 1.40E-02 | 1.04 [ 1.01 , 1.08 ] | 1.19E-02 | 1.04 [ 1.01 , 1.07 ] | 2.39E-02 | 1.04 [ 1.00 , 1.07 ] | 2.65E-02 | 1.04 [ 1.00 , 1.07 ] | 2.72E-02 |
| Lifestyle | PC2 | | | 1.00 [ 0.97 , 1.03 ] | 9.24E-01 | 1.00 [ 0.96 , 1.03 ] | 8.02E-01 | 1.00 [ 0.97 , 1.03 ] | 8.53E-01 | 1.00 [ 0.97 , 1.03 ] | 8.63E-01 | 0.99 [ 0.96 , 1.03 ] | 7.26E-01 | 0.99 [ 0.96 , 1.02 ] | 6.39E-01 | 0.99 [ 0.96 , 1.02 ] | 6.41E-01 |
| Lifestyle | PC3 | | | 1.03 [ 1.00 , 1.06 ] | 9.15E-02 | 1.02 [ 0.99 , 1.05 ] | 2.47E-01 | 1.02 [ 0.98 , 1.05 ] | 3.29E-01 | 1.02 [ 0.98 , 1.05 ] | 3.37E-01 | 1.01 [ 0.98 , 1.05 ] | 4.06E-01 | 1.02 [ 0.98 , 1.05 ] | 3.53E-01 | 1.02 [ 0.98 , 1.05 ] | 3.53E-01 |
| Social_variable | ave_total_household_income18000~30999 | | | | | 1.06 [ 0.97 , 1.15 ] | 1.96E-01 | 1.05 [ 0.96 , 1.15 ] | 2.45E-01 | 1.05 [ 0.96 , 1.15 ] | 2.55E-01 | 1.05 [ 0.96 , 1.15 ] | 2.68E-01 | 1.05 [ 0.96 , 1.15 ] | 2.57E-01 | 1.05 [ 0.96 , 1.15 ] | 2.58E-01 |
| Social_variable | ave_total_household_income31000~51999 | | | | | 1.11 [ 1.01 , 1.22 ] | 3.38E-02 | 1.10 [ 1.00 , 1.21 ] | 4.86E-02 | 1.10 [ 1.00 , 1.21 ] | 5.07E-02 | 1.10 [ 1.00 , 1.21 ] | 5.49E-02 | 1.11 [ 1.00 , 1.21 ] | 4.31E-02 | 1.11 [ 1.00 , 1.21 ] | 4.36E-02 |
| Social_variable | ave_total_household_income>52000 | | | | | 1.08 [ 0.97 , 1.20 ] | 1.70E-01 | 1.07 [ 0.96 , 1.19 ] | 2.28E-01 | 1.07 [ 0.96 , 1.18 ] | 2.52E-01 | 1.06 [ 0.95 , 1.18 ] | 2.76E-01 | 1.07 [ 0.96 , 1.19 ] | 2.32E-01 | 1.07 [ 0.96 , 1.19 ] | 2.33E-01 |
| Social_variable | tdi | | | | | 1.00 [ 0.99 , 1.01 ] | 7.04E-01 | 1.00 [ 0.98 , 1.01 ] | 5.88E-01 | 1.00 [ 0.99 , 1.01 ] | 6.77E-01 | 1.00 [ 0.99 , 1.01 ] | 7.34E-01 | 1.00 [ 0.99 , 1.01 ] | 6.59E-01 | 1.00 [ 0.99 , 1.01 ] | 8.19E-01 |
| Social_variable | own_or_rentrent | | | | | 0.96 [ 0.84 , 1.08 ] | 4.98E-01 | 0.98 [ 0.86 , 1.11 ] | 7.39E-01 | 0.98 [ 0.87 , 1.11 ] | 8.08E-01 | 0.99 [ 0.87 , 1.12 ] | 8.67E-01 | 0.98 [ 0.86 , 1.11 ] | 7.58E-01 | 0.98 [ 0.86 , 1.11 ] | 7.32E-01 |
| Social_variable | accommodation_typeFlat | | | | | 1.07 [ 0.95 , 1.20 ] | 2.70E-01 | 1.04 [ 0.92 , 1.16 ] | 5.25E-01 | 1.04 [ 0.92 , 1.16 ] | 5.39E-01 | 1.04 [ 0.92 , 1.16 ] | 5.30E-01 | 1.05 [ 0.93 , 1.17 ] | 4.51E-01 | 1.05 [ 0.93 , 1.18 ] | 4.05E-01 |
| Social_variable | accommodation_typeCare home/temporary or sheltered | | | | | 0.94 [ 0.53 , 1.63 ] | 8.32E-01 | 0.93 [ 0.52 , 1.62 ] | 8.15E-01 | 0.94 [ 0.53 , 1.63 ] | 8.30E-01 | 0.94 [ 0.53 , 1.64 ] | 8.42E-01 | 0.93 [ 0.52 , 1.62 ] | 8.08E-01 | 0.93 [ 0.52 , 1.62 ] | 8.04E-01 |
| Social_variable | number_in_household>=3 | | | | | 0.92 [ 0.85 , 0.99 ] | 4.01E-02 | 0.98 [ 0.90 , 1.06 ] | 6.65E-01 | 0.98 [ 0.90 , 1.06 ] | 6.13E-01 | 0.98 [ 0.90 , 1.06 ] | 6.26E-01 | 0.98 [ 0.90 , 1.06 ] | 6.84E-01 | 0.98 [ 0.90 , 1.07 ] | 7.00E-01 |
| Social_variable | employmentEmployed | | | | | 0.99 [ 0.88 , 1.11 ] | 8.90E-01 | 0.98 [ 0.87 , 1.10 ] | 7.90E-01 | 0.98 [ 0.87 , 1.10 ] | 7.83E-01 | 0.98 [ 0.87 , 1.10 ] | 7.63E-01 | 0.99 [ 0.88 , 1.11 ] | 8.80E-01 | 0.99 [ 0.88 , 1.11 ] | 8.87E-01 |
| Social_variable | employmentRetired | | | | | 0.90 [ 0.79 , 1.02 ] | 1.24E-01 | 0.90 [ 0.78 , 1.02 ] | 1.13E-01 | 0.89 [ 0.78 , 1.02 ] | 1.04E-01 | 0.89 [ 0.78 , 1.01 ] | 9.40E-02 | 0.89 [ 0.77 , 1.01 ] | 9.07E-02 | 0.89 [ 0.77 , 1.01 ] | 8.97E-02 |
| Social_variable | educationIntermediate | | | | | 1.06 [ 0.96 , 1.15 ] | 2.42E-01 | 1.04 [ 0.95 , 1.14 ] | 4.05E-01 | 1.03 [ 0.94 , 1.13 ] | 4.69E-01 | 1.03 [ 0.94 , 1.13 ] | 5.11E-01 | 1.04 [ 0.95 , 1.14 ] | 4.17E-01 | 1.04 [ 0.95 , 1.14 ] | 4.18E-01 |
| Social_variable | educationHigh | | | | | 1.12 [ 1.01 , 1.23 ] | 3.86E-02 | 1.08 [ 0.98 , 1.20 ] | 1.36E-01 | 1.07 [ 0.96 , 1.18 ] | 2.11E-01 | 1.06 [ 0.96 , 1.18 ] | 2.51E-01 | 1.08 [ 0.97 , 1.19 ] | 1.73E-01 | 1.08 [ 0.97 , 1.19 ] | 1.69E-01 |
| Female_specific | menarche_age | | | | | | | 0.99 [ 0.97 , 1.01 ] | 2.23E-01 | 0.98 [ 0.97 , 1.00 ] | 1.21E-01 | 0.99 [ 0.97 , 1.00 ] | 1.40E-01 | 0.99 [ 0.97 , 1.00 ] | 1.67E-01 | 0.99 [ 0.97 , 1.00 ] | 1.67E-01 |
| Female_specific | menopause_statusYes | | | | | | | 1.05 [ 0.94 , 1.18 ] | 3.66E-01 | 1.06 [ 0.94 , 1.18 ] | 3.57E-01 | 1.05 [ 0.93 , 1.17 ] | 4.41E-01 | 1.06 [ 0.94 , 1.18 ] | 3.35E-01 | 1.06 [ 0.94 , 1.18 ] | 3.38E-01 |
| Female_specific | live_births_num1~3 | | | | | | | 0.84 [ 0.77 , 0.91 ] | 5.26E-05 | 0.84 [ 0.77 , 0.91 ] | 4.90E-05 | 0.84 [ 0.77 , 0.91 ] | 5.20E-05 | 0.84 [ 0.77 , 0.91 ] | 5.15E-05 | 0.84 [ 0.77 , 0.91 ] | 4.74E-05 |
| Female_specific | live_births_num>3 | | | | | | | 0.78 [ 0.67 , 0.91 ] | 1.65E-03 | 0.78 [ 0.67 , 0.90 ] | 1.53E-03 | 0.78 [ 0.67 , 0.91 ] | 1.89E-03 | 0.78 [ 0.67 , 0.90 ] | 1.49E-03 | 0.78 [ 0.67 , 0.90 ] | 1.42E-03 |
| Female_specific | stillbirth_statusYes | | | | | | | 0.98 [ 0.92 , 1.05 ] | 6.35E-01 | 0.98 [ 0.92 , 1.05 ] | 6.36E-01 | 0.98 [ 0.92 , 1.05 ] | 6.39E-01 | 0.98 [ 0.92 , 1.05 ] | 5.61E-01 | 0.98 [ 0.92 , 1.05 ] | 5.65E-01 |
| Female_specific | oc_statusYes | | | | | | | 1.08 [ 0.99 , 1.17 ] | 8.92E-02 | 1.08 [ 0.99 , 1.17 ] | 8.37E-02 | 1.07 [ 0.99 , 1.16 ] | 9.95E-02 | 1.07 [ 0.99 , 1.17 ] | 9.25E-02 | 1.07 [ 0.99 , 1.16 ] | 9.41E-02 |
| Female_specific | HRT_statusYes | | | | | | | 0.91 [ 0.85 , 0.98 ] | 1.25E-02 | 0.91 [ 0.85 , 0.98 ] | 1.26E-02 | 0.92 [ 0.85 , 0.98 ] | 1.53E-02 | 0.92 [ 0.85 , 0.98 ] | 1.70E-02 | 0.92 [ 0.85 , 0.98 ] | 1.68E-02 |
| Health_risk | breastfed_as_babyYes | | | | | | | | | 1.05 [ 0.98 , 1.12 ] | 1.64E-01 | 1.05 [ 0.98 , 1.12 ] | 1.71E-01 | 1.05 [ 0.98 , 1.12 ] | 1.66E-01 | 1.05 [ 0.98 , 1.12 ] | 1.64E-01 |
| Health_risk | maternal_smokingYes | | | | | | | | | 0.95 [ 0.89 , 1.02 ] | 1.69E-01 | 0.95 [ 0.89 , 1.02 ] | 1.80E-01 | 0.95 [ 0.89 , 1.02 ] | 1.70E-01 | 0.95 [ 0.89 , 1.02 ] | 1.70E-01 |
| Health_risk | BMI25-30 | | | | | | | | | 0.95 [ 0.88 , 1.03 ] | 2.62E-01 | 0.95 [ 0.87 , 1.03 ] | 2.19E-01 | 0.95 [ 0.87 , 1.03 ] | 2.17E-01 | 0.95 [ 0.87 , 1.03 ] | 2.17E-01 |
| Health_risk | BMI30-40 | | | | | | | | | 0.84 [ 0.74 , 0.96 ] | 1.29E-02 | 0.85 [ 0.74 , 0.96 ] | 1.58E-02 | 0.84 [ 0.73 , 0.95 ] | 1.03E-02 | 0.84 [ 0.73 , 0.95 ] | 1.02E-02 |
| Health_risk | BMI>40 | | | | | | | | | 0.59 [ 0.44 , 0.78 ] | 3.82E-04 | 0.60 [ 0.45 , 0.80 ] | 7.25E-04 | 0.59 [ 0.44 , 0.78 ] | 3.95E-04 | 0.59 [ 0.44 , 0.78 ] | 3.96E-04 |
| Biomarkers | bio_triglycerides | | | | | | | | | | | 1.00 [ 0.96 , 1.04 ] | 9.95E-01 | 1.00 [ 0.96 , 1.04 ] | 9.58E-01 | 1.00 [ 0.96 , 1.04 ] | 9.49E-01 |
| Biomarkers | bio_hba1c | | | | | | | | | | | 1.00 [ 0.99 , 1.00 ] | 3.87E-01 | 1.00 [ 0.99 , 1.00 ] | 3.01E-01 | 1.00 [ 0.99 , 1.00 ] | 3.01E-01 |
| Biomarkers | bio_glucose | | | | | | | | | | | 0.99 [ 0.95 , 1.02 ] | 5.19E-01 | 0.98 [ 0.95 , 1.02 ] | 4.12E-01 | 0.98 [ 0.95 , 1.02 ] | 4.10E-01 |
| Biomarkers | bio_HDL | | | | | | | | | | | 1.08 [ 0.99 , 1.19 ] | 9.69E-02 | 1.09 [ 0.99 , 1.19 ] | 8.41E-02 | 1.09 [ 0.99 , 1.19 ] | 8.43E-02 |
| Biomarkers | bio_SHBG | | | | | | | | | | | 1.00 [ 1.00 , 1.00 ] | 4.62E-02 | 1.00 [ 1.00 , 1.00 ] | 4.94E-02 | 1.00 [ 1.00 , 1.00 ] | 4.92E-02 |
| Biomarkers | bio_IGF1 | | | | | | | | | | | 1.01 [ 1.00 , 1.01 ] | 4.99E-02 | 1.01 [ 1.00 , 1.01 ] | 4.72E-02 | 1.01 [ 1.00 , 1.01 ] | 4.77E-02 |
| Medical_risk | family_historyYes | | | | | | | | | | | | | 1.47 [ 1.34 , 1.62 ] | 1.17E-14 | 1.47 [ 1.34 , 1.62 ] | 1.17E-14 |
| Medical_risk | CardiovascularYes | | | | | | | | | | | | | 0.99 [ 0.91 , 1.08 ] | 8.52E-01 | 0.99 [ 0.91 , 1.08 ] | 8.53E-01 |
| Medical_risk | HypertensionYes | | | | | | | | | | | | | 1.26 [ 1.17 , 1.36 ] | 2.54E-09 | 1.26 [ 1.17 , 1.36 ] | 2.52E-09 |
| Medical_risk | DiabetesYes | | | | | | | | | | | | | 0.99 [ 0.85 , 1.15 ] | 9.48E-01 | 1.00 [ 0.85 , 1.15 ] | 9.52E-01 |
| Medical_risk | RespiratoryYes | | | | | | | | | | | | | 1.09 [ 1.01 , 1.18 ] | 2.73E-02 | 1.09 [ 1.01 , 1.18 ] | 2.72E-02 |
| Environment_exposure | ap_Nitrogen_oxides | | | | | | | | | | | | | | | 1.00 [ 0.99 , 1.00 ] | 5.00E-01 |
| Environment_exposure | ap_pm10 | | | | | | | | | | | | | | | 1.00 [ 0.98 , 1.02 ] | 9.49E-01 |
| Environment_exposure | ap_pm2.5 | | | | | | | | | | | | | | | 1.01 [ 0.95 , 1.07 ] | 6.82E-01 |
| Environment_exposure | np_ave_24.hour_sound_level | | | | | | | | | | | | | | | 1.00 [ 0.99 , 1.01 ] | 9.65E-01 |

# Appendix: Response of Group 5 to the comments from the peer review

Overall, very interesting!

- Family history + menopause + subtype analysis is very interesting

Good comments ! Thank you very much !

- Imputation process explained well – justification needed for chosen biomarkers

Good comments ! Thank you very much !

- Doing a Cox regression sounds like a really good idea, but it would be redundant to have both logistic and Cox regression

Yes, we do not plan to conduct Cox regression this time.

Inclusion/Exclusion criteria

- What is justification for excluding males? – report breast cancer cases in males

This is because the number of males with breast cancer is too few in our dataset. It limits our further analysis, like subtype analysis.

- Consider plotting the distributions of age at diagnosis and time to diagnosis
    - o Potential need to exclude people diagnosed within a year of study entry or sensitivity analysis stratifying groups by time to diagnosis/age at diagnosis

Yes, we consider the factor of time to diagnosis. We remove people who had breast cancer before recruitment and developed breast cancer within one year since recruitment. Doing this in order to reduce the possibility of associations distortion.

- Did you exclude all other cancer incidence or all prevalent cancer cases?
    - o Excluding all other cancer diagnosis during follow-up may result in super healthy control population, not representative of the general population

We exclude all prevalent cancer cases. Good points! Yes, it would lose somewhat representativeness because of "healthy-control". But if we did not exclude them, it would also introduce bias, like people with cancer or bad illness are more likely to expose risk factors. This is a process of weighing pros and cons.

- Confirm whether you excluded participants who withdrew consent

Yes, we confirm we had excluded participants who withdrew consent.

Tables:

- Table cuts off abruptly: e.g., sociodemographic factors should be grouped together

It's not an abrupt cut-off. We separated sociodemographic factors to social group and demographic group, which was the same with Marc's paper in risk factors for Covid-19. Besides, if we include all factors in only one group, it is quite a lot in our research.

- Highlight p-values below threshold

We have modified.

- Spelling/grammar mistakes:

o 'Daytime of noise pollution' --> Daytime noise pollution or Daytime noise level
o '16-hour sound level of noise pollution' --> 16-hour noise level (etc.)
o 'Not have family breast cancer history' --> No family breast cancer history
o 'Not have stillbirth' --> 'Not had stillbirth'
o "Premenopausal" vs. "Postmenopausal"

We have modified.

LASSO:

- How did you adjust for age and BMI in the multivariate analyses?

We did not adjust for age and BMI in logistic lasso.

- Did you consider a stability analysis or looking at prediction performances (AUC)?

We considered to split the dataset into train and test part (80%: train, 20%: test). Then we conducted ROC curve and calculated AUC.

Sensitivity analysis:

- Have you run any regressions for sensitivity analyses? – better to go beyond t-test

Yes, we added.

- Justify sensitivity analysis of stillbirth vs. Non stillbirth
    - Limits to only participants with experience of childbirth
- Do you have ages at first birth or is this what the sensitivity analysis of stillbirth is trying to capture?

We read articles about the associations between stillbirth (pregnancy loss) and breast cancer. But this issue is still controversy. Therefore, we conducted sensitivity analysis to see whether there are differences between them.

- Would be interesting to see sensitivity analysis for: Oral contraceptive use, Age at menarche, Stage of diagnosis, Socioeconomic status indicators etc.

Finally, we included sensitivity analysis for Oral contraceptive use, Age at menarche, menopause, bc history and stillbirth.

Discussion

- Remember to contextualize in terms of breast cancer screening programs

Thank you for your suggestion.

- Around 10% of breast cancer cases in the UK are diagnosed at a later stage which is pretty good when compared to other countries

Thank you.

- You can be pretty sure that you've got valid breast cancer numbers, i.e., less likely to have unreported cases

Thank you.

Other suggestions:

- Rename female-specific as reproductive

We think using female-specific more accurate. This group not only includes reproductive factors.

- Dietary intake composite scores: losing more than 50% of information; consider including more PCs

PCA is a widely used method in nutritional epidemiology to derive dietary patterns from habitual diet. What we showed in percentage of variance explained was in line with most similar research. Although losing more than 50% of information, this method was still useful because it did not aim to capture as much as information but it could provide joint/combined effects from different diet items.

- Use a correlation plot (circle plot with arrows) to show contribution of variables to each PC

Thanks for your suggestion. We used 3 PCs, so it's hard to visualize in only one plot.

- Forest plots: Use different colour darkness etc. and jitter points to better distinguish between models

We have modified.