# SUPPLEMENTARY MATERIALS

**Supplementary Methods**

Let D indicates breast cancer status (1=Yes, 0=No), Age[i] indicates *i*thage groups, and **X** indicate all other risk factors (covariates). In population-based cohort studies, a logistic regression model can be used to model disease risk:

$logit\left[Pr\left(Age\_{i},X\right)\right]$ = $log\left[\frac{Pr\left(Age\_{i},X\right)}{Pr\left(Age\_{i},X\right)}\right]$

$$=α\_{0}+α\_{Age\_{1}}×Age\_{1}+α\_{Age\_{2}}×Age\_{2}+…+α\_{Age\_{i}}×Age\_{i}+γ\_{cohort}X [1]$$

In case-control studies, let S indicates whether an individual in the population could be selected as case/control (S=1 if selected and S=0 if not). Then

$$logit\left[Pr\left(Age\_{i},X,S\right)\right]=β\_{0}+β\_{Age\_{1}}×Age\_{1}+β\_{Age\_{2}}×Age\_{2}+…+β\_{Age\_{i}}×Age\_{i}+γ\_{c-c}X [2]$$

Based on Bayes’ rule,

$$Pr\left(Age\_{i},X\right)=Pr\left(D=1,Age\_{i},X\right)×Pr\left(Age\_{i},X\right) [3]$$

We apply equation 3 to equation 2, and get

$$Logit\left[Pr\left(Age\_{i},X,S\right)\right]=log\left[\frac{Pr\left(Age\_{i},X,S\right)}{Pr\left(Age\_{i},X,S\right)}\right] =log\left[\frac{Pr\left(Age\_{i},X\right)}{Pr\left(Age\_{i},X\right)}\right]=log\left[\frac{Pr\left(D=1,Age\_{i},X\right)×Pr\left(Age\_{i},X\right) }{Pr\left(D=0,Age\_{i},X\right)×Pr\left(Age\_{i},X\right) }\right]=log\left[\frac{Pr\left(D=1,Age\_{i},X\right)}{Pr\left(D=0,Age\_{i},X\right)}\right]+log\left[\frac{Pr\left(Age\_{i},X\right)}{Pr\left(Age\_{i},X\right)}\right] [4]$$

This is equivalent to

$$α\_{0}+α\_{Age\_{1}}×Age\_{1}+α\_{Age\_{2}}×Age\_{2}+…+α\_{Age\_{i}}×Age\_{i}+γ\_{cohort}X=β\_{0}+β\_{Age\_{1}}×Age\_{1}+β\_{Age\_{2}}×Age\_{2}+…+β\_{Age\_{i}}×Age\_{i}+γ\_{c-c}X-log\left[\frac{Pr\left(D=1,Age\_{i},X\right)}{Pr\left(D=0,Age\_{i},X\right)}\right]$$

Here we need the assumption that the selection of both case and control are not associated with any of the covariates (**X**), which means $Pr\left(D=1,Age\_{i},X\right)=Pr\left(D=1,Age\_{i}\right)$ and $Pr\left(D=0,Age\_{i},X\right)=Pr\left(D=0,Age\_{i}\right)$. This assumption is reasonable for case-control studies that are not matched on covariates, such as our study. Then,

$$ α\_{0}+α\_{Age\_{1}}×Age\_{1}+α\_{Age\_{2}}×Age\_{2}+…+α\_{Age\_{i}}×Age\_{i}+γ\_{cohort}X=β\_{0}+β\_{Age\_{1}}×Age\_{1}+β\_{Age\_{2}}×Age\_{2}+…+β\_{Age\_{i}}×Age\_{i}+γ\_{c-c}X-log\left[\frac{Pr\left(D=1,Age\_{i}\right)}{Pr\left(D=0,Age\_{i}\right)}\right] [5] $$

Let $D\_{i}$ and $C\_{i}$ indicate the total numbers of cases and unaffected controls in the age group *i* of the population. Incidence rate in age group *i* can be expressed as $\frac{D\_{i}}{D\_{i}+C\_{i}}.$ The numbers of selected cases and controls in age group *i* for a case-control study are $d\_{i}$ and $c\_{i}$. The probabilities of being selected as cases and controls are $\frac{d\_{i}}{D\_{i}}$ and $\frac{c\_{i}}{C\_{i}}$ , respectively.

Then, applying these to equation 5, we have

$$α\_{0}+α\_{Age\_{1}}×Age\_{1}+α\_{Age\_{2}}×Age\_{2}+…+α\_{Age\_{i}}×Age\_{i}+γ\_{cohort}X=β\_{0}+β\_{Age\_{1}}×Age\_{1}+β\_{Age\_{2}}×Age\_{2}+…+β\_{Age\_{i}}×Age\_{i}+γ\_{c-c}X-log\left[\frac{d\_{i}}{c\_{i}}×\frac{\left(1-Incidence\_{age\_{i}}\right)}{Incidence\_{age\_{i}}}\right] [6] $$

It is well-known that log odds ratios from a case-control study ($γ\_{c-c})$ and log odds ratios from a cohort study ($γ\_{cohort})$ are the same (1). Then we can use the adjustment term, $-log\left[\frac{d\_{i}}{c\_{i}}×\frac{\left(1-Incidence\_{age\_{i}}\right)}{Incidence\_{age\_{i}}}\right]$, to derive the intercept ($α\_{0}$) and coefficients for age groups ($α\_{Age\_{i}}$) in the cohort logistic model, i.e. $α\_{0}$ = $β\_{0}-log\left[\frac{d\_{0}}{c\_{0}}×\frac{\left(1-Incidence\_{age\_{0}}\right)}{Incidence\_{age\_{0}}}\right]$, $α\_{Age\_{1}}= β\_{Age\_{1}}-log\left[\frac{d\_{1}}{c\_{1}}×\frac{\left(1-Incidence\_{age\_{1}}\right)}{Incidence\_{age\_{1}}}\right]+log\left[\frac{d\_{0}}{c\_{0}}×\frac{\left(1-Incidence\_{age\_{0}}\right)}{Incidence\_{age\_{0}}}\right]$. Set age group of 40-45 year-olds as the reference group ($β\_{Age\_{5}}=0$). Age-specific breast cancer incidences were obtained from Ibadan Cancer Registry via personal communication with Drs. Maxwell Parkin and Olufemi Ogunbiyi and these rates will be published in the IARC monographic book Cancer in Africa, volume II. **Supplementary** **Table 2** illustrates the procedure to calculate new intercept $α\_{0}$ and $α\_{Age\_{i}}$.

To consider competing risk from other diseases, we obtained female Nigerian mortality rates from World Health Organization website (<http://apps.who.int/gho/data/view.main.61200?lang=en>, Assessed on 3/21/2016). Mortality rates in 2013 were used (**Supplementary** **Table 3**). We calculated the competing mortality rates as total mortality rate minus breast cancer-specific mortality rate. Since breast cancer-specific mortality in Nigeria is not available, we used breast cancer mortality-to-incidence rate ratio (MR:IR) for African countries, 0.69, according to Kamangar et al (2). Sensitivity analyses using different MR:IR gave very similar results because breast cancer mortality rates were relatively small compared to total mortality rates.

**References for the Supplementary Methods:**

1. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. Biometrika 1979;66(3):403-411.

2. Kamangar F, Dores GM, Anderson WF. Patterns of cancer incidence, mortality, and prevalence across five continents: defining priorities to reduce cancer disparities in different geographic regions of the world. J Clin Oncol 2006;24(14):2137-50.

# Supplementary Tables

**Supplementary Table 1**. The step by step procedure to establish the relative risk prediction model (N=2692)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | LR chi2 (df) for model | Akaike's information criterion (df) | Bayesian Information criterion (df) | The best model above each row | The difference among each pairs of LR chi2 (df) | *P* |
| M1 | 431.78 (47) | 3352.98 (48) | 3635.90 (48) | - | - | - |
| M2 | 428.67 (44) | 3350.08 (45)  | 3615.33 (45) | M1 | 3.11 (3) | 0.375 |
| M3 | 422.75 (39) | 3346.01 (40) | 3581.78 (40) | M2 | 5.92 (5) | 0.314 |
| M3C | 428.38 (42) | 3346.38 (43) | 3599.83 (43) | M3 | 5.63 (3) | 0.131 |
| M4 | 433.27 (32) | 3321.40 (33) | 3516.00 (33) | M3 | -10.52 (7) | 1.000 |
| M4C | 433.82 (35) | 3326.94 (36) | 3539.13 (36) | M4 | -0.55 (3) | 0.907 |
| M5 | 407.75 (25) | 3333.01 (26) | 3486.26 (26) | M4 | 25.52 (7) | 0.001 |
| M5A | 424.13 (26) | 3318.62 (27) | 3477.77 (27) | M4 | 9.13 (6) | 0.166 |
| M5C | 423.98 (28) | 3322.78 (29) | 3493.71 (29) | M5A | -0.16 (2) | 1.000 |
| M6 | 439.25 (30) | 3311.50 (31) | 3494.23 (31) | M5A | 15.12 (4) | 0.005 |
| M6C1 | 449.20 (33) | 3307.56 (34) | 3507.97 (34) | M6 | 9.94 (3) | 0.019 |
| M6A  | 429.96 (24) | 3308.80 (25) | 3456.16 (25) | M6 | 9.30 (6) | 0.158 |
| M6C2  | 440.14 (27) | 3304.61 (28) | 3469.65 (28) | M6A | 10.19 (3) | 0.017 |
| M7  | 445.66 (41) | 3327.09 (42) | 3574.66 (42) | M6A | 15.70 (17) | 0.545 |
| M7A | 436.62 (35) | 3324.13 (36) | 3536.33 (36) | M6A | 6.67 (11) | 0.825 |

**Note:**

Model 1: variables including current age (5-year interval categories, <25, 25~29.9, 30~34.9, 35~39.9, 40~44.9, 45~ 49.9, 50~54.9, 55~59.9, 60~64.9, 65~69.9, 70~74.9, ≥75), age at menarche (8~11.9, 12~12.9, 13~13.9, 14~14.9, 15~15.9, 16~16.9, ≥17), benign breast diseases (yes, no), family history of breast cancer in first degree relatives (yes, no), parity (0, 1, 2, 3, 4, 5, 6, 7, ≥8), age at first live birth (<20, 20~24.9, 25~29.9, ≥30, non-birth), total duration of breast feeding (<12 month, 12~23.9, 24~35.9, 36~47.9, 48~59.9, 60~71.9, 72~83.9, 84~95.9, ≥96), height (<150 cm, 150~159, 160~169, ≥170), body mass index (BMI, <18.5, 18.5~24.9, 25~29.9, ≥30 kg/m2), waist-hip ratio (<0.80, 0.80~0.84, ≥0.85), and alcohol consumption (yes, no).

Model 2: Based on Model 1, we excluded age at first live birth (<20, 20~24.9, 25~29.9, ≥30, non-birth) due to its non-significance in multi-variables logistic regression model.

Model 3: Based on Model 2, we changed age at menarche from categorical variable to continuous variable (per year).

Model 3C: Based on Model 2, we changed age at menarche from categorical variable to continuous variable (cubic spline with five knots).

Model 4: Based on Model 3, we changed total duration of breast feeding from categorical variable to continuous variable (every 12 months).

Model 4C: Based on Model 3, we changed total duration of breast feeding from categorical variable to continuous variable (cubic spline with five knots).

Model 5: Based on Model 4, we changed parity from categorical variable to continuous variable (per child).

Model 5A: Based on Model 4, we changed parity from categorical variable to continuous variable (two variables: first live birth (yes, no), each additional live birth (continuous variable))

Model 5C: Based on Model 4, we changed parity from categorical variable to continuous variable (cubic spline with five knots).

Model 6: Based on Model 4, we changed height from categorical variable to continuous variable ([height-160]/10).

Model 6C1: Based on Model 4, we changed height from categorical variable to continuous variable (cubic spline with five knots).

Model 6A: Based on Model 5A, we changed height from categorical variable to continuous variable ([height-160]/10).

Model 6C2: Based on Model 5A, we changed height from categorical variable to continuous variable (cubic spline with five knots).

Model 7: Based on Model 6, we added the interaction item between benign breast diseases and current age.

Model 7A: Based on Model 6A, we added the interaction item between benign breast diseases and current age.

The difference among each pairs of LR chi2 and their degree of freedom are in the upper right part of the table, corresponding p-values are in the lower left

Abbreviation: df, degree of freedom; LR, likelihood ratio

**Supplementary** **Table 2.**  Breast cancer incidence rates and calculation of incidence rate adjusted intercept

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Age | Incidence rate\* | # of case | # of control | Log adjustment term [ln] | β0 | βage | β0+ βage - Log term | α0 | αage |
| 20- | 0.000028 | 9 | 153 | 7.650 | 1.168 | -3.489 | -9.971 | -5.934 | -4.037 |
| 25- | 0.000087 | 59 | 212 | 8.070 | 1.168 | -1.600 | -8.502 | -5.934 | -2.568 |
| 30- | 0.000309 | 151 | 314 | 7.350 | 1.168 | -0.869 | -7.051 | -5.934 | -1.117 |
| 35- | 0.000529 | 273 | 327 | 7.364 | 1.168 | -0.150 | -6.346 | -5.934 | -0.412 |
| 40- | 0.000771 | 297 | 317 | 7.102 | 1.168 | 0 | -5.934 | -5.934 | 0 |
| 45- | 0.001324 | 306 | 263 | 6.777 | 1.168 | 0.350 | -5.259 | -5.934 | 0.675 |
| 50- | 0.001144 | 254 | 222 | 6.907 | 1.168 | 0.401 | -5.338 | -5.934 | 0.596 |
| 55- | 0.001439 | 176 | 145 | 6.736 | 1.168 | 0.621 | -4.947 | -5.934 | 0.987 |
| 60- | 0.000928 | 132 | 139 | 6.930 | 1.168 | 0.449 | -5.313 | -5.934 | 0.621 |
| 65- | 0.001075 | 75 | 67 | 6.947 | 1.168 | 0.658 | -5.121 | -5.934 | 0.813 |
| 70- | 0.000767 | 50 | 43 | 7.323 | 1.168 | 0.734 | -5.421 | -5.934 | 0.513 |
| 75+ | 0.000446 | 29 | 23 | 7.947 | 1.168 | 0.930 | -5.849 | -5.934 | 0.085 |

\*The incidence rates were obtained from Ibadan Cancer Registry via personal communication with Drs. Maxwell Parkin and Olufemi Ogunbiyi, and these rates will be published in the IARC monographic book Cancer in Africa, volume II.

**Supplementary** **Table 3**. Competing mortality rates used to estimate absolute risks in the Nigerian Breast Cancer Study (1998-2015)\*

| Age group | Breast cancerIncidence rate | Total mortality rate | BC-specific Mortality rate | Mortality rate excluding breast cancer |
| --- | --- | --- | --- | --- |
| 20- | 2.8 | 424.9 | 1.9 | 423.0 |
| 25- | 8.7 | 554.4 | 6.0 | 548.4 |
| 30- | 30.9 | 673.5 | 21.3 | 652.2 |
| 35- | 52.9 | 869.2 | 36.5 | 832.7 |
| 40- | 77.1 | 967.5 | 53.2 | 914.3 |
| 45- | 132.4 | 1057.0 | 91.4 | 965.6 |
| 50- | 114.4 | 1308.3 | 78.9 | 1229.4 |
| 55- | 143.9 | 1629.3 | 99.3 | 1530.0 |
| 60- | 92.8 | 2329.4 | 64.0 | 2265.4 |
| 65- | 107.5 | 3572.3 | 74.2 | 3498.1 |
| 70- | 76.7 | 5685.3 | 52.9 | 5632.4 |
| 75+ | 44.6 | 9000.1 | 30.8 | 8969.3 |

 \*All rates are presented in 100,000 person-years.

**Supplementary Table 4**. Multivariable regression parameters with 95% CIs for breast cancer in the training data set (N=2692)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variables | Coefficient | Chi-square | *P* value | OR | 95% Confidence Intervals |
| Intercept\*  | -5.547 | - | - | - | - |
| Age group\* (ref: 40~44.9) |  | 179.20 | <0.001 |  |  |
| <25 | -4.484 |  |  | 0.02 | 0.01-0.05 |
| 25~29.9 | -2.634 |  |  | 0.19 | 0.12-0.30 |
| 30~34.9 | -1.122 |  |  | 0.42 | 1.30-0.58 |
| 35~39.9 | -0.335 |  |  | 0.93 | 0.69-1.25 |
| 45~49.9 | 0.620 |  |  | 1.34 | 1.00-1.81 |
| 50~54.9 | 0.552 |  |  | 1.43 | 1.05-1.95 |
| 55~59.9 | 1.040 |  |  | 1.96 | 1.37-2.81 |
| 60~64.9 | 0.566 |  |  | 1.48 | 1.02-2.16 |
| 65~69.9 | 0.963 |  |  | 2.24 | 1.37-3.67 |
| 70~74.9 | 0.721 |  |  | 2.56 | 1.40-4.70 |
| ≥75 | 0.178 |  |  | 2.78 | 1.30-5.93 |
| Age at menarche (per year) | -0.050 | 6.00 | 0.014 | 0.95 | 0.91-0.99 |
| First live birth | -0.785 | 16.67 | <0.001 | 0.46 | 0.31-0.67 |
| Each additional live birth | 0.123 | 9.45 | 0.002 | 1.13 | 1.05-1.22 |
| Breastfeeding (per 12 months) | -0.134 | 35.54 | <0.001 | 0.87 | 0.84-0.91 |
| Benign breast diseases | 0.590 | 14.78 | <0.001 | 1.80 | 1.34-2.44 |
| Family history of breast cancer | 0.436 | 3.97 | 0.046 | 1.55 | 1.01-2.37 |
| Height-160 (per 10 cm)  | 0.519 | 66.71 | <0.001 | 1.68 | 1.48-1.90 |
| Body mass index (ref: 18.5~24.9 kg/m2) |  | 24.63 | <0.001 |  |  |
| <18.5 | 0.346 |  |  | 1.41 | 0.98-2.04 |
| 25~29.9 | -0.312 |  |  | 0.73 | 0.60-0.90 |
| ≥30 | -0.412 |  |  | 0.66 | 0.53-0.83 |
| Alcohol consumption | 0.442 | 7.04 | 0.008 | 1.56 | 1.12-2.16 |

\*Intercept and regression coefficients (odds ratios) for each age group have been adjusted using breast cancer incidence rates from Ibadan Cancer Registry.

**Supplementary Table 5**. Projected probability (%) of developing breast cancer until follow-up to 80 year-old

|  |  |  |
| --- | --- | --- |
| Initial age (year-old) | Year of follow-up | Initial relative risk  |
| 1 | 2 | 5 | 8 | 10 | 15 |
| 20 | 5 | 0.01 | 0.03 | 0.07 | 0.11 | 0.14 | 0.2 |
|  | 10 | 0.1 | 0.1 | 0.3 | 0.4 | 0.6 | 0.8 |
|  | 20 | 0.4 | 0.9 | 2.2 | 3.5 | 4.3 | 6.4 |
|  | 30 | 1.3 | 2.6 | 6.4 | 10.0 | 12.3 | 17.7 |
|  | 40 | 2.3 | 4.5 | 10.7 | 16.5 | 20.1 | 28.3 |
|  | 50 | 2.9 | 5.6 | 13.3 | 20.3 | 24.5 | 34.0 |
|  | 60 | 3.1 | 6.1 | 14.3 | 21.7 | 26.2 | 36.0 |
| 30 | 5 | 0.2 | 0.3 | 0.8 | 1.2 | 1.5 | 2.3 |
|  | 10 | 0.4 | 0.8 | 2.0 | 3.2 | 3.9 | 5.9 |
|  | 20 | 1.3 | 2.6 | 6.4 | 10.0 | 12.4 | 17.9 |
|  | 30 | 2.3 | 4.6 | 11.0 | 16.9 | 20.6 | 29.1 |
|  | 40 | 2.9 | 5.8 | 13.7 | 20.9 | 25.3 | 35.1 |
|  | 50 | 3.2 | 6.3 | 14.8 | 22.4 | 27.0 | 37.2 |
| 40 | 5 | 0.4 | 0.8 | 1.9 | 3.0 | 3.7 | 5.5 |
|  | 10 | 1.0 | 2.0 | 4.8 | 7.6 | 9.4 | 13.8 |
|  | 20 | 2.1 | 4.1 | 9.9 | 15.3 | 18.7 | 26.6 |
|  | 30 | 2.7 | 5.4 | 12.9 | 19.8 | 24.0 | 33.5 |
|  | 40 | 3.0 | 5.9 | 14.1 | 21.4 | 25.9 | 35.9 |
| 50 | 5 | 0.6 | 1.1 | 2.7 | 4.3 | 5.4 | 8.0 |
|  | 10 | 1.2 | 2.4 | 5.8 | 9.2 | 11.3 | 16.5 |
|  | 20 | 2.0 | 3.9 | 9.3 | 14.5 | 17.8 | 25.3 |
|  | 30 | 2.2 | 4.4 | 10.7 | 16.5 | 20.1 | 28.4 |
| 60 | 5 | 0.4 | 0.9 | 2.2 | 3.4 | 4.3 | 6.4 |
|  | 10 | 0.9 | 1.7 | 4.3 | 6.8 | 8.4 | 12.3 |
|  | 20 | 1.2 | 2.4 | 5.9 | 9.3 | 11.5 | 16.7 |
| 70 | 5 | 0.3 | 0.7 | 1.7 | 2.6 | 3.3 | 4.9 |
|  | 10 | 0.5 | 0.9 | 2.3 | 3.7 | 4.6 | 6.8 |

**Supplementary Table 6**. Reclassification of the BWHS, Gail-White, and Gail-Black models with the NBCS model in categories of 5-year breast cancer risk (N=2993)

|  |  |  |
| --- | --- | --- |
| 5-year risk (%) | 5-year risk (%): NBCS | Net reclassification improvement (%) |
| 0.00~0.89 | 0.90~1.65 | ≥1.66 |
| **BWHS model** |  |  | 8.26 |
| Cases |  |  |  | -3.90 |
| 0.00~0.89 | 918 | 117 | 19 |  |
| 0.90~1.65 | 227 | 142 | 49 |  |
| ≥1.66 | 3 | 14 | 24 |  |
| Controls  |   |   |   | 12.16 |
| 0.00~0.89 | 1101 | 46 | 5 |  |
| 0.90~1.65 | 226 | 73 | 10 |  |
| ≥1.66 | 6 | 9 | 4 |  |
| **Gail-White model** |  |  |  | 13.45 |
| Cases |  |  |  | 22.63 |
| 0.00~0.89 | 1124 | 261 | 78 |  |
| 0.90~1.65 | 16 | 11 | 14 |  |
| ≥1.66 | 0 | 0 | 0 |  |
| Controls  |  |  |  | -9.18 |
| 0.00~0.89 | 1322 | 123 | 15 |  |
| 0.90~1.65 | 8 | 4 | 4 |  |
| ≥1.66 | 0 | 0 | 0 |  |
| **Gail-Black model** |  |  | 14.19 |
| Cases |  |  |  | 23.99 |
| 0.00~0.89 | 1146 | 273 | 89 |  |
| 0.90~1.65 | 2 | 0 | 3 |  |
| ≥1.66 | 0 | 0 | 0 |  |
| Controls  |   |   |   | -9.80 |
| 0.00~0.89 | 1331 | 128 | 19 |  |
| 0.90~1.65 | 2 | 0 | 0 |  |
| ≥1.66 | 0 | 0 | 0 |  |

Abbreviation: BWHS: Black Women Health Study; NBCS: Nigerian Breast Cancer Study

We excluded the participants either younger than 35 (n = 898) or elder than 70 year-old (n = 145), to ensure the comparability among different models.

**Supplementary Table 7**. Distribution of the study population following the variables’ classification of the BWHS models (N=4036)

| Risk factors | **No. of case (%)** | **No. of control (%)** | **Total (%)** |
| --- | --- | --- | --- |
| Family history |  |  |  |
| None | 1650 (91. 1) | 2104 (94.6) | 3754 (93.0) |
| First-degree relative | 95 (5.3) | 68 (3.1) | 163 (4.0) |
| Second-degree relative | 66 (3.6) | 53 (2.4) | 119 (3.0) |
| Benign breast diseases |  |  |  |
| No | 1604 (88.6) | 2081 (93.5) | 3685 (91.3) |
|  Yes | 207 (11.4) | 144 (6.5) | 351 (8.7) |
| Current body mass index, kg/m2  |  |  |  |
| <20 | 937 (51.7) | 1211 (54.4) | 2148 (53.2) |
| 20-24 | 653 (36.1) | 746 (33.5) | 1399 (34.7) |
| ≥25 | 221 (12.2) | 268 (12.0) | 489 (12.1) |
| Age at menarche, years |  |  |  |
| <12 | 28 (1.6) | 49 (2.2) | 77 (1.9)  |
| 12-13 | 337 (18.6) | 390 (17.5) | 727 (18.0) |
| ≥14 | 1446 (79.9) | 1786 (80.3) | 3232 (80.1) |
| Age at first birth, years |  |  |  |
| Nulliparous or <25 | 1186 (65.9) | 1501 (67.6) | 2687 (66.8) |
| ≥25 | 614 (34.1) | 720 (32.4) | 1334 (33.2) |
| Oral contraceptive use |  |  |  |
| No | 1296 (71.7) | 1477 (66.6) | 2773 (68.9) |
| Yes | 512 (28.3) | 741 (33.4) | 1253 (31.1) |
| Height |  |  |  |
| <5 feet,5 inches | 1366 (73.8) | 1885 (84.7) | 3221 (79.8) |
| ≥5 feet,5 inches | 475 (26.2) | 340 (15.3) | 815 (20.2) |

**Supplementary Table 8**. Distribution of the study population following the variables’ classification of the Gail models (N=4036)

| Risk factors | **No. of case (%)** | **No. of control (%)** | **Total (%)** |
| --- | --- | --- | --- |
| Age group at menarche |  |  |  |
| ≥14 (0) | 1446 (79.9) | 1786 (80.3) | 3232 (80.1) |
| 12-13 (1) | 337 (18.6) | 390 (17.5) | 727 (18.0) |
| <12 (2) | 28 (1.6) | 49 (2.2) | 77 (1.9) |
| Number of previous breast biopsies |  |  |  |
|  Age <50 y |  |  |  |
|  0 (0) | 1073 (59.3) | 1584 (71.2) | 2657 (65.8) |
|  1 (1) | 22 (1.2) | 2 (0.1) | 24 (0.6) |
|  ≥2 (2) | 0 (0.0) | 0 (0.0) | 0 (0.0) |
|  Age ≥50 y |  |  |  |
|  0(0) | 705 (38.9) | 636 (28.6) | 1341 (33.2) |
|  1(1) | 11 (0.6) | 3 (0.1) | 14 (0.3) |
|  ≥2 (2) | 0 (0.0) | 0 (0.0) | 0 (0.0) |
| Age group at first live birth |  |  |  |
| <20 y (0) |  |  |  |
| Number of first relatives with breast cancer=0 (0) | 362 (20.0) | 511 (23.0) | 873 (21.6) |
| Number of first relatives with breast cancer=1 (1) | 26 (1.4) | 18 (0.8) | 44 (1.1) |
| Number of first relatives with breast cancer≥2 (2) | 0 (0.0) | 0 (0.0) | 0 (0.0) |
| 20-24 y (1) |  |  |  |
| Number of first relatives with breast cancer=0 (0) | 642 (35.5) | 813 (36.5) | 1455 (36.1) |
| Number of first relatives with breast cancer=1 (1) | 36 (2.0) | 14 (0.6) | 50 (1.2) |
| Number of first relatives with breast cancer≥2 (2) | 1 (0.1) | 0 (0.0) | 1 (0.1) |
| 25-29 y or nulliparous(2) |  |  |  |
| Number of first relatives with breast cancer=0 (0) | 521 (28.8) | 660 (29.7) | 1181 (29.3) |
| Number of first relatives with breast cancer=1 (1) | 28 (1.6) | 31 (1.4) | 59 (1.5) |
| Number of first relatives with breast cancer≥2 (2) | 0 (0.0) | 0 (0.0) | 0 (0.0) |
| ≥30 y (3) |  |  |  |
| Number of first relatives with breast cancer=0 (0) | 191 (10.6) | 173 (7.8) | 364(9) |
| Number of first relatives with breast cancer=1 (1) | 4 (0.2) | 5 (0.2) | 9(0.2) |
| Number of first relatives with breast cancer≥2 (2) | 0 (0.0) | 0(0.0) | 0(0.0) |

# Supplementary Figures



**Supplementary Figure 1**. Variables included in Gail model for white women, Gail model for African American women, model based on Black women health study, and model based on Nigerian breast cancer study.



**Supplementary Figure 2**. Plots summarizing the refinement of 5-year risk predictions by comparing observed proportion of women with breast cancer to the proportion expected within deciles of predicted risk. We excluded the participants either younger than 35 (n = 898) or elder than 70 year-old (n = 145), to ensure the comparability among different models. If a model is well calibrated, open circles should fall along dashed line to indicate that the observed proportion of patients in each risk group agrees closely with the proportions predicted from the model. P values were calculated from the Hosmer-Lemeshow test, with p<0.05 indicating significant difference between observed and expected numbers. BWHS: Black Women Health Study; NBCS: Nigerian Breast Cancer Study