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

Philip S. Crooke1; Christina Justenhoven2, Hiltrud Brauch2 for the GENICA Consortium; Sheila Dawling3, Nady Roodi3, Kathryn S. P. Higginbotham4, W. Dale Plummer5, Peggy A. Schuyler5, Melinda E Sanders3, David L. Page3, Jeffrey R. Smith4, William D. Dupont5, and Fritz F. Parl3

Departments of 1Mathematics, 3Pathology, 4Genetic Medicine, and 5Biostatistics, Vanderbilt University, Nashville, TN 37232 and 2Dr. Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Auerbachstrasse 112, 70376 Stuttgart, and University Tübingen, Germany

Financial support: Supported by NIH grants U54CA113007, 1P50 CA098131-01, R01 CA050468, P30 CA068485, and the Vanderbilt Integrative Cancer Biology Center. The GENICA research work was supported by the Federal Ministry of Education and Research (BMBF) Germany grants 01KW9975/5, 01KW9976/8, 01KW9977/0 and 01KW0114, the Robert Bosch Foundation of Medical Research, Stuttgart, Department of Internal Medicine, Evangelische Kliniken Bonn GmbH, Johanniter Krankenhaus, Bonn, Institute of Pathology, Medical Faculty of the University of Bonn, Deutsches Krebsforschungszentrum, Heidelberg, and Institute for Prevention and Occupational Medicine of the German Social Accident Insurance (IPA), Bochum, Germany.

Corresponding author: Fritz F. Parl, MD, PhD, Department of Pathology, Vanderbilt University, TVC 4918; Nashville, TN 37232; email fritz.parl@vanderbilt.edu

Running Title: Estrogen genotypic-phenotypic model of breast cancer risk

Key Words: estrogen, breast, carcinogenesis, risk, model
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-hydroxyestradiol (4-OHE₂), expressed as area under the curve metric (4-OHE₂-AUC). The model also incorporates phenotypic factors (age, body mass index, hormone replacement therapy, oral contraceptives, family history), which plausibly influence estrogen metabolism and the production of 4-OHE₂. 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-OHE₂-AUC among control subjects had a risk of breast cancer that was 2.30 times that of women at the 10th control 4-OHE₂-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.
Introduction

Abundant experimental and epidemiological 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-OHE₂; 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 (E₂) leads to the formation of 4-OHE₂ 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 such factors 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 in relation to breast cancer risk prediction (27). The model incorporates the main components of mammary estrogen metabolism, i.e., the parent hormone E₂ 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-methyltransferase (COMT) (Fig. 1). The oxidative estrogen metabolism pathway begins with the conversion of E2 by CYP1A1 and CYP1B1 into the catechol estrogens 2- and 4-hydroxyestradiol (2-OHE2, 4-OHE2). These same enzymes further oxidize the catechol estrogens to highly reactive estrogen quinones (E2-2,3-Q, E2-3,4-Q), which can form Michael addition products with deoxyribonucleosides. One of the quinones, E2-3,4-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-methoxyestradiol, 4-MeOE2, 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 E2 into the main metabolites, such as 4-OHE2, the precursor for the mutagenic E2-2,3-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 E2-3,4-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 E₂-3,4-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 (E₁, E₂, E₃), hydroxyestrogens (2-OHE₁, 2-OHE₂, 4-OHE₁, 4-OHE₂, 16α-OHE₁) and methoxyestrogens (2-MeOE₁, 2-MeOE₂, 4-MeOE₁, 4-MeOE₂) in normal and malignant breast revealed the highest concentration of 4-OHE₂ 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-OHE₂ 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-OHE₂, but not 2-OHE₂, induced renal cancer in Syrian hamster (35). Analysis of renal DNA demonstrated that 4-OHE₂ significantly increased 8-hydroxyguanosine levels, whereas 2-OHE₂ did not cause oxidative DNA damage (36). In addition to the induction of renal cancer in the hamster model, 4-OHE₂ is capable of inducing uterine adenocarcinoma, a hormonally related cancer, in mice. Administration of E₂, 2-OHE₂, and 4-OHE₂ induced endometrial carcinomas in 7%, 12%, and 66% of treated CD-1 mice, respectively (14). Examination of microsomal E₂ hydroxylation activity in human breast cancer showed significantly higher 4-OHE₂/2-OHE₂ ratios in tumor tissue than in adjacent normal breast tissue, while the latter tissue samples contained four-fold higher levels of 4-OHE₂ than normal tissue from benign breast biopsies (6, 37).
In the present study, we applied our model to two 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 of breast cancer, body mass index, use of hormone replacement therapy and use of oral contraceptives. 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 08/2000 and 09/2004 as described previously (38, 39). In brief, there are 1143 incident breast cancer cases and 1155 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), oral contraception use (OC) (never, 0 < OC ≤ 5, 5 < OC ≤ 10, OC > 10 years), hormone replacement therapy use (HRT) (never, 0 < HRT ≤ 10, HRT > 10 years), and body mass index (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 TM, Gentra Systems, Inc., Minneapolis, MN). DNA samples were available for 1021 cases (89%) and 1015 controls (88%). Genotyping was performed 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 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 8897 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 performed a nested case-control study of women from this sub-cohort who were postmenopausal at exit. We selected two 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 performed 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, San Diego, CA) for genotyping, using five microliters (≥ 250 ng) of each extracted archival DNA. Each 96-well plate of DNA genotyped contained an average of 4.2 (range 1 to 6) reference saliva DNA samples of study subjects for whom DNA from matching blocks was under evaluation. This enabled assessment of genotype accuracy. The concordance 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 E₂ to the catechol estrogens 2-OHE₂ and 4-OHE₂ by CYP1A1 and CYP1B1. The catechol estrogens are either methylated by COMT to methoxyestrogens (e.g., 4-MeOE₂) or further oxidized by the CYPs to quinones, e.g., E₂-3,4-Q, the main quinone that forms depurinating estrogen DNA adducts. Because of its reactivity, E₂-3,4-Q is short-lived and cannot be readily measured. Instead we chose its immediate precursor, 4-OHE₂, 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, \( A + E \xrightleftharpoons[k_2]{k_1} C \xrightarrow[k_3]{k_2} B + E \), is an enzymatic reaction from a reactant \( A \) to a product \( B \) with \( E \) being the enzyme and \( C \) the \( AE \) complex. Since 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 \( k_1, k_2, k_3 \). 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.

\[
\frac{d(E_2)}{dt} = -\frac{k_{cat1}E_{CYP1B1}E_2}{K_{m1} + E_2} - \frac{k_{cat2}E_{CYP1A1}E_2}{K_{m2} + E_2} - \frac{k_{cat1}E_{CYP1B1}E_2}{K_{m3} + E_2}, \quad E_2(0) = E_{20}^{effective}
\]

\[
\frac{d(2-OHE_2)}{dt} = \frac{k_{cat2}E_{CYP1A1}E_2}{K_{m2} + E_2} + \frac{k_{cat1}E_{CYP1B1}E_2}{K_{m3} + E_2}, \quad 2-OHE_2(0) = 0
\] (1)
\[
\frac{d(4-\text{OHE}_2)}{dt} = \frac{k_{cat} E_{\text{CYP1B1}_1} E_2}{K_m + E_2} - \frac{k_{cat} E_{\text{COMT}} 4-\text{OHE}_2}{K_m + 4-\text{OHE}_2}, \quad 4-\text{OHE}_2(0) = 0
\]

The Michaelis-Menten parameters, \( k_{cat} \) and \( K_m \), in the model have been experimentally measured for each of the three enzymes, CYP1A1, CYP1B1, and COMT ((29-32) and additional unpublished results). The value of \( E_2^{\text{effective}} \) is estimated for each patient as described below. As described in our earlier publication, there are two common non-synonymous polymorphisms in CYP1A1 (codons 461 and 462), four in CYP1B1 (codons 48, 119, 432, and 453), and one in COMT (codon 108), giving rise to four 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 two 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: \( E_2 \xrightarrow{\text{CYP1B1}} 4-\text{OHE}_2 \) with 432Val – 453Asn.

The rate of formation of 4-OHE\(_2\) is approximated by using the kinetic parameters for each of the four 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:

$$\frac{d(4\text{-OHE}_2)}{dt} = \sum_{i=1}^{4} p_i \left( \frac{1.17 E_{\text{CYP1B1}} E_2}{23 + E_2} + \frac{6.00 E_{\text{CYP1B1}} E_2}{19 + E_2} + \frac{3.80 E_{\text{CYP1B1}} E_2}{10 + E_2} + \frac{1.78 E_{\text{CYP1B1}} E_2}{32 + E_2} \right)$$

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

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

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 $A_{\text{menarche}}$ increases, $A_{\text{menopause}}$ decreases and/or $P$ increases, the exposure time to high levels of estrogen during pre-menopause decreases, leading to smaller 4-OHE$_2$-AUC values. The calculation of the average AUC value was then performed as the definite integral: $4\text{-OHE}_2$-AUC = $\frac{1}{T_{\text{reaction}}} \int_{0}^{T_{\text{reaction}}} 4\text{-OHE}_2(t) dt$. In the case of
the pre-menopausal GENICA data, we set $T_{\text{reaction}} = T = 120 \text{ min}$ 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 two key components: (1) the initial estrogen level ($E_2(0) = E_{20}^{\text{effective}}$) and (2) the reaction time over which the 4-OHE$_2$-AUC is calculated. The model uses both categorical (e.g., family history) and quantitative (e.g., age of menarche) inputs to calculate the area under the curve for each individual. For example, the initial estrogen level is affected by BMI, intake of OC or HRT, and family history (FH). Each of these covariates has several categories. In the GENICA study of postmenopausal women these categories for BMI were: (i) BMI $<$ 20, (ii) $20 \leq$ BMI $<$ 25, (iii) $25 \leq$ BMI $<$ 30, and (iv) BMI $\geq$ 30; for HRT: (i) never, (ii) $0 \leq$ 10 years, and (iii) $> 10$ years; and for FH the presence or absence of a first degree family history of breast cancer. The cumulative effect of more than one phenotypic risk factor was assumed to be multiplicative. For example, in the case of post-menopausal women, if the multiplier for BMI is $M_{\text{BMI}}$, the multiplier for HRT is $M_{\text{HRT}}$, and the multiplier for family history is $M_{\text{FH}}$, then the effective initial estrogen level for these three factors is:

$$E_{20}^{\text{effective}} = M_{\text{BMI}} M_{\text{HRT}} M_{\text{FH}} E_{20}.$$  

The value of $E_{20}$ was chosen to be 10, which is consistent with the initial estradiol concentration that was used in the experiments to measure the kinetic constants ($k_{\text{cat}}$ and $K_m$) in the estrogen metabolism pathway. The multipliers for these three categories 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 one less than the number of
categories for the variable. Hence, the total number of weights for the three categorical
variables (BMI, HRT, and FH) is \(3 + 2 + 1 = 7\). The dependent variable in this model was
the patient’s case-control status, while the model’s linear predictor was
\[
\beta_0 + \beta_1 f(x_i; K; \alpha_1, \alpha_2, \ldots, \alpha_7),
\]
where \(x_i\) 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; \(K\) denotes a vector of enzyme kinetic
constants; \(\beta_0, \beta_1, \alpha_1, \alpha_2, \ldots, \alpha_7\) are model and regression parameters; and
\(f(x_i; K; \alpha_1, \alpha_2, \ldots, \alpha_7)\) is the 4-OHE2-AUC derived from our estrogen metabolism pathway
model. The seven parameters \(\alpha_1, \alpha_2, \ldots, \alpha_7\) 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-OHE2-
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 since 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 \(K\)
nor age of menarche, age of menopause, and effects of parity, which enter the model
through the definition of \(T_{reaction}\).
For the pre-menopausal GENICA dataset a similar model was used except that oral contraceptive (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 post-menopausal study, except that the parameters for HRT were re-estimated 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. In order 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 non-linear relationship between these two 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 adjusted odds ratio for developing breast cancer as a function of 4-OHE$_2$-AUC.
The denominator of these odds-ratio 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 1354 women (465 cases, 885 controls) and complete data was available on 345 women (111 cases, 234 controls). Bootstrapping was used to estimate the confidence intervals for the odds ratio curves in Figure 3.

In the GENICA cohort for both premenopausal and postmenopausal patients, the distributions of 4-OHE2-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 (ROCs) obtained from our models were derived using the trapezoidal rule. Ninety-five percent confidence intervals for these areas were estimated using the method of DeLong et al. (54). All analyses were performed using Stata version 11 (55).

Conventional and conditional logistic regression were used to assess the effects of the SNPs in our genotypic-phenotypic model on breast cancer odds ratios 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 one parameter per SNP. The odds ratio for each SNP is adjusted for the other SNPs in the model and represents the breast cancer odds ratio
for women with a heterozygous genotype relative to women with a homozygous wild-type genotype. In this additive model, this odds ratio 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 75$^{th}$ percentiles of 4-OHE$_2$-AUC values in premenopausal cases and controls were 2.03 and 1.73, respectively, while for postmenopausal women they were 1.00 and 0.901, respectively. Figures 2B and 2C show how breast cancer odds ratios 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 90$^{th}$ 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 10$^{th}$ control 4-OHE$_2$-AUC percentile (95% confidence interval (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-OHE2-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 odds ratios associated with the five genotypic components of our model in pre- and postmenopausal women from the GENICA study and in postmenopausal women from the NBC. None of these odds ratios 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 epidemiological 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 one or two 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 epidemiological findings. We chose the two principal components of the model, namely estrogen 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 E₂ and E₁ by having an unsaturated B ring. The amount of human estrogens is much lower, e.g., E₂ accounts for only 1.5% of estrogens present in Premarin (63). In spite of 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-OHE₂, 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 E₂. 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, family history 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 family history, we used a weighting factor, $M_{FH}$, to optimize separation of cases and controls.

Benign breast disease encompasses a spectrum of histological entities, usually subdivided into nonproliferative lesions, proliferative lesions without atypia, and atypical hyperplasia (41, 70). Analysis of the original NBC demonstrated that the latter two types of lesions have clinically significant pre-malignant potential (41). In a more recent study of 9087 women followed for a median of 15 years, the relative risk of breast cancer associated with proliferative changes without atypia was 1.88 (95% confidence interval 1.66–2.12) and increased for atypical hyperplasia to 4.24 (95% confidence interval 3.26-5.41) (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 to the absence of proliferative lesions.
Figures 3B and 3C should be compared to Figure 2C. The odds ratio curves from the NBC in Figure 3 were derived using the weights for BMI and FH derived from the GENICA data set. There are nine 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 $\geq 20$ cases and controls per parameter to avoid over-fitting (71). Hence, model over-fitting 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 odds ratio 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 odds ratios associated with equivalent percentile values. For example, adding a proliferative disease history increases the 90th vs. 10th 4-OHE$_2$-AUC percentile odds ratio from 1.81 to 1.83. In marked contrast to Figures 2B, 2C, 3B and 3C, 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 due to 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 family history of breast cancer (23). The Gail model incorporates primary and secondary family history 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 two 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 et al. (72) reported that the Gail model achieves an area under the receiver operating curve (ROC) of 0.534. The addition of seven 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, i.e., patient age, age at menarche and first birth, number of biopsies, presence of atypical hyperplasia in these biopsies, and family history (21, 24) for a direct comparison of our new model with the Gail model. The area under the ROC curve associated with our 4-OHE2-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 three genes, albeit of primary importance for mammary estrogen metabolism. Another important gene, CYP19A1, encodes aromatase, the main enzyme producing E2 and E1 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 E2 concentration to be converted by CYP1A1, CYP1B1, and COMT to carcinogenic metabolites. Among the phase II conjugating enzymes, COMT is the sole methylating enzyme while there are potentially three glutathione-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 glutathione (GSH) via the catalytic action of GSTs. 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 two 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 non-synonymous GSTP1 polymorphisms have been described with altered catalytic activity towards 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 non-carcinogenic 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 E2 and E1. 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 who used HRT had an increased risk of breast cancer (RR 2.24; 95% CI 1.59 – 3.14) compared to 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, since 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 two 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.
Acknowledgements

The GENICA Consortium includes Christina Justenhoven and Hiltrud Brauch: Dr. Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart, and University of Tübingen, Germany; Yon-Dschun Ko and Christian Baisch: Department of Internal Medicine, Evangelische Kliniken Bonn gGmbH, Johanniter Krankenhaus, Bonn, Germany; Ute Hamann: Molecular Genetics of Breast Cancer, Deutsches Krebsforschungszentrum (DKFZ), Heidelberg, Germany; Volker Harth, Sylvia Rabstein, Anne Spickenheuer, Beate Pesch and Thomas Brüning: Institute for Prevention and Occupational Medicine of the German Social Accident Insurance (IPA), Bochum, Germany; Susanne Haas and Hans-Peter Fischer: Institute of Pathology, Medical Faculty of the University of Bonn, Germany.
**Table 1. Phenotypic and Genotypic Factors examined in GENICA and NBC**

<table>
<thead>
<tr>
<th>Phenotypic Factor</th>
<th>GENICA</th>
<th>NBC</th>
<th>Phenotypic Factor</th>
<th>GENICA</th>
<th>NBC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>yes continuous</td>
<td>yes continuous</td>
<td>Menopausal Status</td>
<td>yes 2</td>
<td>yes 2</td>
</tr>
<tr>
<td>Menarche Age</td>
<td>yes continuous</td>
<td>yes continuous</td>
<td>Menopause Age</td>
<td>yes continuous</td>
<td>yes continuous</td>
</tr>
<tr>
<td>Parity</td>
<td>yes continuous</td>
<td>yes continuous</td>
<td>Age at First Birth</td>
<td>yes continuous</td>
<td>yes continuous</td>
</tr>
<tr>
<td>Primary Family History</td>
<td>yes 2</td>
<td>yes 2</td>
<td>Hormone Replacement Therapy</td>
<td>yes 3</td>
<td>yes 5</td>
</tr>
<tr>
<td>Body Mass Index</td>
<td>yes 4</td>
<td>yes 4</td>
<td>Oral Contraceptives</td>
<td>yes 4</td>
<td>yes</td>
</tr>
<tr>
<td>Proliferative Disease</td>
<td>no</td>
<td>yes 3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genotypic Factor</th>
<th></th>
<th></th>
<th>Genotypic Factor</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CYP1A1 rs1799814</td>
<td>C/A</td>
<td>Thr461Asn</td>
<td>CYP1B1 rs1056836</td>
<td>G/C</td>
<td>Val432Leu</td>
</tr>
<tr>
<td>CYP1A1 rs1048943</td>
<td>A/G</td>
<td>Ile462Val</td>
<td>CYP1B1 rs1800440</td>
<td>A/G</td>
<td>Asn453Ser</td>
</tr>
<tr>
<td>COMT rs4680</td>
<td>G/A</td>
<td>Val108Met</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Kinetic parameters for wild-type and variants of CYP1A1, CYP1B1 and COMT

<table>
<thead>
<tr>
<th>Reaction</th>
<th>Enzyme</th>
<th>Allele</th>
<th>(k_{cat})</th>
<th>(K_m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(E_2 \rightarrow 2)-OHE(_2)</td>
<td>CYP1A1</td>
<td>461Thr-462Ile (wt)</td>
<td>1.10</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>461Asn-462Ile</td>
<td>0.80</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>461Thr-462Val</td>
<td>2.70</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>461Asn-462Val</td>
<td>1.30</td>
<td>23</td>
</tr>
<tr>
<td>(E_2 \rightarrow 2)-OHE(_2)</td>
<td>CYP1B1</td>
<td>48Arg-119Ala-432Val-453Asn (wt)</td>
<td>1.30</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48Gly-119Ala-432Val-453Asn</td>
<td>3.20</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48Arg-119Ser-432Val-453Asn</td>
<td>2.30</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48Arg-119Ala-432Leu-453Asn</td>
<td>2.30</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48Arg-119Ala-432Val-453Ser</td>
<td>2.80</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48Gly-119Ser-432Leu-453Ser</td>
<td>2.50</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48Arg-119Ala-432Leu-453Ser</td>
<td>3.90</td>
<td>136</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48Gly-119Ser-432Leu-453Asn</td>
<td>2.00</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48Gly-119Ser-432Val-453Asn</td>
<td>2.00</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48Gly-119Ser-432Val-453Ser</td>
<td>2.10</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48Arg-119Ser-432Leu-453Ser</td>
<td>0.81</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48Gly-119Ser-432Leu-453Ser</td>
<td>0.47</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48Arg-119Ala-432Leu-453Ser</td>
<td>0.47</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48Gly-119Ser-432Leu-453Asn</td>
<td>0.19</td>
<td>9.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48Gly-119Ser-432Val-453Ser</td>
<td>0.65</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48Gly-119Ala-432Val-453Ser</td>
<td>0.23</td>
<td>75</td>
</tr>
<tr>
<td>(E_2 \rightarrow 4)-OHE(_2)</td>
<td>CYP1B1</td>
<td>48Arg-119Ala-432Val-453Asn (wt)</td>
<td>1.17</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48Gly-119Ala-432Val-453Asn</td>
<td>6.00</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48Arg-119Ser-432Val-453Asn</td>
<td>3.80</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48Arg-119Ala-432Leu-453Asn</td>
<td>1.85</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48Arg-119Ala-432Val-453Ser</td>
<td>4.50</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48Gly-119Ser-432Leu-453Ser</td>
<td>4.40</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48Arg-119Ala-432Leu-453Ser</td>
<td>3.31</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48Gly-119Ser-432Leu-453Asn</td>
<td>1.77</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48Gly-119Ser-432Val-453Ser</td>
<td>1.78</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48Gly-119Ser-432Val-453Ser</td>
<td>2.05</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48Arg-119Ser-432Leu-453Asn</td>
<td>0.70</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48Arg-119Ser-432Leu-453Ser</td>
<td>1.90</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48Gly-119Ala-432Leu-453Ser</td>
<td>1.80</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48Gly-119Ala-432Leu-453Asn</td>
<td>0.73</td>
<td>7.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48Arg-119Ser-432Val-453Ser</td>
<td>2.70</td>
<td>17</td>
</tr>
<tr>
<td>(4)-OHE(_2) (\rightarrow) (4)-MeOE(_2)</td>
<td>COMT</td>
<td>108Val (wt)</td>
<td>3.40</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>108Met</td>
<td>2.04</td>
<td>24</td>
</tr>
</tbody>
</table>
Table 3. Effect of individual genotypic factors on breast cancer odds ratios in the GENICA and NBC populations

<table>
<thead>
<tr>
<th>Gene</th>
<th>SNP</th>
<th>Odds Ratio</th>
<th>95% Confidence Interval</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GENICA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Premenopausal women</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>CYP1A1</em></td>
<td>rs1799814</td>
<td>1.20</td>
<td>( 0.61 – 2.37 )</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>rs1048943</td>
<td>0.52</td>
<td>( 0.24 – 1.13 )</td>
<td>0.10</td>
</tr>
<tr>
<td><em>CYP1B1</em></td>
<td>rs1056836</td>
<td>1.17</td>
<td>( 0.88 – 1.56 )</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>rs1800440</td>
<td>1.36</td>
<td>( 0.93 – 1.99 )</td>
<td>0.12</td>
</tr>
<tr>
<td><em>COMT</em></td>
<td>rs4680</td>
<td>0.97</td>
<td>( 0.75 – 1.26 )</td>
<td>0.82</td>
</tr>
<tr>
<td>Postmenopausal women</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>CYP1A1</em></td>
<td>rs1799814</td>
<td>0.75</td>
<td>( 0.53 – 1.07 )</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>rs1048943</td>
<td>0.89</td>
<td>( 0.61 – 1.31 )</td>
<td>0.57</td>
</tr>
<tr>
<td><em>CYP1B1</em></td>
<td>rs1056836</td>
<td>1.02</td>
<td>( 0.87 – 1.20 )</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>rs1800440</td>
<td>1.07</td>
<td>( 0.87 – 1.32 )</td>
<td>0.50</td>
</tr>
<tr>
<td><em>COMT</em></td>
<td>rs4680</td>
<td>0.93</td>
<td>( 0.81 – 1.07 )</td>
<td>0.32</td>
</tr>
<tr>
<td><strong>NBC</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Postmenopausal women</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>CYP1A1</em></td>
<td>rs1799814</td>
<td>1.16</td>
<td>( 0.82 – 1.64 )</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>rs1048943</td>
<td>1.10</td>
<td>( 0.75 – 1.63 )</td>
<td>0.62</td>
</tr>
<tr>
<td><em>CYP1B1</em></td>
<td>rs1056836</td>
<td>1.08</td>
<td>( 0.90 – 1.29 )</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>rs1800440</td>
<td>1.01</td>
<td>( 0.79 – 1.28 )</td>
<td>0.97</td>
</tr>
<tr>
<td><em>COMT</em></td>
<td>rs4680</td>
<td>0.98</td>
<td>( 0.83 – 1.16 )</td>
<td>0.86</td>
</tr>
</tbody>
</table>
**Figure Legends**

**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 E₂ to catechol estrogens 2-OHE₂ and 4-OHE₂. The catechol estrogens are either methylated by COMT to methoxyestrogens (e.g., 4-MeOE₂ as main fraction) or further oxidized by the CYPs to quinones, e.g., E₂-3,4-Q, the main quinone that forms depurinating estrogen DNA adducts. Each of the three genes CYP1A1, CYP1B1, and COMT is genotyped for all subjects and the SNP genotype data 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-OHE₂, for each haplotype configuration as well as the weighted average of all 4-OH-E₂ 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, BMI) and exposure time (age, ages at menarche and menopause, parity), as well as by family history of breast cancer and proliferative disease. The weight of each phenotypic factor on hormone concentration is determined by maximum likelihood estimation. Thus, the combined genotypic-phenotypic model allows the calculation of 4-OHE₂ produced by each woman, expressed as 4-OHE₂ area under the curve metric (4-OHE₂-AUC), for a personalized risk estimate of developing breast cancer.
**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-OHE₂-AUC values for cases (red) and controls (blue). The boxes span the 4-OHE₂-AUC values from the 25\textsuperscript{th} to 75\textsuperscript{th} percentiles, the vertical line within each box indicates the median value. Whiskers extend to the extreme values or to 1.5 times the inter-quartile range. Outliers beyond the whiskers are plotted individually. (B) Graph of breast cancer odds ratios as a function of 4-OHE₂-AUC values in premenopausal women. The denominator of these odds ratios are the breast cancer odds associated with a woman whose 4-OHE₂-AUC value equals the median value for a premenopausal control. (C) The analogous odds ratio 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-OHE₂-AUC values for cases and controls (see Figure 2 for additional explanation). (B & C) Graphs of breast cancer odds ratios as a function of 4-OHE₂-AUC values in postmenopausal women. The denominator of these odds ratios are the breast cancer odds associated with a woman whose 4-OHE₂-AUC value equals the median value for a postmenopausal control. Graphs B and C use 4-OHE₂-AUC models that exclude and include, respectively, each patient’s history of benign proliferative breast disease.
References


Nashville Breast Cohort
Postmenopausal Women

A
Proliferative Disease Included in Model

Control

No

Case

Control

Yes

Case

$P = 3.5 \times 10^{-4}$

$P = 2.6 \times 10^{-7}$

B
Model Excludes Proliferative Disease

$P = 7.6 \times 10^{-4}$

95% CI

C
Model Includes Proliferative Disease

$P = 9.5 \times 10^{-7}$

95% CI