# Cancer Epidemiology, Biomarkers & Prevention

## Estrogen Metabolism and Exposure in a Genotypic–Phenotypic Model for Breast Cancer Risk Prediction

Philip S. Crooke, Christina Justenhoven, Hiltrud Brauch, et al.

| | |
|---|---|
| **Updated Version** | Access the most recent version of this article at:<br>doi:10.1158/1055-9965.EPI-11-0060 |

| | |
|---|---|
| **Cited Articles** | This article cites 76 articles, 32 of which you can access for free at:<br>http://cebp.aacrjournals.org/content/20/7/1502.full.html#ref-list-1 |

| | |
|---|---|
| **E-mail alerts** | Sign up to receive free email-alerts related to this article or journal. |
| **Reprints and Subscriptions** | To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org. |
| **Permissions** | To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org. |

*Research Article*

# Estrogen Metabolism and Exposure in a Genotypic–Phenotypic Model for Breast Cancer Risk Prediction

Philip S. Crooke[1], Christina Justenhoven[5], Hiltrud Brauch[5], and for the GENICA Consortium;
Sheila Dawling[2], Nady Roodi[2], Kathryn S. P. Higginbotham[3], W. Dale Plummer[4], Peggy A. Schuyler[4],
Melinda E. Sanders[2], David L. Page[2], Jeffrey R. Smith[3], William D. Dupont[4], and Fritz F. Parl[2]

## Abstract

**Background:** Current models of breast cancer risk prediction do not directly reflect mammary estrogen metabolism or genetic variability in exposure to carcinogenic estrogen metabolites.

**Methods:** We developed a model that simulates the kinetic effect of genetic variants of the enzymes CYP1A1, CYP1B1, and COMT on the production of the main carcinogenic estrogen metabolite, 4-hydroxy-estradiol ($4\text{-}OHE_2$), expressed as area under the curve metric ($4\text{-}OHE_2\text{-}AUC$). The model also incorporates phenotypic factors (age, body mass index, hormone replacement therapy, oral contraceptives, and family history), which plausibly influence estrogen metabolism and the production of $4\text{-}OHE_2$. We applied the model to two independent, population-based breast cancer case–control groups, the German GENICA study (967 cases, 971 controls) and the Nashville Breast Cohort (NBC; 465 cases, 885 controls).

**Results:** In the GENICA study, premenopausal women at the 90th percentile of $4\text{-}OHE_2\text{-}AUC$ among control subjects had a risk of breast cancer that was 2.30 times that of women at the 10th control $4\text{-}OHE_2\text{-}AUC$ percentile (95% CI: 1.7–3.2, $P = 2.9 \times 10^{-7}$). This relative risk was 1.89 (95% CI: 1.5–2.4, $P = 2.2 \times 10^{-8}$) in postmenopausal women. In the NBC, this relative risk in postmenopausal women was 1.81 (95% CI: 1.3–2.6, $P = 7.6 \times 10^{-4}$), which increased to 1.83 (95% CI: 1.4–2.3, $P = 9.5 \times 10^{-7}$) when a history of proliferative breast disease was included in the model.

**Conclusions:** The model combines genotypic and phenotypic factors involved in carcinogenic estrogen metabolite production and cumulative estrogen exposure to predict breast cancer risk.

**Impact:** The estrogen carcinogenesis–based model has the potential to provide personalized risk estimates. *Cancer Epidemiol Biomarkers Prev; 20(7); 1502–15. ©2011 AACR.*

## Introduction

Abundant experimental and epidemiologic evidence has implicated estrogens as prime risk factors for the development of breast cancer (1–4). Experiments on estrogen metabolism (5–7), formation of DNA adducts (8, 9), mutagenicity (10, 11), cell transformation (12, 13), and carcinogenicity (14, 15) have implicated certain estrogen metabolites, especially the catechol estrogen 4-hydroxyestradiol ($4\text{-}OHE_2$; Fig. 1) to react with DNA via its quinone, causing mutations and initiating cancer. Estrogen-DNA adducts have been detected in normal and malignant human breast tissues (16–18) and we have provided direct experimental evidence that oxidative metabolism of 17β-estradiol ($E_2$) leads to the formation of $4\text{-}OHE_2$ and deoxyribonucleoside adducts (19). Epidemiologic studies have indicated that breast cancer risk is higher in women with early menarche and late menopause, who have longer exposure to estrogens (20). Therefore, current models of breast cancer risk prediction are mainly based on cumulative estrogen exposure and incorporate factors such as current age, age at menarche, and age at first live birth (21–26). While these traditional exposure data are valuable in risk calculation, they do not directly reflect mammary estrogen metabolism. Furthermore, current models do not address genetic variability between women in exposure to carcinogenic estrogen metabolites, including catechols and quinones.

We previously sought to address these limitations by developing a kinetic–genetic model of estrogen exposure

**Authors' Affiliations:** Departments of [1]Mathematics, [2]Pathology, [3]Genetic Medicine, and [4]Biostatistics, Vanderbilt University, Nashville, Tennessee; and [5]Dr Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Auerbachstrasse 112, Stuttgart, and University Tübingen, Germany

**Corresponding Author:** Fritz F. Parl, Department of Pathology, Vanderbilt University, TVC 4918, Nashville, TN 37232. Phone: 615-343-9117; Fax: 1-615-343-9563; E-mail: fritz.parl@vanderbilt.edu
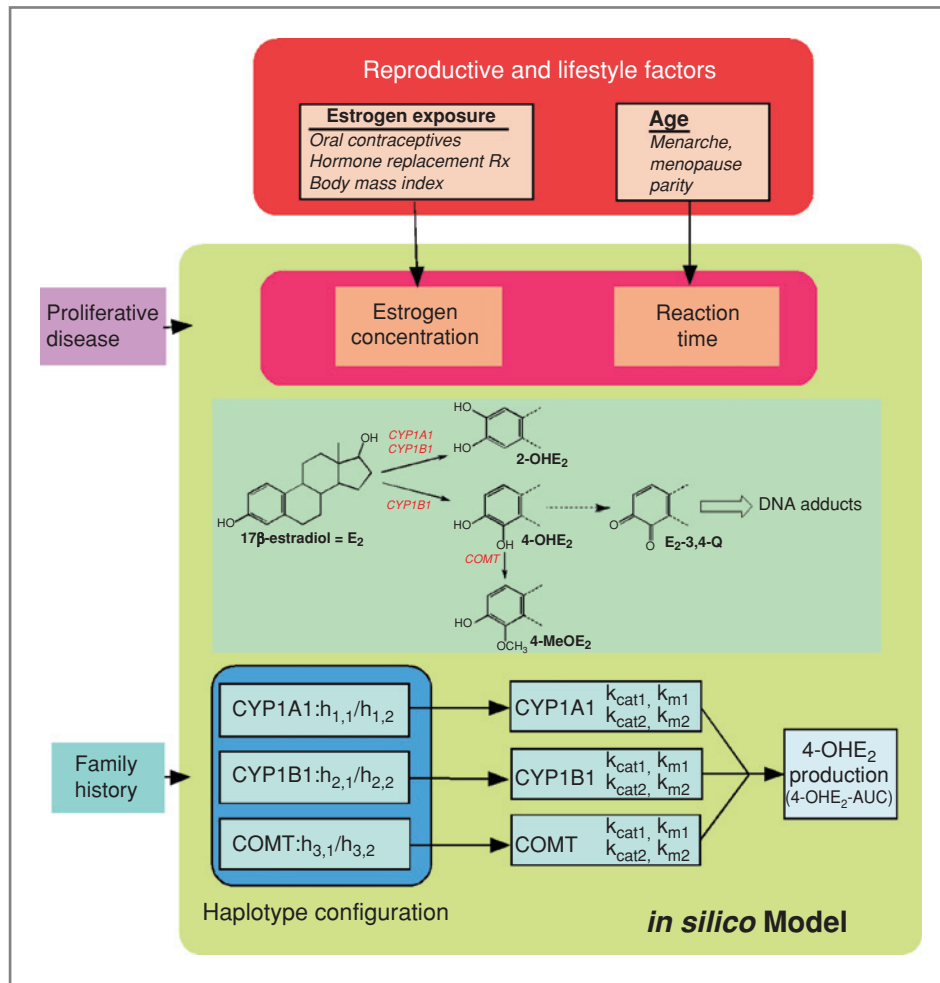
American Association for **Cancer Research**

**Figure 1.** Genotypic–phenotypic model of estrogen metabolism and exposure (yellow box represents genotype component with phenotypic factors outside). In the center is the estrogen metabolism pathway used to generate the *in silico* genotype model. The pathway is initiated by CYP1A1 and CYP1B1, which catalyze the oxidation of $E_2$ to catechol estrogens $2\text{-}OHE_2$ and $4\text{-}OHE_2$. The catechol estrogens are either methylated by COMT to methoxyestrogens (e.g., $4\text{-}MeOE_2$ as main fraction) or further oxidized by the CYPs to quinones, for example, $E_2\text{-}3,4\text{-}Q$, the main quinone that forms depurinating estrogen DNA adducts. Each of the 3 genes CYP1A1, CYP1B1, and COMT is genotyped for all subjects and the SNP genotype data are used to derive the haplotype configuration for each subject. The model then uses the kinetic constants in a system of nonlinear differential equations to calculate the production of the main carcinogenic estrogen metabolite, $4\text{-}OHE_2$, for each haplotype configuration as well as the weighted average of all $4\text{-}OHE_2$ production values, using the probabilities of haplotype configurations as weights. The genotype model is influenced by traditional risk factors, which are thought to affect hormone concentration (OC, HRT, and BMI) and exposure time (age, ages at menarche and menopause, parity), as well as by FH of breast cancer and proliferative disease. The weight of each phenotypic factor on hormone concentration is determined by MLE. Thus, the combined genotypic–phenotypic model allows the calculation of $4\text{-}OHE_2$ produced by each woman, expressed as $4\text{-}OHE_2\text{-}AUC$, for a personalized risk estimate of developing breast cancer.

in relation to breast cancer risk prediction (27). The model incorporates the main components of mammary estrogen metabolism, that is, the parent hormone $E_2$ and the principal enzymes expressed in breast tissue, the phase I cytochrome P450 (CYP) enzymes CYP1A1 and CYP1B1 and the phase II enzyme catechol-*O*-methyltransferaset (COMT; Fig. 1). The oxidative estrogen metabolism pathway begins with the conversion of $E_2$ by CYP1A1 and CYP1B1 into the catechol estrogens $2\text{-}OHE_2$ and $4\text{-}OHE_2$. These same enzymes further oxidize the catechol estrogens to highly reactive estrogen quinones ($E_2\text{-}2,3\text{-}Q$, $E_2$-$3,4\text{-}Q$), which can form Michael addition products with

deoxyribonucleosides. One of the quinones, $E_2\text{-}3,4\text{-}Q$, has been shown to cause depurinating estrogen-DNA adducts and mutations in breast epithelium (9, 28). The genotoxicity of oxidative estrogen metabolism is mitigated by COMT, which catalyzes the methylation of catecholestrogens to methoxyestrogens (e.g., 4-methoxy-estradiol, $4\text{-}MeOE_2$, as main fraction) thereby limiting the catechol estrogens available for conversion to estrogen quinones. Using experimentally determined rate constants for these enzyme reactions (29–32), the model allowed kinetic simulation of the conversion of $E_2$ into the main metabolites, such as $4\text{-}OHE_2$, the precursor for

the mutagenic $E_2$-3,4-Q. The simulations showed excellent agreement with experimental results and provided a quantitative assessment of the metabolic interactions. Using rate constants of genetic variants of CYP1A1, CYP1B1, and COMT, the model further allowed examination of the kinetic impact of enzyme polymorphisms on the entire estrogen metabolic pathway. Furthermore, the model identified those genetic variants in CYP1A1, CYP1B1, and COMT that produce the largest quantities of catechols and quinones. Application of the model to a breast cancer case–control population (221 invasive breast cancer cases, 217 controls) defined the estrogen quinone $E_2$-3,4-Q as a potential breast cancer risk factor. This exploratory analysis identified a subset of women at increased breast cancer risk based on their enzyme isoform and consequent $E_2$-3,4-Q production (27). These results suggest that traditional breast cancer risk prediction may be enhanced by incorporation of inherited differences in estrogen metabolism (33).

Comparison of intra-tissue concentrations of estrogens ($E_1$, $E_2$, $E_3$), hydroxyestrogens (2-OHE$_1$, 2-OHE$_2$, 4-OHE$_1$, 4-OHE$_2$, and 16α-OHE$_1$) and methoxyestrogens (2-MeOE$_1$, 2-MeOE$_2$, 4-MeOE$_1$, and 4-MeOE$_2$) in normal and malignant breast revealed the highest concentration of 4-OHE$_2$ in malignant tissue (34). The concentration (1.6 nmol/g tissue) determined by combined high-performance liquid chromatography and gas chromatography–mass spectrometry was more than twice as high as that of any other compound. Such a high level in neoplastic mammary tissue suggests a mechanistic role of 4-OHE$_2$ in tumor development. This is supported by experimental evidence, which indicates that 4-catechol estrogens are more carcinogenic than the 2-OH isomers. Treatment with 4-OHE$_2$, but not 2-OHE$_2$, induced renal cancer in Syrian hamster (35). Analysis of renal DNA showed that 4-OHE$_2$ significantly increased 8-hydroxyguanosine levels, whereas 2-OHE$_2$ did not cause oxidative DNA damage (36). In addition to the induction of renal cancer in the hamster model, 4-OHE$_2$ is capable of inducing uterine adenocarcinoma, a hormonally related cancer, in mice. Administration of $E_2$, 2-OHE$_2$, and 4-OHE$_2$ induced endometrial carcinomas in 7%, 12%, and 66% of treated CD-1 mice, respectively (14). Examination of microsomal $E_2$ hydroxylation activity in human breast cancer showed significantly higher 4-OHE$_2$/2-OHE$_2$ ratios in tumor tissue than in adjacent normal breast tissue, whereas the latter tissue samples contained 4-fold higher levels of 4-OHE$_2$ than normal tissue from benign breast biopsies (6, 37).

In the present study, we applied our model to 2 independent, population–based case–control studies of genetic breast cancer risk factors to validate and extend prediction of the model. Furthermore, we incorporated traditional risk factors into the model including age, family history (FH) of breast cancer, body mass index (BMI), use of hormone replacement therapy (HRT), and use of oral contraceptives (OC). Thus, we developed a more refined risk prediction model that integrates known

reproductive and lifestyle factors with predicted exposure to oxidative estrogen metabolites as determined by inherited variation in critical genes involved in the estrogen metabolism pathway. This new genotypic–phenotypic model incorporates interactions between common risk factors, each of low penetrance when considered alone, but of greater attributable risk when considered in synergistic combination. The combined genotypic–phenotypic model led to the identification of high-risk women.

## Materials and Methods

### Study populations

The participants of the population-based case–control gene environment interaction and breast cancer (GENICA) study from the Greater Bonn Region, Germany, were recruited between August 2000 and September 2004 as described previously (38, 39). In brief, there are 1,143 incident breast cancer cases and 1,155 population controls matched in 5-year age classes. Cases and controls were eligible if they were of Caucasian ethnicity, current residents of the study region, and below 80 years of age. Information on known and proposed risk factors was collected via in person interviews. The response rate for cases was 88% and for controls 67%. Characteristics of the study population regarding potential breast cancer risk factors include age at diagnosis (< 50, ≥ 50 years), menopausal status (premenopausal, postmenopausal), breast cancer in mothers and sisters (yes, no), OC (never, 0 < OC ≤ 5, 5 < OC ≤ 10, OC > 10 years), HRT use (never, 0 < HRT ≤ 10, HRT > 10 years), and BMI (BMI < 20, 20 ≤ BMI < 25, 25 ≤ BMI < 30, BMI ≥ 30). The GENICA study was approved by the Ethics Committee of the University of Bonn. All study participants gave written informed consent. Genomic DNA was extracted from heparinized blood samples (Puregene, Gentra Systems, Inc.). DNA samples were available for 1,021 cases (89%) and 1,015 controls (88%). Genotyping was conducted by matrix-assisted laser desorption/ionization time-of-flight spectrometry (MALDI-TOF MS) and PCR-based fragment length polymorphism genotyping as previously described (40). Phenotypic and genotypic factors examined in the GENICA study are summarized in Table 1. The GENICA data were used as training set for the model.

The Nashville Breast Cohort (NBC) is an ongoing retrospective cohort study of 16,946 women who underwent a breast biopsy revealing benign parenchyma or fibroadenoma at Vanderbilt, St. Thomas, and Baptist Hospitals in Nashville, Tennessee since 1954 (41, 42). Subjects provided written informed consent under approved institutional review board protocols. To be eligible for inclusion in this cohort, a woman could not have had a diagnosis of breast cancer prior to her entry biopsy. Additional details on the NBC are given elsewhere (41, 42). Subjects were followed by telephone

**Table 1.** Phenotypic and genotypic factors examined in GENICA and NBC

| | GENICA | | NBC | |
|---|---|---|---|---|
| | Available | Categories | Available | Categories |
| *Phenotypic factor* | | | | |
| Age | Yes | continuous | yes | Continuous |
| Menopausal status | Yes | 2 | yes | 2 |
| Menarche age | Yes | continuous | yes | Continuous |
| Menopausal age | Yes | continuous | yes | Continuous |
| Parity | Yes | continuous | yes | Continuous |
| Age at first birth | Yes | continuous | yes | Continuous |
| Primary family history | Yes | 2 | yes | 2 |
| Body mass index | Yes | 4 | yes | 4 |
| Hormone replacement therapy | Yes | 3 | yes | 5 |
| Oral contraceptives | Yes | 4 | yes | |
| Proliferative disease | No | | yes | 3 |
| *Genotypic factor* | | | | |
| CYP1A1 | | | | |
| rs1799814 | C/A | | Thr461Asn | |
| rs1048943 | A/G | | Ile462Val | |
| CYP1B1 | | | | |
| rs1056836 | G/C | | Val432Leu | |
| rs1800440 | A/G | | Asn453Ser | |
| COMT | | | | |
| rs4680 | G/A | | Val108Met | |

interviews or, if deceased, with their next of kin, through medical record reviews, and through searches of the National Death Index and the Tennessee Cancer Registry. Successful follow-up has been obtained on 90% of the women who met the entry criteria for the NBC and who were biopsied before 1990. There were 8,897 Caucasian women among this group whose entry biopsy formalin-fixed, paraffin-embedded (FFPE) tissue blocks were available and who were eligible for this study, of which 575 had developed breast cancer on follow-up. We conducted a nested case–control study of women from this subcohort who were postmenopausal at exit. We selected 2 controls per case from the risk set of those who had not been diagnosed with breast cancer by the follow-up time when their case developed this disease. These controls were selected without replacement. Controls were matched to cases by age and year of entry biopsy. Successful DNA extractions from benign archival entry biopsy specimens were conducted for 465 postmenopausal Caucasian cases and for 885 of their matched controls. The proportion of subjects with successful DNA extractions was 96%. We employed the Illumina GoldenGate assay (Illumina) for genotyping, using 5 μL ($\geq$ 250 ng) of each extracted archival DNA. Each 96-well plate of DNA genotyped contained an average of 4.2 (range 1–6) reference saliva DNA samples of study subjects for whom DNA from matching blocks was under evaluation. This enabled assessment of genotype accuracy. The concor-

dance rate for subject saliva DNA–FFPE DNA pairs was 99.95% to date, across 227,819 duplicate genotype pairs (43). Drs. Page and Sanders conducted the histologic review of patients' entry biopsy slides using criteria of the Cancer Committee of the College of American Pathologists, without knowledge of subsequent cancer outcome (44–47).

## Mathematical model for estrogen metabolism pathway

For the purpose of the present study, we focused on the principal reactions of the estrogen metabolism pathway shown in Figure 1, starting with the oxidation of $E_2$ to the catechol estrogens 2-OHE$_2$ and 4-OHE$_2$ by CYP1A1 and CYP1B1. The catechol estrogens are either methylated by COMT to methoxyestrogens (e.g., 4-MeOE$_2$) or further oxidized by the CYPs to quinones, for example, $E_2$-3,4-Q, the main quinone that forms depurinating estrogen DNA adducts. Because of its reactivity, $E_2$-3,4-Q is short-lived and cannot be readily measured. Instead we chose its immediate precursor, 4-OHE$_2$, because it is reliably quantified and known to be carcinogenic (9). The reactions in the pathway are modeled by a system of nonlinear ordinary differential equations. We assume that each reaction in the pathway,

$$A + E \underset{k_2}{\overset{k_1}{\rightleftarrows}} C \overset{k_3}{\longrightarrow} B + E,$$
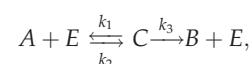
is an enzymatic reaction from a reactant $A$ to a product $B$ with $E$ being the enzyme and $C$ the $AE$ complex. Because the individual reaction rates, $k_1$, $k_2$, $k_3$, are often difficult to measure, we use the quasi-steady–state assumption:

$$\frac{db}{dt} = \frac{k_{cat}ea}{K_m + a}.$$

where $a$, $b$, and $e$ are the concentrations of $A$, $B$, and $E$, respectively, and $k_{cat}$ and $K_m$ are kinetic constants that depend on $k_1, k_2$, and $k_3$. Using these assumptions for each reaction depicted in Figure 1, we have the following system of nonlinear differential equations for the components in the pathway.

$$\frac{d(E_2)}{dt} = -\frac{k_{cat_1}E_{CYP1B1}E_2}{K_{m_1} + E_2} - \frac{k_{cat_2}E_{CYP1A1}E_2}{K_{m_2} + E_2}$$
$$- \frac{k_{cat_3}E_{CYP1B1}E_2}{K_{m_3} + E_2}, E_2(0) = E_{2_0}^{effective}$$

$$\frac{d(2\text{-}OHE_2)}{dt} = \frac{k_{cat_2}E_{CYP1A1}E_2}{K_{m_2} + E_2} + \frac{k_{cat_3}E_{CYP1B1}E_2}{K_{m_3} + E_2},$$
$$2\text{-}OHE_2(0) = 0 \qquad (1)$$

$$\frac{d(4\text{-}OHE_2)}{dt} = \frac{k_{cat_1}E_{CYP1B1}E_2}{K_{m_1} + E_2}$$
$$- \frac{k_{cat_5}E_{COMT}4\text{-}OHE_2}{K_{m_5} + 4\text{-}OHE_2}, 4\text{-}OHE_2(0) = 0$$

The Michaelis–Menten parameters, $k_{cat}$ and $K_m$, in the model have been experimentally measured for each of the 3 enzymes, CYP1A1, CYP1B1, and COMT [(29–32) and additional unpublished results]. The value of $E_{2_0}^{effective}$ is estimated for each patient as described below. As described in our earlier publication, there are 2 common nonsynonymous polymorphisms in CYP1A1 (codons 461 and 462), 4 in CYP1B1 (codons 48, 119, 432, and 453), and 1 in COMT (codon 108), giving rise to 4 alleles of CYP1A1, 16 of CYP1B1, and 2 of COMT, respectively (27). The corresponding $k_{cat}$ and $K_m$ for wild-type and variant enzymes for the pathway reactions are summarized in Table 2. In the model, we combined the genotypic and kinetic information to determine the amount of 4-OHE$_2$ for each woman as summarized in Figure 1.

We do not have complete information about all these SNPs for each woman in the 2 data sets under consideration in this paper because CYP1B1 codons 48 and 119 were not assessed. To circumvent this limitation, we extracted values of the kinetic parameters in Table 2 by using a nonlinear averaging procedure. To understand the averaging algorithm, consider the reaction: $E_2 \xrightarrow{CYP1B1} 4\text{-}OHE_2$ with 432Val-453Asn. The rate of formation of 4-OHE$_2$ is approximated by using the kinetic parameters for each of the 4 CYP1B1 haplotypes (48Arg-119Ala-432Val-453Asn, 48Gly-119Ala-432Val-453Asn, 48Arg-119Ser-432Val-453Asn, and 48Gly-119Ser-432Val-453Asn) coupled with its frequency within the population. Thus, reaction kinetics of the rate of 4-OHE$_2$ production from E$_2$ for a woman with the 432Val-453Asn

genetic characterization is approximated by the differential equation:

$$\frac{d(4\text{-}OHE_2)}{dt} =$$
$$\frac{p_1 \frac{1.17E_{CYP1B1}E_2}{23 + E_2} + p_2 \frac{6.00E_{CYP1B1}E_2}{19 + E_2} + p_3 \frac{3.80E_{CYP1B1}E_2}{10 + E_2} + p_4 \frac{1.78E_{CYP1B1}E_2}{32 + E_2}}{\sum_{i=1}^{4} p_i} \qquad (2)$$

where $p_i$, $i = 1, 2, 3, 4$, are the relative frequencies of haplotypes in the population. With this method of assigning the kinetic parameters for each woman, we can numerically calculate the time trajectories of each component of the metabolism pathway.

Another component of the model is the reaction time ($T_{reaction}$) over which the cumulative 4-OHE2 exposure occurs. The model assumes that $T_{reaction}$ is a function of the age at menarche ($A_{menarche}$), age at menopause ($A_{menopause}$), and number of children (parity, $P$). To capture these factors, we defined the effective time: $T_{reaction} = M_1 T$ where

$$M_1 = \frac{A_{menopause} - (A_{menarche} + P)}{A_{menopause}}$$
$$= 1 - \frac{A_{menarche} + P}{A_{menopause}} \le 1$$

and $T = 120$. The form for $M_1$ was chosen to reflect the influence of ages of menarche and menopause and the number of children on the exposure time to higher levels of estrogen. In particular, as $A_{menarche}$ increases, $A_{menopause}$ decreases and/or $P$ increases, the exposure time to high levels of estrogen during premenopause decreases, leading to smaller 4-OHE$_2$-AUC values. The calculation of the average AUC value was then conducted as the definite integral:

$$4\text{-}OHE_2\text{-}AUC = \frac{1}{T_{reaction}} \int_0^{T_{reaction}} 4\text{-}OHE_2(t)dt.$$

In the case of the premenopausal GENICA data, we set $T_{reaction} = T = 120$ min because these women do not have an established age of menopause.

### Effect of phenotypic risk factors on model

As depicted in Figure 1, we incorporated traditional risk factors into the model by considering their effects on 2 key components: (i) the initial estrogen level $\left(E_2(0) = E_{2_0}^{effective}\right)$ and (ii) the reaction time over which the 4-OHE$_2$-AUC is calculated. The model uses both categorical (e.g., FH) and quantitative (e.g., age of menarche) inputs to calculate the AUC for each individual. For example, the initial estrogen level is affected by BMI, intake of OC or HRT, and FH. Each of these covariates has several categories. In the GENICA study of postmenopausal women these categories for BMI were: (i) BMI < 20, (ii) $20 \le$ BMI < 25, (iii) $25 \le$ BMI < 30, and (iv) BMI $\ge$ 30; for HRT: (i) never,

**Table 2.** Kinetic parameters for wild-type and variants of CYP1A1, CYP1B1, and COMT

| Reaction | Enzyme | Allele | $k_{cat}$ | $K_m$ |
|---|---|---|---|---|
| $E_2 \rightarrow 2\text{-}OHE_2$ | CYP1A1 | 461Thr-462Ile (wt*) | 1.10 | 17 |
| | | 461Asn-462Ile | 0.80 | 23 |
| | | 461Thr-462Val | 2.70 | 18 |
| | | 461Asn-462Val | 1.30 | 23 |
| $E_2 \rightarrow 2\text{-}OHE_2$ | CYP1B1 | 48Arg-119Ala-432Val-453Asn (wt) | 1.30 | 45 |
| | | 48Gly-119Ala-432Val-453Asn | 3.20 | 29 |
| | | 48Arg-119Ser-432Val-453Asn | 2.30 | 18 |
| | | 48Arg-119Ala-432Leu-453Asn | 2.30 | 73 |
| | | 48Arg-119Ala-432Val-453Ser | 2.80 | 39 |
| | | 48Gly-119Ser-432Leu-453Ser | 2.50 | 29 |
| | | 48Arg-119Ala-432Leu-453Ser | 3.90 | 136 |
| | | 48Gly-119Ser-432Leu-453Asn | 2.00 | 57 |
| | | 48Gly-119Ser-432Val-453Asn | 2.00 | 64 |
| | | 48Gly-119Ser-432Val-453Ser | 2.10 | 67 |
| | | 48Arg-119Ser-432Leu-453Asn | 0.81 | 51 |
| | | 48Arg-119Ser-432Leu-453Ser | 0.47 | 24 |
| | | 48Gly-119Ala-432Leu-453Ser | 0.47 | 31 |
| | | 48Gly-119Ala-432Leu-453Asn | 0.19 | 9.1 |
| | | 48Arg-119Ser-432Val-453Ser | 0.65 | 36 |
| | | 48Gly-119Ala-432Val-453Ser | 0.23 | 75 |
| $E_2 \rightarrow 4\text{-}OHE_2$ | CYP1B1 | 48Arg-119Ala-432Val-453Asn (wt) | 1.17 | 23 |
| | | 48Gly-119Ala-432Val-453Asn | 6.00 | 19 |
| | | 48Arg-119Ser-432Val-453Asn | 3.80 | 10 |
| | | 48Arg-119Ala-432Leu-453Asn | 1.85 | 37 |
| | | 48Arg-119Ala-432Val-453Ser | 4.50 | 17 |
| | | 48Gly-119Ser-432Leu-453Ser | 4.40 | 15 |
| | | 48Arg-119Ala-432Leu-453Ser | 3.31 | 55 |
| | | 48Gly-119Ser-432Leu-453Asn | 1.77 | 30 |
| | | 48Gly-119Ser-432Val-453Asn | 1.78 | 32 |
| | | 48Gly-119Ser-432Val-453Ser | 2.05 | 33 |
| | | 48Arg-119Ser-432Leu-453Asn | 0.70 | 27 |
| | | 48Arg-119Ser-432Leu-453Ser | 1.90 | 13 |
| | | 48Gly-119Ala-432Leu-453Ser | 1.80 | 12 |
| | | 48Gly-119Ala-432Leu-453Asn | 0.73 | 7.2 |
| | | 48Arg-119Ser-432Val-453Ser | 2.70 | 17 |
| | | 48Gly-119Ala-432Val-453Ser | 0.81 | 28 |
| $4\text{-}OHE2 \rightarrow 4\text{-}MeOE2$ | COMT | 108Val (wt) | 3.40 | 24 |
| | | 108Met | 2.04 | 24 |

(ii) $0 \leq 10$ years, and (iii) $> 10$ years; and for FH, the presence or absence of a first degree FH of breast cancer. The cumulative effect of more than 1 phenotypic risk factor was assumed to be multiplicative. For example, in the case of postmenopausal women, if the multiplier for BMI is $M_{BMI}$, the multiplier for HRT is $M_{HRT}$, and the multiplier for FH is $M_{FH}$, then the effective initial estrogen level for these 3 factors is: $E_{2_0}^{effective} = M_{BMI}M_{HRT}M_{FH}E_{2_0}$. The value of $E_{2_0}$ was chosen to be 10, which is consistent with the initial estradiol concentration that was used in the experiments to measure the kinetic constants ($k_{cat}$ and $K_m$) in the estrogen metabolism pathway. The multipliers for these 3 cate-

gories were determined by making them parameters in a logistic regression model. Each multiplier has $k$ possible values (which we call weights). The number of weights for each categorical variable is 1 less than the number of categories for the variable. Hence, the total number of weights for the 3 categorical variables (BMI, HRT, and FH) is $3 + 2 + 1 = 7$. The dependent variable in this model was the patient's case–control status, whereas the model's linear predictor was $\beta_0 + \beta_1 f(\mathbf{x}_i; \mathbf{K}; \alpha_1, \alpha_2, ..., \alpha_7)$, where $\mathbf{x}_i$ denotes the $i$th patient's genotype, age at either cancer diagnosis or selection as a control, age of menarche, age of menopause, reproductive history, BMI, HRT and FH; $\mathbf{K}$ denotes a vector of enzyme kinetic

constants; $\beta_0, \beta_1, \alpha_1, \alpha_2, ..., \alpha_7$ are model and regression parameters; and $f(\mathbf{x}_i; \mathbf{K}; \alpha_1, \alpha_2, ..., \alpha_7)$ is the 4-OHE$_2$-AUC derived from our estrogen metabolism pathway model. The 7 parameters $\alpha_1, \alpha_2, ..., \alpha_7$ are associated with the categories for BMI, HRT and FH. Maximum likelihood estimates (MLE) of these parameters are obtained from our logistic regression model. For any set of arguments, the value of the 4-OHE$_2$-AUC function, $f$, is obtained by solving the nonlinear system of differential equations (1) for our estrogen metabolism pathway model. This complicates the derivation of the MLEs of the model parameters because the gradient of the likelihood surface cannot be written in closed form. To derive these estimates we used a hybrid search method involving the Neider–Mead method, differential evolution, and simulated annealing (48–50) as implemented by *Mathematica* (Version 7.0.1.0, Wolfram Research, Inc.). MLE optimization did not include the experimentally determined enzyme kinetic constants $\mathbf{K}$ nor age of menarche, age of menopause, and effects of parity, which enter the model through the definition of $T_{reaction}$.

For the premenopausal GENICA data set, a similar model was used except that OC usage (categorized as never, $0 < OC \leq 5$, $5 < OC \leq 10$, $OC > 10$ years) was used in place of HRT. Phenotypic weights and genotypic factors for postmenopausal women in the NBC were the same as those used in the GENICA postmenopausal study, except that the parameters for HRT were reestimated from the NBC data. We did this because the NBC data has different categories of HRT usage: ($0 \leq HRT < 0.125$, $0.125 \leq HRT < 1$, $1 \leq HRT < 5$, $5 \leq HRT < 10$, $HRT \geq 10$ years). An additional model was tested for postmenopausal NBC women in which their histories of benign breast disease (no proliferative disease, proliferative disease without atypia, and atypical hyperplasia) were included.

## Statistical methods

In the GENICA study, cases and controls were frequency matched by age. In the NBC, cases were individually matched to controls based on their age at entry biopsy and year of entry biopsy (43). For this reason, we analyzed these data using regular logistic regression for the GENICA data and conditional logistic regression for the NBC. To adjust for residual confounding, these analyses were adjusted for age in the GENICA study and age at entry biopsy and year of entry biopsy in the NBC. Restricted cubic splines were used to model the relationship between the log-odds of breast cancer and the 4-OHE$_2$-AUC values adjusted for these residual confounding variables. These analyses found no evidence of a nonlinear relationship between these 2 variables. For this reason 4-OHE$_2$-AUC was entered directly into our regression models without higher-order spline covariates. These models were used in Figures 2 and 3 to plot the
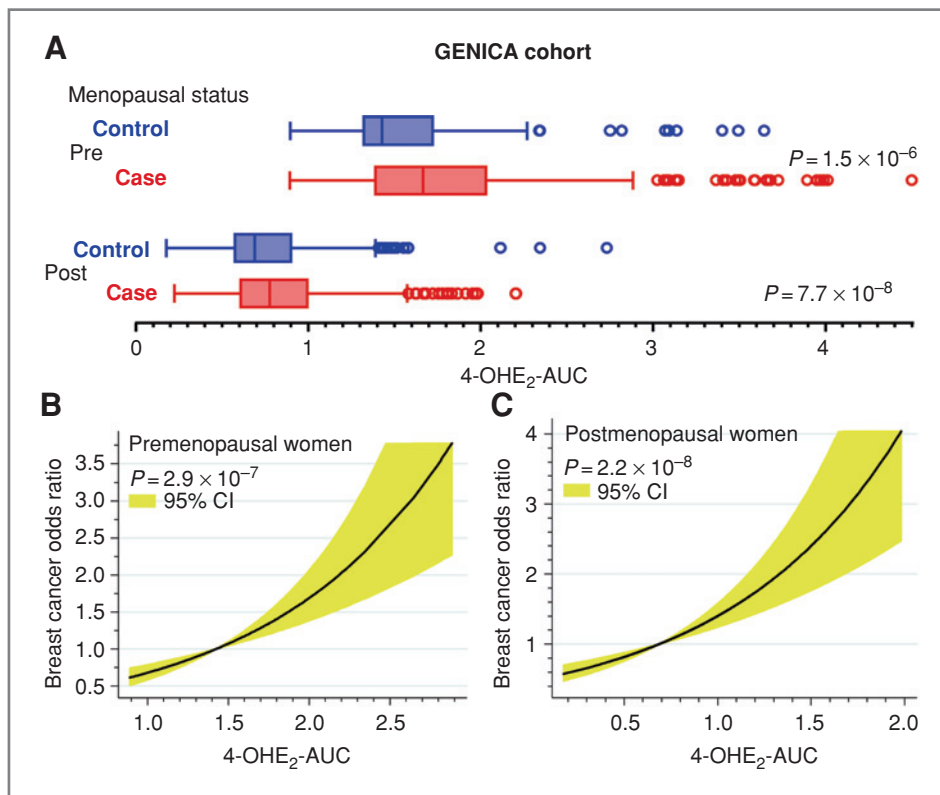


**Figure 2.** Application of model to GENICA case–control study of 218 premenopausal cases, 213 premenopausal controls, 749 postmenopausal cases, and 758 postmenopausal controls. A, box-and-whisker plots of 4-OHE$_2$–AUC values for cases (red) and controls (blue). The boxes span the 4-OHE$_2$–AUC values from the 25th to 75th percentiles, the vertical line within each box indicates the median value. Whiskers extend to the extreme values or to 1.5 times the interquartile range. Outliers beyond the whiskers are plotted individually. B, graph of breast cancer ORs as a function of 4-OHE$_2$–AUC values in premenopausal women. The denominator of these ORs are the breast cancer odds associated with a woman whose 4-OHE$_2$-AUC value equals the median value for a premenopausal control. C, the analogous OR curve for postmenopausal women.

**Figure 3.** Application of model to postmenopausal women from the Nashville Cohort study of 465 cases and 885 controls. A, box-and-whisker plot 4-OHE$_2$-AUC values for cases and controls (see Figure 2 for additional explanation). B and D, graphs of breast cancer ORs as a function of 4-OHE$_2$-AUC values in postmenopausal women. The denominator of these ORs is the breast cancer odds associated with a woman whose 4-OHE$_2$-AUC value equals the median value for a postmenopausal control. Graphs B and C use 4-OHE$_2$-AUC models that exclude and include, respectively, each patient's history of benign proliferative breast disease.
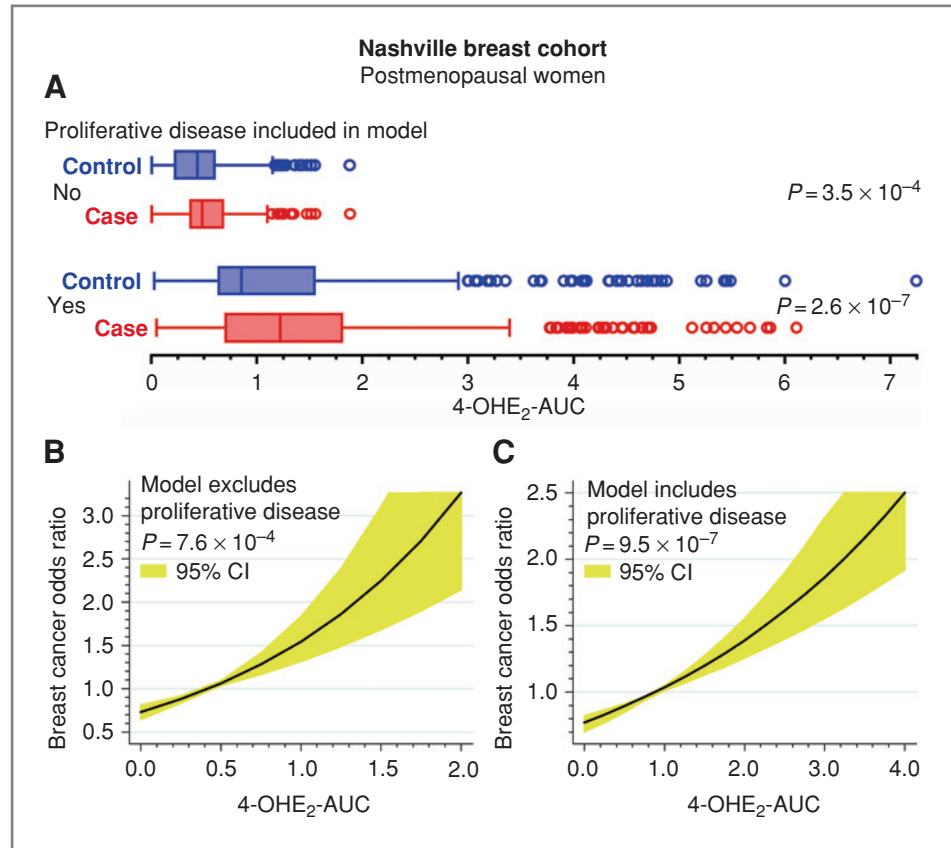
adjusted OR for developing breast cancer as a function of 4-OHE$_2$-AUC. The denominator of these OR curves were women with 4-OHE$_2$-AUC values equal to the median 4-OHE$_2$-AUC value among controls.

In the GENICA study there were no missing values for any of the covariates used to derive the 4-OHE$_2$-AUC. In the NBC, multiple imputation was used to adjust for missing covariates (51–53). The imputed calculations used 1,354 women (465 cases, 885 controls) and complete data was available on 345 women (111 cases, 234 controls). Bootstrapping was used to estimate the CIs for the OR curves in Figure 3.

In the GENICA cohort for both premenopausal and postmenopausal patients, the distributions of 4-OHE$_2$-AUC values in cases and controls were compared using the Wilcoxon rank-sum test. In the NBC, these distributions were compared using a multiply imputed $t$ test (53).

The areas under receiver operating curves (ROC) obtained from our models were derived using the trapezoidal rule. Ninety-five percent CIs for these areas were estimated using the method of DeLong and colleagues (54). All analyses were conducted using Stata version 11 (55).

Conventional and conditional logistic regressions were used to assess the effects of the SNPs in our genotypic–phenotypic model on breast cancer ORs of women in the GENICA and NBC studies, respectively. These risks are adjusted for residual confounding in the same way as in the analyses that produced Figures 2 and 3. An additive model was used with 1 parameter per SNP. The OR for each SNP is adjusted for the other SNPs in the model and represents the breast cancer OR for women with a heterozygous genotype relative to women with a homozygous wild-type genotype. In this additive model, this OR also equals that of women with a homozygous variant genotype relative to women who are heterozygous.

## Results

The model was applied to the GENICA study and results are shown in Figure 2 for both premenopausal and postmenopausal women. 4-OHE$_2$-AUC values in cases were significantly higher than in controls for both pre- and postmenopausal women (Figure 2A). Although the median curve areas were higher for cases than controls in both pre- and postmenopausal subjects, more marked differences were observed at higher percentiles. For example, the 75th percentiles of 4-OHE$_2$-AUC values in premenopausal cases and controls were 2.03 and 1.73, respectively, whereas for postmenopausal women they were 1.00 and 0.901, respectively. Figure 2B and C show how breast cancer ORs increase with increasing 4-OHE$_2$-AUC values in pre- and postmenopausal

women, respectively. These graphs are adjusted for age and are derived with respect to women with 4-OHE$_2$-AUC values equal to the median value for control subjects (1.32 and 0.575 for pre- and postmenopausal women, respectively). Premenopausal women at the 90th percentile of 4-OHE$_2$-AUC among control subjects had a risk of breast cancer that was 2.30 times that of women at the 10th control 4-OHE$_2$-AUC percentile (95% CI: 1.7–3.2, $P = 2.9 \times 10^{-7}$). This relative risk was 1.89 (95% CI 1.5–2.4, $P = 2.2 \times 10^{-8}$) in postmenopausal women.

Parameter estimates obtained from the GENICA postmenopausal model were used to test this model on postmenopausal women from the NBC. (There were insufficient premenopausal women in the NBC to separately assess the effects of 4-OHE$_2$-AUC in these patients.) Models were run that either included or excluded a history of benign proliferative disease. Figure 3 shows the results of these analyses. The top boxplots in Figure 3A and the plot in Figure 3B were derived from a model that excluded a history of benign proliferative disease. In the bottom boxplots of Figure 3A and in the plot in Figure 3C, this history was included in the model. In the NBC, this relative risk in postmenopausal women was 1.81 (95% CI: 1.3–2.6, $P = 7.6 \times 10^{-4}$), which increased to 1.83 (95% CI: 1.4–2.3, $P = 9.5 \times 10^{-7}$) when a history of proliferative breast disease was included in the model. Table 3 shows the ORs associated with the 5 genotypic components of our model in pre- and postmenopausal women from the GENICA study and in postmenopausal women from the NBC. None of these ORs were significantly different from one.

## Discussion

We present a genotypic–phenotypic model for breast cancer risk prediction, which incorporates the main components of mammary estrogen metabolism, enzyme variants, and traditional risk factors related to estrogen exposure. In contrast to the relatively small number of functional studies of estrogen metabolism, multiple epidemiologic studies have investigated breast cancer risk in relation to genetic variation in the critical enzymes involved in estrogen metabolism with inconsistent findings (56, 57). Such studies were limited by their ability to consider only 1 or 2 of the enzymes in the estrogen metabolic pathway. Even those studies that examined all of the component enzymes were not able to assess underlying metabolic interactions in the pathway (38, 58–60). The drawback of any purely genetic assessment is the lack of consideration about functional interactions inherent in complex metabolic pathways such as the estrogen metabolism pathway. A pathway-based functional and quantitative approach is necessary to overcome the current limitation in genotype assessment (61). Our original estrogen metabolism pathway-based model (27) not only combined kinetic and genetic data, but also provided the opportunity to incorporate traditional risk factors tied to estrogen exposure. In attempting to answer the important question how best to incorporate these risk factors into our kinetic–genetic model, we were guided by biological reasoning, experimental data, and epidemiologic findings. We chose the 2 principal components of the model, namely estrogen

**Table 3.** Effect of individual genotypic factors on breast cancer ORs in the GENICA and NBC populations

| Gene | SNP | OR | 95% CI | P |
|---|---|---|---|---|
| *GENICA* | | | | |
| Premenopausal women | | | | |
| CYP1A1 | rs1799814 | 1.20 | (0.61–2.37) | 0.61 |
| | rs1048943 | 0.52 | (0.24–1.13) | 0.10 |
| CYP1B1 | rs1056836 | 1.17 | (0.88–1.56) | 0.27 |
| | rs1800440 | 1.36 | (0.93–1.99) | 0.12 |
| COMT | rs4680 | 0.97 | (0.75–1.26) | 0.82 |
| Postmenopausal women | | | | |
| CYP1A1 | rs1799814 | 0.75 | (0.53–1.07) | 0.11 |
| | rs1048943 | 0.89 | (0.61–1.31) | 0.57 |
| CYP1B1 | rs1056836 | 1.02 | (0.87–1.20) | 0.77 |
| | rs1800440 | 1.07 | (0.87–1.32) | 0.50 |
| COMT | rs4680 | 0.93 | (0.81–1.07) | 0.32 |
| *NBC* | | | | |
| Postmenopausal women | | | | |
| CYP1A1 | rs1799814 | 1.16 | (0.82–1.64) | 0.41 |
| | rs1048943 | 1.10 | (0.75–1.63) | 0.62 |
| CYP1B1 | rs1056836 | 1.08 | (0.90–1.29) | 0.40 |
| | rs1800440 | 1.01 | (0.79–1.28) | 0.97 |
| COMT | rs4680 | 0.98 | (0.83–1.16) | 0.86 |

level and reaction time, to connect to the traditional risk factors (Figure 1).

It is obvious that the estrogen level is increased in women receiving exogenous estrogens in form of OC or HRT. Moreover, there is general agreement that the risk associated with OC and HRT depends on the duration of exposure, being lowest in women who never used OC or HRT (62). In the United States, the most commonly prescribed HRT is Premarin, a complex mixture of estrogens, in particular the equine estrogens equilin and equilenin, which differ structurally from $E_2$ and $E_1$ by having an unsaturated B ring. The amount of human estrogens is much lower, for example, $E_2$ accounts for only 1.5% of estrogens present in Premarin (63). Despite the structural difference, equine estrogens are metabolized by CYP1B1 and CYP1A1 to the catechol 4-OH-equilenin, which contains aromatic A and B rings. Like 4-OHE$_2$, 4-OH-equilenin is further metabolized to its quinone, and cell culture experiments showed that 4-OH-equilenin via its quinone induced DNA damage in breast cancer cell lines and cellular transformation *in vitro* (64, 65). Thus, all estrogens including equine estrogens used in HRT are metabolized via the same CYP-mediated oxidative pathway to generate catechols and quinones, which, in turn, cause DNA damage. However, equine estrogens appear to be metabolized less efficiently than human estrogens, which may explain why Premarin seems to have a weaker effect on risk of breast cancer than endogenous $E_2$. HRT and OC were documented in both GENICA and NBC, although specific issues, such as timing of exposure (e.g., age at first use, time since first use, time since last use) were not recorded and therefore could not be addressed in our model. In general, it was our intent in designing the model to capture each risk factor without attempting to specify every possible subgroup.

Besides input from exogenous OC and HRT, a variety of other factors influence the estrogen concentration, especially body weight and exercise. The Endogenous Hormones and Breast Cancer Collaborative Group concluded that the increase in breast cancer risk with increasing BMI among postmenopausal women was largely the result of the associated increase in estrogens (66). Because of the importance of body weight and obesity, we included BMI as an integral component into the model utilizing data available in the GENICA and NBC groups. Exercise has also a well-known effect on estrogen concentration and breast cancer risk, especially in postmenopausal women (67). We did not include exercise in the model because neither GENICA nor NBC had collected exercise data. If such data were available in another study population, we could readily integrate exercise as a phenotypic factor into a future model via its effect on estrogen concentration.

Family history of breast cancer is associated with 10% to 20% of breast cancer cases and within that group approximately one half (5%–10% of all cases) are strongly hereditary, for example, linked to germline mutations in genes such as *BRCA1* and *BRCA2* (68). It has been recognized that *BRCA1* and *BRCA2* mutations exhibit variable penetrance, which is likely accounted for by other susceptibility genes among carriers (69). Thus, FH results from the combined input of high- and low-penetrance genes. There were no known patients with *BRCA1* or *BRCA2* mutations in either GENICA or NBC. To reflect FH, we used a weighting factor, $M_{FH}$, to optimize separation of cases and controls.

Benign breast disease encompasses a spectrum of histologic entities, usually subdivided into nonproliferative lesions, proliferative lesions without atypia, and atypical hyperplasia (41, 70). Analysis of the original NBC showed that the latter 2 types of lesions have clinically significant premalignant potential (41). In a more recent study of 9,087 women followed for a median of 15 years, the relative risk of breast cancer associated with proliferative changes without atypia was 1.88 (95% CI: 1.66–2.12) and increased for atypical hyperplasia to 4.24 (95% CI: 3.26–5.41; ref. 70). As expected, the inclusion of proliferative disease as a risk factor in our model improved risk prediction for the NBC and the model showed a progressive risk increase for proliferative disease without atypia and atypical hyperplasia compared with the absence of proliferative lesions.

Figure 3B and C should be compared with Figure 2C. The OR curves from the NBC in Figure 3 were derived using the weights for BMI and FH derived from the GENICA data set. There are 9 parameters in the GENICA model that are being fit to the data and there are over 200 premenopausal cases and controls and over 700 postmenopausal cases and controls. This gives us over 20 premenopausal and over 70 postmenopausal cases and controls per parameter. Typical rules of thumb are that you should have 20 or more cases and controls per parameter to avoid overfitting (71). Hence, model overfitting should not be serious concern, particularly for the postmenopausal women. In the postmenopausal GENICA women, the breast cancer odds associated with women at the 90th control 4-OHE$_2$-AUC percentile was 1.89 times that of women at the 10th control 4-OHE$_2$-AUC percentile. This OR was reduced to 1.81 (a 4% reduction) in postmenopausal NBC women using the model that excluded proliferative disease. Hence, the test set analysis of the NBC women provides considerable validation of the GENICA model for postmenopausal women. Adding a history of proliferative disease to the 4-OHE$_2$-AUC model (Figure 3C) changes the range of 4-OHE$_2$-AUC values and increases the level of statistical significance but does not greatly affect the ORs associated with equivalent percentile values. For example, adding a proliferative disease, history increases the 90th versus 10th 4-OHE$_2$-AUC percentile OR from 1.81 to 1.83. In marked contrast to Figure 2B and C, Figure 3B and C and Table 3 shows no evidence of elevated breast cancer risk associated with the SNPs in our genotypic–phenotypic model. It is thus plausible that the variation in breast cancer risk shown in these figures is because of variation

in the patient's 4-OHE$_2$-AUC rather than to variation in the individual SNPs that are used in this model.

Several models are currently available to predict the risk of breast cancer, of which the Claus and Gail models are used most often (22, 26). The Claus model, which is based on assumptions of the prevalence of high-penetrance genes for susceptibility to breast cancer, is only applicable for women with a FH of breast cancer (23). The Gail model incorporates primary and secondary FH as well as the age at menarche, the age at first live birth, the number of breast biopsies, the presence of atypical hyperplasia in these biopsies, and race (21, 24). Both of these models were developed on the basis of data from much larger study populations than the 2 study groups available to us. The advantage of our genotypic–phenotypic model is the underlying biologic reasoning inherent in a pathway-based model and the integration of endogenous and exogenous risk factors.

In a recent study, Wacholder and colleagues (72) reported that the Gail model achieves an area under the ROC of 0.534. The addition of 7 SNPs associated with breast cancer increased the ROC to 0.586. We used the NBC, which includes information on most of the risk categories of the Gail model, that is, patient age, age at menarche and first birth, number of biopsies, presence of atypical hyperplasia in these biopsies, and FH (21, 24) for a direct comparison of our new model with the Gail model. The area under the ROC curve associated with our 4-OHE$_2$-AUC model that includes proliferative disease was 0.588 (95% CI: 0.56–0.62). This was slightly greater than, but not significantly different from that associated with the Gail model 0.558 (95% CI: 0.53–0.59). Hence, while these models can identify women at increased breast cancer risk, none of them are particularly effective at predicting who will develop breast cancer.

A shortcoming of our current model is the omission of functional SNPs outside the coding region and the inclusion of only 3 genes, albeit of primary importance for mammary estrogen metabolism. Another important gene, CYP19A1, encodes aromatase, the main enzyme producing E$_2$ and E$_1$ from androgen precursors. Haplotype-tagging SNPs and common haplotypes spanning the coding and proximal 5′ region of CYP19A1 were shown to be significantly associated with a 10% to 20% increase in endogenous estrogen levels in postmenopausal women (73). The future addition of CYP19A1 in form of haplotype-tagging SNPs would extend the range of our model by including information about the input E$_2$ concentration to be converted by CYP1A1, CYP1B1, and COMT to carcinogenic metabolites. Among the phase II conjugating enzymes, COMT is the sole methylating enzyme, whereas there are potentially 3 glutathione (GSH)-conjugating enzymes, GSTA1, GSTM1, and GSTP1. COMT catalyzes the methylation of catechol estrogens to methoxy estrogens, which lowers the catechol estrogens available for conversion to estrogen quinones. In turn, the estrogen quinones undergo conjugation with GSH via the catalytic action of glutathione S-transferases (GST). The formation of GSH-estrogen conjugates would reduce the level of estrogen quinones and thereby lower the potential for DNA damage. Based on protein levels, GSTP1 is the most important member of the GST family expressed in breast tissue (74). The 2 other GST isoforms, GSTM1 and GSTA1, are expressed at lower levels. In fact, about 50% of Caucasian women possess the GSTM1 null genotype and therefore completely lack GSTM1 expression in all tissues including breast (75). Based on these considerations, we cloned wild-type GSTP1 cDNA and prepared the purified, recombinant enzyme to assess its role in the estrogen metabolic pathway. We showed that GSTP1 converted the estrogen quinones into estrogen-GSH conjugates (31). Several nonsynonymous GSTP1 polymorphisms have been described with altered catalytic activity toward polycyclic aromatic hydrocarbon carcinogens (76, 77). With regard to estrogen substrates, it is presently unknown whether the variants differ from wild-type GSTP1 in their ability to convert carcinogenic estrogen quinones to nontoxic estrogen-GSH conjugates. In future experiments, we could determine the kinetic rate constants for the variant GSTP1 isoforms and utilize them to account for genetic differences between women in the production of these noncarcinogenic estrogen metabolites.

Another limitation of our model is the lack of actual estrogen metabolite measurements. However, it would be difficult if not impractical to obtain a sufficient number of samples to truly reflect a woman's lifetime endogenous and exogenous estrogen exposure. Thus, we derived the overall exposure by taking into account her total years of ovulation as a function of current age, age at menarche, age at menopause, numbers of full-term pregnancies, and the use of OC and HRT. Our estimates could be improved by taking into account genetic information related to the CYP19A1 gene, which encodes aromatase as sole enzyme producing the parent estrogens E$_2$ and E$_1$. As mentioned above, certain CYP19A1 haplotypes were shown to be associated with increased endogenous estrogen levels in postmenopausal women (73).

In a discussion of mathematical modeling, A.M. Turing wrote: "This model will be a simplification and an idealization, and consequently a falsification. It is to be hoped that features retained for discussion are those of greatest importance in the present state of knowledge." (78) The genotypic–phenotypic approach to modeling reflects this paradigm. The model contains different facets that can be manipulated to strengthen its predictive powers. Furthermore, its flexibility allows one to change the metabolism pathway and/or the phenotypic parameters. For example, incorporation of additional enzymes (e.g., CYP19A1, GSTP1) and their variants into the pathway is easily accomplished by adding suitable differential equations with appropriate kinetic constants to the set of differential equations of the metabolism pathway. Similarly, if another phenotypic parameter became available (e.g.,

alcohol consumption with categorical data), it could readily be incorporated into the model. Regular alcohol consumption has been linked to an increase in breast cancer risk. A meta-analysis of 98 studies showed an excess risk of 22% for drinkers versus nondrinkers with a dose–response relationship among women who drink moderate to high levels of alcohol (79, 80). Thus, the relationship between alcohol and breast cancer appears to be causal but the mechanism for this association is not well understood. One potential mechanism is the influence of alcohol intake on estrogen metabolism. Animal experiments have shown that ethanol consumption increases hepatic aromatase activity, which, in turn, could increase the conversion of androgens to estrogens (81). Indeed, several studies observed a positive correlation between alcohol intake in women and both blood and urinary estrogen concentrations but other studies found no correlation or even an inverse association (82, 83). However, postmenopausal women receiving HRT experienced a significant and sustained increase in circulating estrogen following ingestion of alcohol (84). Women drinking 20 g/day or more and who used HRT had an increased risk of breast cancer (RR 2.24; 95% CI: 1.59–3.14) compared with nondrinkers who never used HRT (85). In light of the latter association, the model could be refined by inclusion of alcohol consumption in the subgroup of women who received HRT. Other, seemingly unrelated, factors can also be tied into the AUC model. For example, because the genotypic–phenotypic model is based on the formation of DNA adducts in the estrogen metabolism pathway, a dynamic system (a submodel) for the enzymatic repair of these adducts can be integrated into the model (86). This would permit one to investigate women who have the genetic machinery that produces high 4-OHE$_2$-AUC values and their accompanying risk, but have effective DNA repair machinery, thus mitigating the breast cancer risk. This flexibility allows us, as stated above by Turing, to experiment with the model by adding and/or removing components to enhance its ability to predict breast cancer risk.

In summary, the current study presents a model for the prediction of breast cancer risk that incorporates the mammary estrogen metabolism pathway, genetic enzyme variants, and traditional risk factors related to estrogen exposure. The model was applied to 2 separate case–control studies and has the potential to give a personalized risk estimate to allow more targeted screening and possibly earlier diagnosis and treatment of the disease.

## Disclosure of Potential Conflicts of Interest

## Grant Support

## References

1. MacMahon B, Feinleib M. Breast cancer in relation to nursing and menopausal history. J Natl Cancer Inst 1960;24:733–53.
2. Parl FF. Estrogens, estrogen receptor and breast cancer. Amsterdam, the Netherlands: IOS Press; 2000.
3. Pike MC, Krailo MD, Henderson BE, Casagrande JT, Hoel DG. 'Hormonal' risk factors, 'breast tissue age' and the age-incidence of breast cancer. Nature 1983;303:767–70.
4. Yager JD, Davidson NE. Estrogen carcinogenesis in breast cancer. N Engl J Med 2006;354:270–82.
5. Devanesan P, Santen RJ, Bocchinfuso WP, Korach KS, Rogan EG, Cavalieri E. Catechol estrogen metabolites and conjugates in mammary tumors and hyperplastic tissue from estrogen receptor-α knockout (ERKO)/Wnt-1 mice: implications for initiation of mammary tumors. Carcinogenesis 2001;22:1573–6.
6. Rogan EG, Badawi AF, Devanesan PD, Meza J, Edney JA, West WW, et al. Relative imbalances in estrogen metabolism and conjugation in breast tissue of women with carcinoma: potential biomarkers of susceptibility to cancer. Carcinogenesis 2003; 24:697–702.
7. Zhu BT, Conney AH. Functional role of estrogen metabolism in target cells: review and perspectives. Carcinogenesis 1998;19:1–27.
8. Cavalieri EL, Stack DE, Devanesan PD, Todorvic R, Dwivedy I, Higginbotham S, et al. Molecular origin of cancer: catechol estrogen-3,4-quinones as endogenous tumor initiators. Proc Natl Acad Sci USA 1997;94:10937–42.
9. Li KM, Todorovic R, Devanesan P, Higginbotham S, Kofeler H, Ramanathan R, et al. Metabolism and DNA binding studies of 4-hydroxyestradiol and estradiol-3,4-quinone in vitro and in female ACI rat mammary gland in vivo. Carcinogenesis 2004;25:289–97.
10. Fernandez SV, Russo IH, Russo J. Estradiol and its metabolites 4-hydroxyestradiol and 2-hydroxyestradiol induce mutations in human breast epithelial cells. Int J Cancer 2006;118:1862–8.
11. Zhao Z, Kosinska W, Khmelnitsky M, Cavalieri EL, Rogan EG, Chakravarti D, et al. Mutagenic activity of 4-hydroxyestradiol, but

not 2-hydroxyestradiol, in BB Rat2 embryonic cells, and the mutational spectrum of 4-hydroxyestradiol. Chem Res Toxicol 2006; 19:475–9.

12. Russo J, Fernandez SV, Russo PA, Fernbaugh R, Sheriff FS, Lareef HM, et al. 17-Beta-estradiol induces transformation and tumorigenesis in human breast epithelial cells. FASEB J 2006;20:1622–34.

13. Shekhar MP, Nangia-Makker P, Wolman SR, Tait L, Heppner GH, Visscher DW. Direct action of estrogen on sequence of progression of human preneoplastic breast disease. Am J Pathol 1998;152:1129–32.

14. Newbold RR, Liehr JG. Induction of uterine adenocarcinoma in CD-1 mice by catechol estrogens. Cancer Res 2000;60:235–7.

15. Yue W, Santen RJ, Wang JP, Li Y, Verderame MF, Bocchinfuso WP, et al. Genotoxic metabolites of estradiol in breast: potential mechanism of estradiol induced carcinogenesis. J Steroid Biochem Mol Biol 2003;86:477–86.

16. Embrechts J, Lemiere F, Van Dongen W, Esmans EL, Buytaert P, Van Marck E, et al. Detection of estrogen DNA-adducts in human breast tumor tissue and healthy tissue by combined nano LC-nano ES tandem mass spectrometry. J Am Soc Mass Spectrom 2003; 14:482–91.

17. Markushin Y, Zhong W, Cavalieri EL, Rogan EG, Small GJ, Yeung ES, et al. Spectral characterization of catechol estrogen quinone (CEQ)-derived DNA adducts and their identification in human breast tissue extract. Chem Res Toxicol 2003;16:1107–17.

18. Zhang Q, Aft RL, Gross ML. Estrogen carcinogenesis: specific identification of estrorgen-modified nucleobase in breast tissue from women. Chem Res Toxicol 2008;21:1509–13.

19. Belous AR, Hachey DL, Dawling S, Roodi N, Parl FF. Cytochrome P450 1B1-mediated estrogen metabolism results in estrogen-deoxyribonucleoside adduct formation. Cancer Res 2007;67:812–7.

20. Henderson BE, Feigelson HS. Hormonal carcinogenesis. Carcinogenesis 2000;21:427–33.

21. www.cancer.gov/bcrisktool.

22. Armstrong K, Eisen A, Weber B. Assessing the risk of breast cancer. New Engl J Med 2000;342:564–71.

23. Claus EB, Risch N, Thompson WD. Autosomal dominant inheritance of early-onset breast cancer. Implications for risk prediction. Cancer 1994;73:643–51.

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

25. Rockhill B, Spiegelman D, Byrne C, Hunter DJ, Colditz GA. Validation of the Gail et al. model of breast cancer risk prediction and implications for chemoprevention. J Natl Cancer Inst 2001;93:358–66.

26. Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. Statist Med 2004;23:1111–30.

27. Crooke PS, Ritchie MD, Hachey DL, Dawling S, Roodi N, Parl FF. Estrogens, enzyme variants, and breast cancer: a risk model. Cancer Epidemiol Biomarkers Prev 2006;15:1620–9.

28. Cavalieri EL, Li KM, Balu N, Saeed M, Devanesan P, Higginbotham S, et al. Catechol ortho-quinones: the electrophilic compounds that form depurinating DNA adducts and could initiate cancer and other diseases. Carcinogenesis 2002;23:1071–7.

29. Dawling S, Hachey DL, Roodi N, Parl FF. In vitro model of mammary estrogen metabolism: structural and kinetic differences in mammary metabolism of catechol estrogens 2- and 4-hydroxyestradiol. Chem Res Toxicol 2004;17:1258–64.

30. Dawling S, Roodi N, Mernaugh RL, Wang XY, Parl FF. Catechol-O-methyltransferase (COMT)-mediated metabolism of catechol estrogens: comparison of wild-type and variant COMT isoforms. Cancer Res 2001;61:6716–22.

31. Hachey DL, Dawling S, Roodi N, Parl FF. Sequential action of phase I and II enzymes cytochrome P450 1B1 and glutathione S-transferase P1 in mammary estrogen metabolism. Cancer Res 2003;63:8492–9.

32. Hanna IH, Dawling S, Roodi N, Guengerich FP, Parl FF. Cytochrome P450 1B1 (CYP1B1) pharmacogenetics: association of polymorphisms with functional differences in estrogen hydroxylation activity. Cancer Res 2000;60:3440–4.

33. Parl FF, Egan KM, Li C, Crooke PS. Estrogen exposure, metabolism, and enzyme variants in a model for breast cancer risk prediction. Cancer Informatics 2009;7:109–21.

34. Castagnetta LAM, Granata OM, Traina A, Ravazzolo B, Amoroso M, Miele M, et al. Tissue content of hydroxyestrogens in relation to survival of breast cancer patients. Clin Cancer Res 2002;8:3146–55.

35. Liehr JG, Fang WF, Sirbasku DA, Ari-Ulubelen A. Carcinogenicity of catechol estrogens in Syrian hamsters. J Steroid Biochem 1986;24:353–6.

36. Han X, Liehr JG. 8-Hydroxylation of guanine bases in kidney and liver DNA of hamsters treated with estradiol: role of free radicals in estrogen-induced carcinogenesis. Cancer Res 1994;54:5515–7.

37. Liehr JG, Ricci MJ. 4-Hydroxylation of estrogens as marker of human mammary tumors. Proc Natl Acad Sci USA 1996;93:3294–6.

38. Justenhoven C, Hamann U, Schubert F, Zapatka M, Pierl CB, Rabstein S, et al. Breast cancer: a candidate gene approach across the estrogen metabolic pathway. Breast Cancer Res Treat 2008;108:137–49.

39. Pesch B, Ko YD, Brauch H, Hamann U, Harth V, Rabstein S, et al. Factors modifying the association between hormone-replacement therapy and breast cancer risk. Eur J Epidemiol 2005;20:699–711.

40. Justenhoven C, Pierl CB, Haas S, Fischer HP, Baisch C, Hamann U, et al. The CYP1B1_1358_GG genotype is associated with estrogen receptor-negative breast cancer. Breast Cancer Res Treat 2008;111:171–7.

41. Dupont WD, Page DL. Risk factors for breast cancer in women with proliferative breast disease. N Engl J Med 1985;312:146–51.

42. Page DL, Dupont WD, Rogers LW, Rados MS. Atypical hyperplastic lesions of the female breast. A long-term follow-up study. Cancer 1985;55:2698–708.

43. Dupont WD, Breyer JP, Bradley KM, Schuyler PA, Plummer WD, Sanders ME, et al. Protein phosphatase 2A subunit gene haplotypes and proliferative breast disease modify breast cancer risk. Cancer 2010;116:8–19.

44. Fitzgibbons PL, Henson DE, Hutter RV. Benign breast changes and the risk for subsequent breast cancer: an update of the 1985 consensus statement. Cancer Committee of the College of American Pathologists. Arch Pathol Lab Med 1998;122:1053–5.

45. Page DL, Anderson TJ. Diagnostic histopathology of the breast. Edinburgh, Scotland: Churchill Livingstone; 1987.

46. Page DL, Jensen RA, Simpson JF. Routinely available indicators of prognosis in breast cancer. Breast Cancer Res Treat 1998;51: 195–208.

47. Page DL, Jensen RA, Simpson JF. Premalignant and malignant disease of the breast: the roles of the pathologist. Mod Pathol 1998;11:120–8.

48. Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. Science 1983;220:671–80.

49. Nelder JA, Mead R. A simplex method for function minimization. Comput J 1965;7:308–13.

50. Storn R, Price K. Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. J Global Optimization 1997;11:341–59.

51. Rubin DB. Multiple imputation for nonresponse in surveys. New York: Wiley; 1987.

52. Schafer JL. Analysis of incomplete multivariate data. Boca Raton, FL: Chapman & Hall/CRC; 1997.

53. StataCorp. Stata statistical software: Release 11. Multiple imputation. College Station, TX: Stata Corporation; 2009.

54. Delong ER, Delong DM, Clarkepearson DI. Comparing the areas under 2 or more correlated receiver operating characteristic curves—a nonparametric approach. Biometrics 1988;44:837–45.

55. StataCorp. Stata statistical software: Release 11. College Station, TX: Stata Corporation; 2009.

56. Dunning AM, Healey CS, Pharoah PDP, Teare MD, Ponder BAJ, Easton DF. A systematic review of genetic polymorphisms and breast cancer risk. Cancer Epidemiol Biomarkers Prev 1999;8:843–54.

57. Mitrunen K, Hirvonen A. Molecular epidemiology of sporadic breast cancer. The role of polymorphic genes involved in oestrogen biosynthesis and metabolism. Mutation Res 2003;544:9–41.

58. Low YL, Li Y, Humphreys K, Thalamuthu A, Li Y, Darabi H, et al. Multi-variant pathway association analysis reveals the importance of genetic determinants of estrogen metabolism in breast and endometrial cancer susceptibility. PLos Genet 2010.p. e1001012.

59. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 2001;69:138–47.

60. Thomas DC. The need for a systematic approach to complex pathways in molecular epidemiology. Cancer Epidemiol Biomarkers Prev 2005;14:557–9.

61. Parl FF, Crooke PS, Conti DV, Thomas DC. Pathway-based methods in molecular cancer epidemiology. In:Rebbeck TR, Ambrosone CB, Shields PG, editors. Molecular epidemiology applications in cancer and other human diseases. New York: Informa Healthcare; 2008. p. 189–204.

62. Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and hormone replacement therapy: collaborative reanalysis of data from 51 epidemiological studies of 52705 women with breast cancer and 108411 women without breast cancer. Lancet 1997;350:1047–59.

63. Shen L, Pisha E, Huang Z, Pezzuto JM, Krol E, Alam Z, et al. Bioreductive activation of catechol estrogen-ortho-quinones: aromatization of the B ring in 4-hydroxyequilenin markedly alters quinoid formation and reactivity. Carcinogenesis 1997;18:1096–101.

64. Pisha E, Lui X, Constantinou AI, Bolton JL. Evidence that a metabolite of equine estrogens, 4-hydroxyequilenin, induces cellular transformation in vitro. Chem Res Toxicol 2001;14:82–90.

65. Zhang F, Swanson SM, Van Breemen RB, Liu X, Yang Y, Gu C, et al. Equine estrogen metabolite of 4-hydroxyequilenin induces DNA damage in the rat mammary tissues: formation of single-strand breaks, apurinic sites, stable adducts, and oxidized bases. Chem Res Toxicol 2001;14:1654–9.

66. Endogenous Hormones and Breast Cancer Collaborative Group: Body Mass index, serum sex hormones, and breast cancer risk in postmenopausal women. J Natl Cancer Inst 2003;95:1218–26.

67. McTiernan A, Tworoger SS, Ulrich CM, Yasui Y, Irwin ML, Rajan KB, et al. Effect of exercise on serum estrogens in postmenopausal women: a 12-month randomized clinical trial. Cancer Res 2004;64:2923–8.

68. Claus EB, Schildkraut JM, THompson WD, Risch NJ. The genetic attributable risk of breast and ovarian cancer. Cancer 1996;77:2318–24.

69. Chenevix-Trench G, Milne RL, Antoniou AC, Couch FJ, Easton DF, Goldgar DE. An international initiative to identify genetic modifiers of cancer risk in BRCA1 and BRCA2 mutation carriers: the consortium of investigators of modifiers of BRCA1 and BRCA2 (CIMBA). Breast Cancer Res 2007;9:104.

70. Hartmann LC, Sellers TA, Frost MH, Lingle WL, Degnim AC, Ghosh K, et al. Benign breast disease and the risk of breast cancer. N Engl J Med 2005;353:229–37.

71. Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer; 2001.

72. Wacholder S, Hartge P, Prentice R, Garcia-Closas M, Feigelson HS, Diver WR, et al. Performance of common genetic variants in breast-cancer risk models. N Engl J Med 2010;362:986–93.

73. Haiman CA, Dossus L, Setiawan VW, Stram DO, Dunning AM, Thomas G, et al. Genetic variation at the CYP19A1 locus predicts circulating estrogen levels but not breast cancer risk in postmenopausal women. Cancer Res 2007;67:1893–7.

74. Kelley MK, Engqvist-Goldstein A, Montali JA, Wheatley JB, Schmidt DE Jr, Kauvar LM. Variability of glutathione S-transferase isoenzyme patterns in matched normal and cancer human breast tissue. Biochem J 1994;304:843–8.

75. Garte S, Gaspari L, Alexandrie AK, Ambrosone C, Autrup H, Aurup JL, et al. Metabolic gene polymorphism frequencies in control populations. Cancer Epidemiol Biomarkers Prev 2001;10:1239–48.

76. Hu X, Xia H, Srivastava SK, Herzog C, Awasthi YC, Ji X, et al. Activity of four allelic forms of glutathione S-transferase hGSTP1–1 for diol epoxides of polycyclic aromatic hydrocarbons. Biochem Biophys Res Commun 1997;238:397–402.

77. Hu X, Xia H, Srivastava SK, Pal A, Wasthi YC, Zimniak P, et al. Catalytic efficiencies of allelic variants of human glutathione S-transferase P1–1 toward carcinogenic anti-diol epoxides of benzo[c]phenanthrene and benzo[g]chrysene. Cancer Res 1998;58:5340–3.

78. Turing AM. The chemical basis of morphogenesis. Philos Trans R Soc Lond B Biol Sci 1952;237:37–72.

79. Key J, Hodgson S, Omar RZ, Jensen TK, Thompson SG, Boobis AR, et al. Meta-analysis of studies of alcohol and breast cancer with consideration of the methodological issues. Cancer Causes Control 2006;17:759–70.

80. Smith-Warner SA, Spiegelman D, Yaun SS, van der Brandt PA, Folsom AR, Goldbohm A, et al. Alcohol and breast cancer in women: a pooled analysis of cohort studies. JAMA 1998;279:535–40.

81. Gordon G, Southren AL, Vittek J, Lieber CS. The effect of alcohol ingestion on hepatic aromatase activity and plasma steroid hormones in the rat. Metabolism 1979;28:20–4.

82. Dorgan JF, Reichman ME, Judd JT, Brown C, Longcope C, Schatzkin A, et al. The relation of reported alcohol ingestion to plasma levels of estrogens and androgens in premenopausal women. Cancer Causes Control 1994;5:53–60.

83. Onland-Moret NC, Peeters PHM, van der Schouw YT, Grobbee DE, van Gils CH. Alcohol and endogenous sex steroid levels in postmenopausal women: a cross-sectional study. J Clin Endocrinol Metab 2005;90:1414–9.

84. McDivit AM, Greendale GA, Stanczyk FZ, Huang MH. Effects of alcohol and cigarette smoking on change in serum estrone levels in postmenopausal women randomly assigned to fixed doses of conjugated equine estrogens with or without a progestin. Menopause 2008;15:382–5.

85. Horn-Ross P, Canchola A, West D, Stewart S, Bernstein L, Deapen D, et al. Patterns of alcohol consumption and breast cancer risk in the California Teachers Study cohort. Cancer Epidemiol Biomarkers Prev 2004;13:405–11.

86. Crooke PS, Parl FP. A mathematical model for DNA damage and repair. J Nucl Acids 2010;352:603–7.