

Breast Cancer Res Treat (2013) 142:187–202  
DOI 10.1007/s10549-013-2719-3

EPIDEMIOLOGY

## Validation of Rosner–Colditz breast cancer incidence model using an independent data set, the California Teachers Study

B. A. Rosner · G. A. Colditz · S. E. Hankinson ·  
J. Sullivan-Halley · J. V. Lacey Jr. ·  
L. Bernstein

Received: 3 July 2013 / Accepted: 28 September 2013 / Published online: 26 October 2013  
© The Author(s) 2013. This article is published with open access at [Springerlink.com](http://Springerlink.com)

**Abstract** To validate an established breast cancer incidence model in an independent prospective data set. After aligning time periods for follow-up, we restricted populations to comparable age ranges (47–74 years), and followed them for incident invasive breast cancer (follow-up 1994–2008, Nurses' Health Study [NHS]; and 1995–2009, California Teachers Study [CTS]). We identified 2026 cases during 540,617 person years of follow-up in NHS, and 1,400 cases during 288,111 person years in CTS. We

fit the Rosner–Colditz log-incidence model and the Gail model using baseline data. We imputed future use of hormones based on type and prior duration of use and other covariates. We assessed performance using area under the curve (AUC) and calibration methods. Participants in the CTS had fewer children, were leaner, consumed more alcohol, and were more frequent users of postmenopausal hormones. Incidence rate ratios for breast cancer showed significantly higher breast cancer in the CTS (IRR = 1.32, 95 % CI 1.24–1.42). Parameters for the log-incidence model were comparable across the two cohorts. Overall, the NHS model performed equally well when applied in the CTS. In the NHS the AUC was 0.60 (s.e. 0.006) and applying the NHS betas to the CTS the performance in the independent data set (validation) was 0.586 (s.e. 0.009). The Gail model gave values of 0.547 (s.e. 0.008), a significant 4 % lower,  $p < 0.0001$ . For women 47–69 the AUC values for the log-incidence model are 0.608 in NHS and 0.609 in CTS; and for Gail are 0.569 and 0.572. In both cohorts, performance of both models dropped off in older women 70–87, and later in follow-up (6–12 years). Calibration showed good estimation against SEER with a non-significant 4 % underestimate of overall breast cancer incidence when applying the model in the CTS population ( $p = 0.098$ ). The Rosner–Colditz model performs consistently well when applied in an independent data set. Performance is stronger predicting incidence among women 47–69 and over a 5-year time interval. AUC values exceed those for Gail by 3–5 % based on AUC when both are applied to the independent validation data set. Models may be further improved with addition of breast density or other markers of risk beyond the current model.

B. A. Rosner · S. E. Hankinson  
Channing Division of Network Medicine, Department of  
Medicine, Harvard Medical School, Boston, MA, USA

B. A. Rosner  
Department of Biostatistics, Harvard School of Public Health,  
Boston, MA, USA

G. A. Colditz (✉)  
Alvin J. Siteman Cancer Center at Barnes-Jewish Hospital and  
Washington University School of Medicine, 660 South Euclid  
Avenue, Campus Box 8100, Saint Louis, MO 63110, USA  
e-mail: [colditzg@wustl.edu](mailto:colditzg@wustl.edu)

S. E. Hankinson  
Division of Biostatistics and Epidemiology, University of  
Massachusetts, Amherst, MA, USA

S. E. Hankinson  
Channing Division of Network Medicine, Department of  
Medicine, Brigham and Women's Hospital, Boston, MA, USA

J. Sullivan-Halley · J. V. Lacey Jr. · L. Bernstein  
Division of Cancer Etiology, Department of Population  
Sciences, Beckman Research Institute of the City of Hope,  
Duarte, CA, USA

**Keywords** Breast cancer · Prediction models ·  
Validation · Calibration · Methods

## Introduction

For over a decade since developing and expanding the Rosner–Colditz model for breast cancer incidence [1, 2], we have sought approaches to estimating performance in an independent validation data set. Although we have conducted internal validation using split sample approaches [3], we have not previously used an independent data set to assess performance. This has largely been due to the need for data on age at each birth for women, an input to spacing of births that directly relates to breast cancer risk in early studies [4] and is confirmed in our model [5] and by others [6]. The closer births are together, the more rapidly breast tissue-aging decreases and the lower total risk accumulates through premenopausal years [7]. In addition, details on age at menopause and type of menopause as well as type and duration of postmenopausal hormone therapy (HT) are important risk factors.

Our approach then is to use an independent data set to estimate performance following the principles outlined in literature addressing validation and application of prediction models in medicine [8, 9]. To date, no model of breast cancer incidence has been implemented as part of routine clinical care where risk estimates might guide level of screening, genetic counseling, or chemoprevention.

As previously noted, the Rosner–Colditz model includes a range of established reproductive factors, body mass index (BMI), and alcohol intake in its basic form [2]. This is one of a large number of breast cancer risk prediction models. In a systematic review and meta-analysis, Meads et al. [10] identified 17 breast cancer risk models with differing sets of modifiable and non-modifiable risk factors, with many omitting age at menopause, type of menopause, and use of postmenopausal hormones, all factors strongly related to future breast cancer risk. Only four models had validation in potentially independent data sets. These models included Gail [11] and also the Rosner–Colditz model [1, 2, 12]. The performance of the Gail model summarized as AUC in a previous validation within the NHS data was 0.58, though both have not been compared in a common independent data set.

Moons and others emphasize a sequence of model development, validation, application, and assessment of performance in application/clinical setting [8, 9]. To date, we find no reports on the last aspect of breast cancer model performance in routine clinical settings. Here we focus on the conduct of validation in an independent data set.

We collaborated with California Teachers Study (CTS) investigators to draw on an independent prospective data set and assess the performance of the Rosner–Colditz model, which was developed and refined in the Nurses' Health Study (NHS). We also compare model performance against the Gail model when both are fit to the independent data set.

## Methods

As noted above, a key issue in identifying an independent prospective study with appropriate risk factor collection included the need for details of age at each pregnancy, a refinement of usual reporting of age at first birth and number of births typical of epidemiologic studies. Details on age and type of menopause were also important since this is omitted from the Gail model despite a long record of being established as a modifier of future breast cancer risk [5, 13, 14]. Other key risk factors not included in the Gail model are duration and type of postmenopausal HT used [15], BMI [16], and alcohol intake [17]. These are all in the Rosner–Colditz log-incidence model.

**CTS** This cohort contains the necessary data collected at baseline in 1995 for the cohort. The CTS approach to questionnaire follow-up, after 2 years, then after 3 more years, then at varying intervals each updating some exposures, together with case ascertainment ongoing annually through the California tumor registry, meant we use baseline data only. We limit the population to women who were postmenopausal at baseline. To compare incidence during common follow-up time periods we use the time frame for CTS from baseline 1995 to 2009.

**NHS** This cohort of women followed from 1976 has routinely updated information every 2 years on reproductive risk factors for breast cancer, family history of breast cancer, use of postmenopausal hormones, and from 1980 onwards alcohol intake. The original Rosner–Colditz model was developed in the broader NHS cohort [1, 2, 5]. For comparability with data available from the CTS, we limit the population for this analysis to women who were postmenopausal at baseline in 1994. Thus the corresponding time available for the NHS is 1994–2008. In 1994, NHS participants were 47–74. Hence, we limit the CTS participants included in the analysis to a comparable age range, excluding their older cohort members.

### Model fitting issues

Limited only to baseline data from the CTS, we modified the Rosner–Colditz model to omit updating. Because this differs from our standard approach of updating exposure information every 2 years [2], we estimate the impact of this modification on overall performance.

Duration of current use of postmenopausal HT is significantly related to incidence of breast cancer [2, 18], and to type of menopause, age at menopause, and time since menopause. These factors are all importantly related to postmenopausal breast cancer incidence. We, therefore, used imputation methods to estimate future duration of use for postmenopausal HT in the CTS [19]. We used a two-step process to estimate use according to type of hormone used

currently, and duration of use. We first fit a model to NHS data to estimate the duration of hormone use from 1994 to the return of the 2006 follow-up questionnaire for each type of HT (estrogen, E, alone and estrogen plus a progestin, E&P). Predictors included menopause type and time since menopause, and duration of use of HT among current users (see Tables 8 and 9). In addition to these characteristics of menopause, parity was positively related to ever use of E alone but not E&P, and positively to duration of use of estrogen alone, but inversely to duration of estrogen plus progestin. BMI was inversely related to ever use of E and E&P, but was unrelated to duration of use of either. Alcohol use was inversely related to ever use of E alone and to ever use of E&P, but not to duration of use of either formulation. We developed this model separately for use of E alone and for use of E&P. We then used this model with baseline CTS data to impute future use by type and duration for participants, taking the average of 5 imputations for each participant. (See Tables 8 and 9 for the imputation models and Appendix 2 for a summary of the imputation strategy.)

#### Time frame

To compare incidence of breast cancer in the two cohorts over a common time frame, we identified common subsets from the two cohorts. We use the CTS baseline in 1995 and 1994 as the start point for inclusion of NHS follow-up. We then draw on the age range of the NHS participants to define a comparable age range for CTS participants. Thus we limit NHS follow-up data to the interval 1994–2008. CTS data for the corresponding years are included with follow-up from 1995 to 2009.

During follow-up of the NHS cohort from 1994 to 2008, we identified 2,026 invasive breast cancer diagnoses among postmenopausal women during 540,617 person years. In the CTS, we identified 1,400 incident invasive breast cancer diagnoses among postmenopausal women during 288,111 person-years.

#### Description of the log-incidence model of breast cancer

We assume that the incidence of breast cancer at time  $t$  ( $I_t$ ) is proportional to the number of cell divisions accumulated throughout life up to age  $t$  (i.e.,  $I_t = kC_t$ ).

$C_t$  is obtained from

$$C_t = C_0 \times \prod_{i=0}^{t-1} (C_{i+1}/C_i) = C_0 \times \prod_{i=0}^{t-1} \lambda_i \quad (1)$$

Thus,  $\lambda_i = \frac{C_{i+1}}{C_i}$  = the rate of increase in  $C_t$  from age  $i$  to age  $i + 1$ .

$\log(\lambda_i)$  is assumed to be a linear function of risk factors that are relevant at age  $i$ . The set of relevant risk factors

and their magnitude and/or direction may vary according to the stage of reproductive life. We fit PROC NLIN of SAS to estimate the parameters of the model with breast cancer risk factors including (1) duration of premenopause, (2) duration postmenopause, (3) type of menopause, natural or surgical (4) parity, (5) age at each birth, (6) current, past HRT use, (7) duration of HT use by type, (8) BMI, premenopause  $\equiv$  BMI<sub>1</sub>, (9) BMI, postmenopause  $\equiv$  BMI<sub>2</sub>, (10) height, (11) benign breast disease (BBD), (12) alcohol intake, (13) family history of breast cancer.

We fit the base model using baseline variables and imputed HT duration without updating exposures and assessed covariates using the CTS comparing their magnitude and direction to the variables in the NHS. We assess the performance of the model from the NHS in the CTS by fitting the NHS model and averaging five imputations of HT use. We fit the Gail model [11] using the formula from page 1880, with the caveat that in each cohort the number of previous biopsies is scored 0 or 1 and the number of relatives with family history is scored 0 or 1. We compare the  $c$ -statistic for Gail versus Rosner–Colditz log-incidence using the Wilcoxon rank sum test [20].

To assess calibration, we use the NHS model to estimate relative risks for individual women in the CTS and combine these with SEER data to estimate absolute risk. We then group the CTS participants by decile of estimated absolute risk and compare observed and expected counts of incident breast cancers and test for trend using Poisson regression approaches (for additional details, see Appendix 1).

To assess calibration, we apply the NHS risk model to the CTS population using imputed data for HRT use over 12 years. Suppose there are  $N$  subjects in the CTS population who are followed for  $T$  person-years. We divide the  $T$  person-years into  $L$  age strata and let  $T_l$  = number of person-years in the  $l$ th age stratum. Based on the NHS risk model, we compute the relative risk for the  $i$ th person at the  $j$ th person-year given by  $RR_{ij}$  compared to a hypothetical person at baseline risk where all covariate values are 0. Let  $h_1^*(l)$  be the age-specific incidence rate for the  $l$ th age group from SEER 1995–2006. We use the methods of Gail (1989) to combine the  $RR_{ij}$  from the NHS model with  $h_1^*(l)$  to estimate  $h_1(l)$  = baseline incidence rate for the  $l$ th age group of CTS. An estimate of the incidence rate for the  $i$ th subject in the  $j$ th person-year is then given by

$$\hat{I}_{ij} = \sum_{l=1}^L h_1(l) \delta_{ijl} RR_{ij}$$

where  $\delta_{ijl} = 1$  if the  $i$ th subject is in age group  $l$  at the  $j$ th person-year, = 0 otherwise.

The corresponding estimate of cumulative incidence for the  $i$ th subject over  $t_i$  person-years is given by

$$E_i = 1 - \exp\left(-\sum_{j=1}^{t_i} \hat{I}_{ij}\right)$$

Let  $O_i = 1$  if the  $i$ th subject develops breast cancer over  $t_i$  person-years, = 0 otherwise.

If the NHS model is well calibrated in the CTS population, then  $O_i$  should follow a Poisson distribution with mean =  $E_i$ . To test this we let  $\mu_i = E(O_i)$  and consider the Poisson regression model

$$\ln(\mu_i) = \alpha + \ln(E_i)$$

A test of the calibration of the model at the individual level is

$$H_0 : \alpha = 0 \text{ vs. } H_1 : \alpha \neq 0$$

which we can perform using a Poisson regression model with intercept only and offset given by  $\ln(E_i)$ .

We also can group the subjects into deciles by cumulative incidence per year (or  $E_i^* = E_i/t_i$ ) and compute the observed ( $O^{(d)}$ ) and expected ( $E^{(d)}$ ) number of cases in the  $d$ th decile and run a Poisson regression at the aggregate level of the form:

$$\ln(\mu^{(d)}) = \alpha + \ln(E^{(d)})$$

where  $\mu^{(d)} = E(O^{(d)})$ .

The individual and aggregate Poisson regression models are actually equivalent. The Poisson regression approach should be a more sensitive model of goodness of fit than the Hosmer–Lemeshow statistic given by

$$\chi_{HL}^2 = \sum_{d=1}^{10} \frac{(O^{(d)} - E^{(d)})^2}{E^{(d)}}$$

which is more similar to a test of heterogeneity than the test for trend approach given by Poisson regression.

Finally, to combine inferences over several imputed data sets, multiple imputation approaches are used to obtain an overall test of calibration based on averaging estimates of  $\alpha$  over several imputations. More detail on the calibration methodology is given in Table 8.

## Results

### Risk factor prevalence differences (Tables 1, 2)

Baseline data for the NHS and CTS are presented in Table 1, for women 47–59 years at baseline, and Table 2, for women 60–74 years of age. The mean age, age at menarche, and age at menopause were comparable in the cohorts as were the prevalence of biopsy confirmed BBD and family history of breast cancer. The CTS included more nulliparous women (25 %) versus 6 % in the NHS for

women 47–59 years, and 18 versus 6 % for women 60–74 years. CTS cohort members versus women in the NHS had an average of 1 fewer births per woman; more current postmenopausal hormone use (age 47–59 years: 70 vs. 56 %, age 60–74 years: 53 vs. 35 %) and longer duration of use; leaner current BMI (age 47–59 years: 25.3 vs. 26.6, age 60–79 years: 25.3 vs. 26.1) and higher current alcohol intake (age 47–59: 7.9 g/day vs. 5.0 g/day, 60–79: 8.2 g/day vs. 5.1 g/day).

### Incidence rates (Table 3)

Age-specific and age-adjusted incidence rates show breast cancer incidence rates are higher in the CTS for women over age 60 years (Table 3). Across all ages, 47–87 years, the age-adjusted incidence rate ratio shows that the CTS has significantly higher incidence (age-adjusted IRR 1.32, 95 % CI 1.24–1.42).

### Comparing parameter estimates in each cohort (Table 4)

The modified model using only baseline data and imputed HT duration of use was fit separately to the NHS and then to the CTS cohort data to compare coefficients side by side (see Table 4). We note a number of important similarities across the two independent cohort studies supporting favorable performance. The magnitude of the coefficient for age at first birth (gynecologic age at first birth) is comparable, being positive in both cohorts. The associated birth index (a summary of total years from each birth to minimum [age, or age at menopause], summed over all births in parous women and = 0 for nulliparous women) shows a strong inverse association of comparable magnitude in both cohorts (–0.0032 in NHS vs. –0.0026 in CTS). Thus, for a typical woman with menarche at age 13, menopause at 50, births at 20, 23, 26, 29, (giving a birth index 102), this translates to a RR 0.72 for the NHS and 0.77 for the CTS. Terms for BBD and family history are comparable as are the association for alcohol and for height and BMI among women not taking HT (estrogen negative time).

We also note some differences between the two cohorts. The magnitude of the association for duration of E&P has a larger magnitude in the CTS,  $b = 0.035$  versus 0.015 in NHS. The term for current use is weaker in the CTS, giving a combined relative risk for a current user with 5 years of use of  $e^{0.202 + 5(0.035)} = e^{0.377} = 1.46$  compared to a never user for the CTS and  $e^{0.368 + 5(0.015)} = e^{0.443} = 1.56$  for the NHS. For current users with 10 years of use, the RRs are 1.74 for the CTS and 1.68 for the NHS. Thus, the overall associations for current users are comparable at longer durations of use. The association for BMI is somewhat weaker during

**Table 1** Comparison of baseline risk factors between NHS and CTS, age 47–59

Variable	NHS			CTS		
	Mean $\pm$ SD	Range	<i>N</i>	Mean $\pm$ SD	Range	<i>N</i>
Age	54.8 $\pm$ 3.1	47–59	18,308	54.0 $\pm$ 3.3	47–59	11,419
Age at menarche	12.4 $\pm$ 1.3	9–21	18,308	12.5 $\pm$ 1.4	10–17	11,419
Age at menopause	47.9 $\pm$ 5.1	21–59	18,308	48.0 $\pm$ 4.8	35–56	11,419
Type of menopause						
Natural	13,910 (76 %)			8,393 (74 %)		
Bilateral oophorectomy	4,398 (24 %)			3,026 (26 %)		
Nulliparous (%)	1,075 (6 %)			2,858 (25 %)		
Age at 1st birth <sup>a</sup>	24.6 $\pm$ 3.0	15–46	17,233	25.6 $\pm$ 4.6	14–46	8,561
Parity						
0	1,075 (6 %)			2,858 (25 %)		
1	1,355 (7 %)			1,695 (15 %)		
2	6,016 (33 %)			4,094 (36 %)		
$\geq 3$	9,862 (53 %)			2,772 (24 %)		
Mean	2.7 $\pm$ 1.4	0–15	18,308	1.7 $\pm$ 1.3	0–10	11,419
Birth index	58.0 $\pm$ 33.9	0–236	18,308	36.3 $\pm$ 32.6	0–316	11,419
Age at 1st birth–age at menarche	12.2 $\pm$ 3.3	1–32	17,233	13.1 $\pm$ 4.8	0–32	8,561
Current PMH use	10,232 (56 %)			7,975 (70 %)		
Past PMH use	2,457 (13 %)			1,168 (10 %)		
Duration E use (years)	1.4 $\pm$ 3.6	0–34.0	18,308	2.0 $\pm$ 4.1	0–20.0	11,419
Duration E&P use (years)	1.2 $\pm$ 2.1	0–14.0	18,308	2.1 $\pm$ 3.0	0–20.0	11,419
Current BMI (kg/m <sup>2</sup> )	26.6 $\pm$ 5.3	12.5–68.7	18,308	25.3 $\pm$ 5.3	16.2–60.6	11,419
BMI at age 18 (kg/m <sup>2</sup> )	21.3 $\pm$ 2.9	13.1–43.3	18,308	21.4 $\pm$ 3.4	14.5–53.3	11,419
Height	64.6 $\pm$ 2.4	48–79	18,308	64.8 $\pm$ 2.6	45–76	11,419
Alcohol (g/day)	5.0 $\pm$ 9.3	0–292.8	18,308	7.9 $\pm$ 9.6	0–112.8	11,419
Alcohol at age 18 (g/day)	3.2 $\pm$ 5.1	0–108.1	18,308	5.3 $\pm$ 8.2	0–157.2	11,419
Benign breast disease (%) (biopsy confirmed)	4,203 (23 %)			2,181 (19 %)		
Family Hx breast cancer (%)	2,049 (11 %)			1,497 (13 %)		

<sup>a</sup> Among parous women

estrogen negative time (postmenopause, non-use of postmenopausal hormones) in the CTS compared to the NHS (0.00038 vs. 0.00195 per BMI unit per year).

Summary model performance in NHS and CTS cohorts (Table 5)

We fit model coefficients from Table 4 to NHS and applied the coefficients from NHS to CTS data for follow-up from 1995 to 2009 as an external validation of the NHS model (see Table 5). The overall performance in the NHS was 0.597 for the full follow-up and 0.586 in CTS. For the first 5-year follow-up interval among women 47–69 years, the risk prediction performance was comparable in both cohorts (0.608 in NHS and 0.609 in CTS) supporting validity of the model. We also observed that in NHS during the first 5-year follow-up period, 1994–1999, performance was higher in

women 47–69 years ( $c = 0.608$ ) than in those 70–87 years ( $c = 0.587$ ). For the second follow-up interval from 2000 to 2008 the model again performed better in younger women  $c = 0.599$  compared to older women  $c = 0.577$ , but in each group performance was lower than in the first time interval. This pattern of performance was also observed when the Gail model was applied to the NHS cohort performance was higher in younger women and in the first versus second follow-up interval.

Applying the NHS log-incidence model to the CTS data, a similar pattern emerged; the performance was better during the first 5 years of follow-up in younger than older women ( $c = 0.609$  for 47–69 year old women vs. 0.564 for 70–87 year old women). During the later follow-up, 2001–2009, the performance was further reduced. The Gail model applied to the CTS data also showed this pattern in the first follow-up interval.

**Table 2** Comparison of baseline risk factors between NHS and CTS, age 60–74

Variable	NHS Mean $\pm$ SD	CTS Range	Variable	NHS Mean $\pm$ SD	CTS Range	Variable
Age	66.0 $\pm$ 3.8	60–74	27,434	66.1 $\pm$ 4.1	60–74	11,222
Age at menarche	12.7 $\pm$ 1.4	9–21	27,434	12.6 $\pm$ 1.4	10–17	11,222
Age at menopause	49.2 $\pm$ 4.8	20–66	27,434	49.8 $\pm$ 4.8	35–56	11,222
Type of menopause						
Natural	21,838 (80 %)			9,097 (81 %)		
Bilateral oophorectomy	5,596 (20 %)			2,125 (19 %)		
Nulliparous (%)	1,744 (6 %)			2,012 (18 %)		
Age at 1st birth <sup>a</sup>	25.7 $\pm$ 3.6	16–47	25,690	25.3 $\pm$ 4.2	14–46	9,210
Parity						
0	1,744 (6 %)			2,012 (18 %)		
1	1,879 (7 %)			1,147 (10 %)		
2	6,281 (23 %)			3,037 (27 %)		
$\geq 3$	17,530 (64 %)			5,026 (45 %)		
Mean	3.2 $\pm$ 1.8	0–16	27,434	2.3 $\pm$ 1.6	0–13	11,222
Birth index	64.8 $\pm$ 39.4	0–259	27,434	52.2 $\pm$ 40.6	0–321	11,222
Age at 1st birth–age at menarche	13.0 $\pm$ 3.8	1–35	25,690	12.6 $\pm$ 4.3	1–34	9,210
Current PMH use	9,474 (35 %)			5,942 (53 %)		
Past PMH use	7,580 (28 %)			1,854 (17 %)		
Duration E use (years)	2.4 $\pm$ 5.2	0–48.4	27,434	3.8 $\pm$ 6.5	0–20.0	11,222
Duration E&P use (years)	0.9 $\pm$ 2.1	0–14.0	27,434	3.2 $\pm$ 4.8	0–20.0	11,222
Current BMI (kg/m <sup>2</sup> )	26.1 $\pm$ 5.0	12.9–69.4	27,434	25.3 $\pm$ 4.8	16.0–60.4	11,222
BMI at age 18 (kg/m <sup>2</sup> )	21.3 $\pm$ 3.0	10.8–59.3	27,434	21.5 $\pm$ 3.1	15.0–48.4	11,222
Height	64.4 $\pm$ 2.4	39–79	27,434	64.4 $\pm$ 2.5	56–74	11,222
Alcohol (g/day)	5.1 $\pm$ 9.6	0–113.4	27,434	8.2 $\pm$ 10.4	0–130.8	11,222
Alcohol at age 18 (g/day)	2.5 $\pm$ 4.9	0–161.3	27,434	3.6 $\pm$ 6.4	0–116.1	11,222
Benign breast disease (%) (biopsy confirmed)	6,133 (22 %)			2,408 (21 %)		
Family Hx breast cancer (%)	3,684 (13 %)			1,687 (15 %)		

<sup>a</sup> Among parous women

Comparing the Gail model to the log-incidence model in the independent CTS data, the AUC for the Gail model performance was 4 % lower overall ( $c = 0.547$  vs.  $0.586$ ,  $p < 0.0001$ ); during the first follow-up period for women 47–69 years ( $c = 0.572$  vs.  $0.609$ , difference in AUC =  $0.037$ ,  $p = 0.008$ ), and in women 70–87 years ( $c = 0.516$  vs.  $0.564$ , difference in AUC =  $0.048$ ,  $p = 0.09$ ). In the later follow-up from 2001 to 2009 these differences persisted.

Comparison of  $c$  statistic for actual NHS data versus the use of imputed values in that cohort (Table 6)

To assess the drop off in model performance induced by not updating exposure variables, we next fit the model to NHS data using first imputed and then updated values for HT duration (see Table 6). Fitting the model to NHS

updated data from 1994 through 2008 (right hand panel of Table 6) we observe an AUC  $c$  statistic value of  $0.616$  (s.e.  $0.006$ ). If instead of using observed updated data, we impute future duration of HT after menopause, the AUC  $c$  statistic is reduced modestly to  $0.600$  (s.e.  $0.006$ ). When assessing performance in the early follow-up from baseline and later follow-up—again the actual data were comparable to imputed data for the first 5 years, but showed reduced performance in the 2000–2008 interval. For example, for women 47–69, the AUC decreased from  $0.641$  with actual updated data to  $0.595$  using imputed data.

Calibration observed and expected counts in CTS by decile of risk, predicted with NHS betas

Finally, we use five imputations to estimate the expected number of cases of breast cancer according to the NHS

**Table 3** Comparison of breast cancer incidence rates between NHS and CTS

Age group	NHS			CTS			
	Cases	<i>p</i> _years	Incidence rate (per 10 <sup>5</sup> py)	Cases	<i>p</i> _years	Incidence rate (per 10 <sup>5</sup> py)	IRR
47–49	7	1,764	396.8	2	2,511	79.6	0.20
50–54	74	23,452	315.5	48	18,792	255.4	0.81
55–59	275	66,427	414.0	177	43,889	403.3	0.97
60–64	434	109,802	395.3	298	60,673	491.2	1.24
65–69	523	127,301	410.8	328	61,539	533.0	1.30
70–74	460	117,401	391.8	274	50,987	537.4	1.37
75–79	217	68,627	316.2	197	32,853	599.6	1.90
80–87	36	25,844	139.3	76	16,867	450.6	3.23
Total	2,026	540,618		1,400	288,111		
Crude			374.8			485.9	1.30
Age-adjusted			374.8			500.5	1.32
Crude IRR				1.30			
95 % CI				1.21	1.39		
Age-adjusted IRR				1.32			
95 % CI				1.24	1.42		

Based on SEER data for white women, 1995–2006

model stratifying the CTS participants by decile of risk. As shown in Table 7, the observed count was slightly lower than the predicted case count. Poisson regression across all women allows estimation of the adjustment factor ( $\alpha$ ) =  $-0.048$ , s.e. ( $\alpha$ ) =  $0.027$ ,  $p = 0.074$ . Overall the model fit is not significantly different from SEER, O/E =  $0.96$  a  $4\%$  underestimate. Thus applying the NHS model with its rich use of exposure across the life course for established breast cancer risk factors, and accounting for the risk factor profile of individual women in the CTS, we fully account for breast cancer incidence in this independent population.

## Discussion

We identified an independent large data set with 1,400 incident invasive breast cancer cases, that allowed evaluation of a breast cancer incidence risk prediction models using a common definition of incident invasive breast cancer, over common time periods, and age groups. Age-standardized breast cancer incidence in the CTS was significantly higher than in NHS. Overall performance of the Rosner–Colditz log-incidence model shows AUC consistent with performance in the original NHS, supporting external validity of the model. In the external validation data set the model outperformed the Gail model by  $3\text{--}5\%$  for differing age groups and follow-up intervals based on

the AUC. Although adaptations had to be made using only baseline data, this approach is comparable to using the tool in clinical practice to predict risk and stratify women to guide prevention interventions. Assessment of the lack of updating but use of imputed duration of hormone use among postmenopausal women showed modest attenuation over a 5-year follow-up interval in the NHS. Calibration against SEER showed good performance and close agreement of predicted with observed incidence.

## General issues on validating

Data availability on key reproductive variables including age at first birth, age at each birth, menopause and type of menopause, as well as history of biopsy confirmed BBD and family history of breast cancer, height, weight, and history of alcohol intake supported use of a common model in comparable data that had been collected with similar methods and would reflect approaches in clinical and epidemiologic practice. Because HT modifies risk of breast cancer, imputing future use among current users was necessary as the CTS does not update data every 2 years as NHS does, and in clinical practice future use is unknown but is important for risk prediction. Summary imputation models are provided that may be of use for clinical application in other settings where future use of hormones will be estimated given past history ascertained at a clinic visit

**Table 4** Relationship between Breast Cancer Risk Factors and Breast Cancer, based on an average of 5 imputations of HT experience over 12 years

Variable	Beta	NHS			California Teachers Study	
		2,026 cases			1,400 cases	
		540,618 person-years			288,111 person-years	
		s.e.	<i>p</i> value	Beta	s.e.	<i>p</i> value
Constant	-7.420	0.352	<0.001	-8.048	0.322	<0.001
Duration of premenopause	0.044	0.009	<0.001	0.056	0.008	<0.001
Duration postmenopause						
Natural menopause	-0.009	0.005	0.069	0.017	0.006	0.002
Bilateral oophorectomy	-0.015	0.006	0.013	0.012	0.008	0.15
Pregnancy history						
Gynecologic age at 1st birth <sup>a</sup>	0.0089	0.0048	0.062	0.0055	0.0041	0.18
Birth index	-0.0032	0.0007	<0.001	-0.0026	0.0008	0.001
BBD						
BBD (yes vs. no)	0.237	0.588	0.69	0.314	0.834	0.71
BBD × age at menarche	0.021	0.024	0.37	0.078	0.039	0.046
BBD × duration of premenopause	-0.002	0.011	0.84	-0.021	0.014	0.14
BBD × duration postmenopause	-0.012	0.006	0.051	-0.018	0.009	0.044
HT use						
Duration oral estrogen alone	0.021	0.007	<0.001	0.016	0.008	0.047
Duration oral estrogen plus progesterone	0.015	0.008	0.056	0.035	0.008	<0.001
Current use	0.368	0.093	<0.001	0.202	0.118	0.087
Past use	0.087	0.065	0.18	0.092	0.098	0.35
BMI (kg/m <sup>2</sup> )						
Estrogen positive <sup>b</sup>	-0.00082	0.00024	<0.001	0.00000	0.00024	0.99
Estrogen negative <sup>c</sup>	0.00195	0.00042	<0.001	0.00038	0.00056	0.50
Height (in.)						
Estrogen positive	0.00035	0.00032	0.27	0.00031	0.00020	0.12
Estrogen negative	0.00033	0.00098	0.74	-0.00030	0.00015	0.049
Alcohol intake (g)						
Premenopause	0.00048	0.00014	<0.001	0.00042	0.00021	0.039
Postmenopause, while on HT	0.00004	0.0003	0.99	0.00015	0.00045	0.74
Postmenopause, while not on HT	0.00013	0.00019	0.48	-0.00032	0.00037	0.39
Family history of breast cancer	0.403	0.059	<0.001	0.346	0.069	<0.001

<sup>a</sup> Age at 1st birth minus age at menarche if parous, = 0 if nulliparous

<sup>b</sup> Either premenopause or postmenopause while on HT

<sup>c</sup> Postmenopause while not on HT

without any updating going forward. As seen in Tables 8 and 9 the imputation performed well in terms of ever use (*c* statistic 0.87) and duration of use of estrogen alone and estrogen plus progestin. Assessment indicates such imputation is robust for 5 years, though predictive performance

may attenuate over longer follow-up or prediction time intervals.

To fit the Gail model we used a common approach in both cohorts and used family history positive without the added detail of more than one relative. An extremely small



**Table 5** *c* Statistics by study, time period, and age group

Time period	Age group	Number of cases	Log-incidence model		Gail model AUC	Difference				
			AUC	SE		SE	AUC	SE	<i>z</i> value	<i>p</i> value
NHS										
1994–2008	47–87	2,026	0.597	0.007	0.562	0.006	0.034	0.007	4.857	1.191E–06
1994–1999	47–69	852	0.608	0.011	0.569	0.010	0.040	0.011	3.636	2.765E–04
1994–1999	70–87	301	0.587	0.016	0.555	0.017	0.032	0.017	1.882	0.060
2000–2008	47–69	461	0.599	0.013	0.572	0.013	0.026	0.012	2.167	0.030
2000–2008	70–87	412	0.577	0.016	0.543	0.014	0.034	0.016	2.125	0.034
California Teachers Study										
1995–2009	47–87	1,400	0.586	0.009	0.547	0.008	0.040	0.009	4.444	8.812E–06
1995–2000	47–69	422	0.609	0.015	0.572	0.014	0.037	0.014	2.643	0.008
1995–2000	70–87	144	0.564	0.025	0.516	0.024	0.048	0.028	1.714	0.086
2001–2009	47–69	431	0.591	0.016	0.537	0.014	0.054	0.015	3.600	0.000
2001–2009	70–87	403	0.565	0.015	0.543	0.014	0.023	0.016	1.438	0.151

*AUC* area under the curve, *SE* standard error

**Table 6** Comparison of *c* statistics with actual updated hormone therapy (HT) data versus imputed HT data by time period and age group, NHS data 1994–2008

Time period	Age group <sup>a</sup>	Imputation of follow-up HT data <sup>c</sup>			Time period	Age group <sup>b</sup>	Actual updated HT data		
		Number of cases	AUC	SE			Number of cases	AUC	SE
1994–2008	47–87	2,026	0.600	0.006	1994–2008	47–87	2,026	0.616	0.006
1994–1999	47–69	852	0.620	0.010	1994–1999	47–69	851	0.618	0.010
1994–1999	70–87	301	0.585	0.016	1994–1999	70–87	302	0.590	0.016
2000–2008	47–69	461	0.595	0.013	2000–2008	47–69	457	0.641	0.013
2000–2008	70–87	412	0.575	0.015	2000–2008	70–87	416	0.626	0.014

*AUC* area under the curve, *SE* standard error

<sup>a</sup> Age group was defined by updating baseline (1994) age by 1 year for each succeeding year

<sup>b</sup> Age group was defined by using actual questionnaire age based on follow-up questionnaires

<sup>c</sup> Based on an average of five imputations of follow-up HT data

**Table 7** Calibration of the NHS model in the California Teachers Study

	Risk decile									
	1	2	3	4	5	6	7	8	9	10
Observed number of cases	71.6	94.6	111.6	119.6	126.0	129.6	152.0	150.4	195.6	248.8
Expected number of cases	66.6	88.6	103.1	116.3	128.6	143.6	160.3	181.4	208.0	272.4
$\alpha$	–0.048									
SE ( $\alpha$ )	0.027									
<i>p</i> value	0.074									

*SE* standard error

fraction of all cohort members have more than one relative with breast cancer, limiting the impact of this truncation of data.

Review of evidence shows many models of breast cancer incidence have been developed, but few are validated, and perhaps even fewer evaluated for performance in clinical

**Table 8** Imputation models for estimating ever/never use of estrogen alone and duration of use of estrogen alone, NHS, 1995–2006 as a function of baseline (1994) covariates

Variable	Ever/never use of estrogen alone ( $n = 45,742$ ) <sup>a</sup>			ln(duration estrogen alone) ( $n = 9,145$ ) <sup>b</sup>		
	Beta	SE	<i>p</i> value	Beta	SE	<i>p</i> value
Constant	−3.078	0.212		0.762	0.105	
Duration of premenopause	−0.002	0.005	0.68	−0.004	0.003	0.15
Duration postmenopause						
Natural menopause	−0.045	0.004	<0.001	−0.020	0.002	<0.001
Bilateral oophorectomy	0.011	0.004	0.005	−0.007	0.002	0.002
Pregnancy history						
Gynecologic age at 1st birth <sup>c</sup>	0.0033	0.0035	0.35	−0.0022	0.0019	0.25
Birth index	0.0031	0.0005	<0.001	0.00052	0.00027	0.049
BBD						
BBD (yes vs. no)	0.863	0.371	0.020	0.213	0.181	0.24
BBD × age at menarche	−0.035	0.017	0.032	−0.0001	0.0085	0.99
BBD × duration of premenopause	−0.011	0.007	0.10	−0.0032	0.0032	0.32
BBD × duration postmenopause	−0.007	0.005	0.16	−0.0080	0.0025	0.001
HT use						
Duration oral estrogen alone	0.235	0.005	<0.001	0.0282	0.0017	<0.001
Duration oral estrogen + progesterone	−0.041	0.007	<0.001	−0.0027	0.0041	0.52
Current use	1.978	0.053	<0.001	0.993	0.037	<0.001
Past use	0.580	0.059	<0.001	0.275	0.041	<0.001
BMI (kg/m <sup>2</sup> )						
Estrogen positive <sup>d</sup>	0.00035	0.00015	0.015	−0.00007	0.00007	0.34
Estrogen negative <sup>e</sup>	−0.00081	0.00045	0.074	−0.00002	0.00027	0.94
Height (in.)						
Estrogen positive <sup>d</sup>	0.00003	0.00019	0.88	−0.00015	0.00009	0.10
Estrogen negative <sup>e</sup>	−0.00024	0.00090	0.79	0.00113	0.00052	0.030
Alcohol intake (g)						
Premenopause	−0.00012	0.00010	0.23	−0.00006	0.00006	0.26
Postmenopause, while on HT	−0.00102	0.00023	<0.001	0.00009	0.00009	0.29
Postmenopause, while not on HT	0.00073	0.00018	<0.001	−0.00013	0.00012	0.29
Family history of breast cancer	0.148	0.046	0.30	−0.041	0.025	0.099
<i>c</i> statistic	0.871			–		
<i>R</i> <sup>2</sup>	–			0.271		

<sup>a</sup> Over 12 years (1995–2006)<sup>b</sup> Among women with duration of estrogen alone >0 over 12 years (1995–2006)<sup>c</sup> Age at 1st birth minus age at menarche if parous, = 0 if nulliparous<sup>d</sup> Either premenopause or postmenopause while on HT<sup>e</sup> Postmenopause while not on HT

settings. This applies more broadly than just breast or other cancer prediction—with limited validation and evaluation of clinical impact of prediction models on disease outcomes. For breast cancer, Meads [10] show the range of variables included is substantial with many models not including menopause, type of menopause, or use of postmenopausal HT, or alcohol intake. Other than the Rosner–Colditz model based on NHS data, only Boyle includes alcohol [21], a known carcinogen for breast cancer [17], and

age at menopause is only included by Rosner–Colditz and Tyrer [22]. Parity and BMI are more broadly included across models [10]. The most complete of the 17 models summarized by Meads is the Rosner–Colditz model with external validity now established in this independent data set. Several models were assessed for performance by Amir et al. [23] in a UK population of 4,536 women attending a “family history and hereditary screening programme”, among whom 52 developed breast cancer. The Tyrer–

**Table 9** Imputation models for estimating ever/never use of estrogen plus progestin (E&P) and duration of use of E&P, NHS, 1995–2006 as a function of baseline (1994) covariates

Variable	Ever/never use of E&P ( $n = 45,742$ ) <sup>a</sup>			ln(duration E&P) ( $n = 11,516$ ) <sup>b</sup>		
	Beta	SE	<i>p</i> value	Beta	SE	<i>p</i> value
Constant	0.405	0.212		1.291	0.115	
Duration of premenopause	−0.041	0.005	<0.001	−0.0061	0.0030	0.038
Duration postmenopause						
Natural menopause	−0.077	0.003	<0.001	−0.018	0.002	<0.001
Bilateral oophorectomy	−0.309	0.008	<0.001	−0.043	0.004	<0.001
Pregnancy history						
Gynecologic age at 1st birth <sup>c</sup>	0.0010	0.0031	0.74	0.0012	0.0016	0.48
Birth index	0.0000	0.0004	0.94	−0.0008	0.0002	<0.001
BBD						
BBD (yes vs. no)	−0.336	0.399	0.40	−0.344	0.211	0.10
BBD × age at menarche	−0.003	0.016	0.84	0.0100	0.0080	0.21
BBD × duration of premenopause	0.009	0.008	0.24	0.0049	0.0041	0.23
BBD × duration postmenopause	0.001	0.004	0.84	−0.0016	0.0025	0.52
HT use						
Duration oral estrogen alone	−0.048	0.007	<0.001	0.012	0.004	0.001
Duration oral estrogen plus progesterone	0.414	0.008	<0.001	0.062	0.003	<0.001
Current use	1.356	0.037	<0.001	0.471	0.023	<0.001
Past use	0.453	0.041	<0.001	−0.179	0.028	<0.001
Body mass index (kg/m <sup>2</sup> )						
Estrogen positive <sup>d</sup>	−0.00032	0.00014	0.027	−0.00003	0.00008	0.65
Estrogen negative <sup>e</sup>	−0.00088	0.00042	0.036	−0.00011	0.00028	0.69
Height (in.)						
Estrogen positive <sup>d</sup>	−0.00026	0.00019	0.19	0.00005	0.00009	0.58
Estrogen negative <sup>e</sup>	0.00031	0.00084	0.71	−0.00079	0.00051	0.12
Alcohol intake (g)						
Premenopause	0.00010	0.00009	0.29	−0.00001	0.00005	0.88
Postmenopause, while on HT	−0.00121	0.00032	<0.001	−0.00001	0.00014	0.96
Postmenopause, while not on HT	0.00016	0.00017	0.37	−0.00023	0.00012	0.055
Family history of breast cancer	−0.23	0.044	<0.001	−0.035	0.024	0.14
<i>c</i> statistic	0.870			–		
<i>R</i> <sup>2</sup>	–			0.207		

<sup>a</sup> Over 12 years (1995–2006)<sup>b</sup> Among women with duration of E&P > 0 over 12 years (1995–2006)<sup>c</sup> Age at 1st birth minus age at menarche if parous, = 0 if nulliparous<sup>d</sup> Either premenopause or postmenopause while on HT<sup>e</sup> Postmenopause while not on HT

Cuzick model [22] had the best performance based on *c* statistic, though the O/E performance was at the level of 0.8 for this model compared to 0.9 for Gail [23]. While Amir and Tyrer–Cuzick have been evaluated in high-risk populations where they are likely to perform better, such a comparison in the general population has not been reported.

For CHD on the other hand, Van Dieren et al. [24] review evidence on model development and evaluation—45 prediction models reported in the literature, 12 specific for patients with diabetes; 31 % validated in independent population of diabetics, and only one evaluated in clinic for its effect on patient management.

## Calibration

While age-standardized incidence rates differ between NHS and CTS the coefficients for risk factors when fitted to the Rosner–Colditz breast cancer incidence model are quite comparable and evaluation of predicted incidence in the calibration analysis shows no significant deviation from SEER incidence, with O/E of 0.96. The range of incidence expected in the SEER calibration study reveals approximately fourfold difference in expected values between lowest and highest decile. This is a non-trivial spread in risk across deciles and is evaluated by the Poisson regression to assess trend in difference between O and E over deciles of risk. The observed lower incidence in NHS may reflect cohort follow-up procedures that do not fully capture incident breast cancers as efficiently as the surveillance through the state tumor registry in California, a state with historically low out migration. As all women should have access to Medicare after age 65, differential screening and access to care should not be an issue when comparing these two cohorts.

## Future issues

Future applications in routine clinical settings will add further modeling issues. For example, as approximately one-third of women report hysterectomy in the United States and because age at menopause is an important risk factor in our model, we will need to impute estimated age at menopause among women with hysterectomy before menopause. We have previously derived an algorithm for use in this setting [25]. Other missing data will also need to be addressed, likely using NHANES data as has been implemented in clinical applications of a risk model for progression of age-related macular degeneration using demographic, genetic, environmental, and ocular factors [26]. Other clinical application data come from the United Kingdom where Evans and colleagues have collected breast risk data in a routine breast screening setting, and report evaluation of the Tyrer and Cuzick breast risk model at the level of distributions of 10-year risk and also assess SNPs in a subset of women. Approximately 34 % of women attending breast screening enrolled and risk estimates were returned to those with 10-year risk above 8 % (107 women). Performance assessment of the tool is ongoing in this routine mammography setting [27]. The breast cancer surveillance consortium generated a risk prediction model among more than 1 million women undergoing mammography [28]. They began with age, race, ethnicity, and breast density (measure with BI-RADS) and adjusted estimates of family history and history of breast biopsy. The model was developed in 60 % of the population and validated in the remaining 40 %, and is

well calibrated, though it does not include any reproductive or lifestyle predictors of breast cancer. While these two examples indicate that risk factors and prediction can be incorporated into mammography services, issues of missing data and real time estimation of risk have yet to be addressed, and the impact of risk presentation on clinical decision making and outcomes of care has not been evaluated.

## Conclusion

Through validation in an independent data set, we have shown that the Rosner–Colditz model performs consistently when applied in that independent setting. Performance is stronger predicting incidence among women 47–69 years and over a 5-year time interval. AUC values are significantly higher than the Gail model in the independent validation data set, and may be further improved with addition of breast density or other markers of risk beyond the current model. Further refinement may be needed to handle missing data in routine clinical settings.

**Acknowledgments** This work was supported by the National Cancer Institute, National Institute of Health (PO1 CA87969). G.A.C. is also supported by an American Cancer Society Clinical Research Professorship and the Breast Cancer Research Foundation. LB, JVL, and JSH are supported by RO1 CA077398. LB is also supported by K05 CA136967. We also acknowledge insightful comments from our colleagues Drs. Tamimi and Willett, programming support of Marion McPhee and Rong Chen and secretarial support of Virginia Piaseczny.

**Conflict of interest** The authors declare they have no conflict of interest.

**Ethical standards** All data collection was conducted with approval of appropriate institutional review boards to protect human subjects with consent and data protection systems in place. Data analysis for this manuscript was conducted on de-identified data sets.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

## Appendix 1: Calibration procedure—CTS validation study

We follow the general calibration procedure of Gail (1989).

Procedure for the *m*th imputation

1. We wish to apply the NHS risk model to the CTS population, where the incidence model is of the form:

$$\ln(I) = \alpha + \sum_{k=1}^K \beta_k x_k$$

2. Suppose there are  $N$  subjects in the CTS population who are followed collectively for  $T$  person–years.
3. Suppose the  $i$ th subject is followed for  $t_i$  person years, where:

$$\sum_{i=1}^N t_i = T,$$

and that  $m_{ij}$  = age of the  $i$ th subject at the  $j$ th person–year.

4. We divide the  $T$  person–years into  $L$  age strata and let  $T_l$  = number of person–years in the  $l$ th age stratum, where

$$T = \sum_{q=1}^L T_q,$$

and

$$\begin{aligned} q = 1 & \text{ if } a_1 \leq m_{ij} < a_2, \\ & = 2 \text{ if } a_2 \leq m_{ij} < a_3, \\ & \dots \\ & = l \text{ if } a_l \leq m_{ij} < a_{l+1}, \\ & \dots \\ & = L \text{ if } a_L \leq m_{ij} < a_{L+1}. \end{aligned}$$

In this case,  $L = 8$  and the age groups are defined by 45–49, 50–54, 55–59, 60–64, 65–69, 70–74, 75–79, 80–87.

5. We define a person as being at baseline risk in the  $l$ th age stratum if  $\mathbf{x} = \mathbf{0}$ .
6. We define covariate values for the  $i$ th person in the  $j$ th person–year by  $x_{ij1}, \dots, x_{ijK}$ , and let

$$RR_{ij} = \exp\left(\sum_{k=1}^K \hat{\beta}_k x_{ijk}\right),$$

$i = 1, \dots, N; j = 1, \dots, t_i.$

7. Let
  - $Y_l$  = observed number of cases in age group  $l$ ,  $l = 1, \dots, L$ ;
  - $O_{ij} = 1$  if the  $i$ th subject becomes a case in the  $j$ th person – year, = 0 otherwise.
  - $\delta_{ijl} = 1$  if  $a_l \leq m_{ij} < a_{l+1}$ , = 0 otherwise,  $i = 1, \dots, N; j = 1, \dots, t_i; l = 1, \dots, L$ .

8. Let
 
$$F_l = \sum_{i=1}^N \sum_{j=1}^{t_i} \frac{O_{ij} \delta_{ijl} / Y_l}{RR_{ij}}, \quad l = 1, \dots, L.$$

9. Let
  - $h_1^*(l)$  = age-specific incidence rate from SEER 1995–2006,  $l = 1, \dots, L$ .
10. Let
  - $h_1(l) = h_1^*(l)F_l$ ,
  - = baseline incidence rate in the  $l$ th age stratum of CTS,  $l = 1, \dots, L$ .
11. An estimate of the incidence rate for the  $i$ th subject in the  $j$ th person–year is given by:
 
$$\hat{I}_{ij} = \sum_{l=1}^L h_1(l) \delta_{ijl} RR_{ij}, \quad i = 1, \dots, N; j = 1, \dots, t_i$$
12. An estimate of the cumulative incidence for the  $i$ th subject over  $t_i$  person–years is given by:
 
$$E_i = 1 - \exp\left(-\sum_{j=1}^{t_i} \hat{I}_{ij}\right), \quad i = 1, \dots, N$$
13. Let
  - $O_i = 1$  if the  $i$ th subject is a case over  $t_i$  person-years,
  - = 0 if the  $i$ th subject is a control over  $t_i$  person-years,  $i = 1, \dots, N$ .
14. Compute
  - $E_i^* = E_i/t_i$  = cumulative incidence per person–year for the  $i$ th subject,  $i = 1, \dots, N$  and rank subjects by decile of  $E_i^*$ .
15. Let
  - $\lambda_{id} = 1$  if subject  $i$  is in the  $d$ th decile of  $E_i^*$ ,
  - = 0 otherwise,  $i = 1, \dots, N; d = 1, \dots, 10$ .
16. Compute
 
$$O^{(d)} = \sum_{i=1}^N O_i \lambda_{id} = \text{observed number of cases in the } d\text{th decile of } E_i^*,$$

$$E^{(d)} = \sum_{i=1}^N E_i \lambda_{id} = \text{expected number of cases in the } d\text{th decile of } E_i^*$$
17. *Poisson regression at the aggregate level*  
 Run Poisson regression of  $O^{(d)}$  using  $\ln(E^{(d)})$  as the offset based on the model:
 
$$\ln(\mu_d) = \alpha + \ln(E^{(d)}), \quad \text{where } \mu_d = E(O^{(d)})$$
 or equivalently  $\mu_d = \exp(\alpha)E^{(d)}$   
 Note in the above model,  $\alpha = 0$  implies that  $\mu_d = E^{(d)}$ .

18. *Poisson regression at the individual level*  
Run Poisson regression of  $O_i$  using  $\ln(E_i)$  as the offset based on the model:

$$\ln(\mu_i) = \alpha + \ln(E_i),$$

or

$$O_i = \exp(\alpha)E_i,$$

where

$$\mu_i = E(O_i).$$

Note  $\alpha = 0$  implies that  $\mu_i = E_i$ . Also, the models in steps 17 and 18 are equivalent because  $O^{(d)}$  is a sum of Poisson distributions which is also Poisson, where

$$\mu_d = \sum_{i=1}^N \mu_i \lambda_{id} = \exp(\alpha) \sum_{i=1}^N E_i \lambda_{id} = \exp(\alpha) E^{(d)}.$$

19. If we average SEER incidence rates for white women from 1995 to 2000 and 2000 to 2006, we obtain:

Age	SEER incidence rate (per 10 <sup>5</sup> person-years)	CTS incidence rate (per 10 <sup>5</sup> person-years)	CTS number of cases
45–49	253.3	64.2	1
50–54	336.8	269.2	47
55–59	417.0	407.1	173
60–64	494.2	493.2	297
65–69	538.8	534.6	328
70–74	579.3	540.4	268
75–79	592.6	582.0	166
80–87	517.4	476.1	55

We used the SEER incidence rates in column 2 as our estimates of  $h_1^*(l)$ ,  $l = 1, \dots, L$

Overall inference over  $m$  imputations

20. For step 16, let  $O_m^{(d)}, E_m^{(d)}, T_m^{(d)}$  = observed count, expected count and person-years for the  $d$ th risk decile in the  $m$ th imputation. The overall estimates of  $O^{(d)}, E^{(d)}$  and  $T^{(d)}$  are given by:

$$\begin{aligned} O^{(d)} &= \sum_{m=1}^M O_m^{(d)} / M, E^{(d)} = \sum_{m=1}^M E_m^{(d)} / M, T^{(d)} \\ &= \sum_{m=1}^M T_m^{(d)} / M. \end{aligned}$$

For steps 17 and 18, we obtain the estimate:

$\hat{\alpha}^{(m)}$  for the  $m$ th imputation and associated estimate

$\text{var}[\hat{\alpha}^{(m)}]$ . The overall point estimate is:

$$\hat{\alpha} = \sum_{m=1}^M \hat{\alpha}^{(m)} / M,$$

with associated variance given by:

$$\begin{aligned} \text{var}(\hat{\alpha}) &= \sum_{m=1}^M \text{var}[\hat{\alpha}^{(m)}] / M \\ &+ [(M + 1) / M] \sum_{m=1}^M [\hat{\alpha}^{(m)} - \hat{\alpha}]^2 / (M - 1), \end{aligned}$$

and test statistic

$$\begin{aligned} z_{\hat{\alpha}} &= \hat{\alpha} / [\text{var}(\hat{\alpha})]^{1/2} \sim N(0, 1) \text{ under } H_0 \text{ that } \alpha = 0, \\ p\text{-value} &= 2 \times [1 - \Phi|z_{\hat{\alpha}}|]. \end{aligned}$$

**Appendix 2: Imputation strategy for ever use and duration of use of postmenopausal hormone therapy**

1. Let

$$\begin{aligned} \ln\left(\frac{P_{i1}}{1 - P_{i1}}\right) &= \alpha_1 + \sum_{k=1}^K \beta_{k1} x_{ik} \\ \ln\left(\frac{P_{i2}}{1 - P_{i2}}\right) &= \alpha_2 + \sum_{k=1}^K \beta_{k2} x_{ik} \end{aligned} \tag{1}$$

where  $p_{i1}$  = Prob(duration estrogen alone > 0, 1994–2006) for the  $i$ th NHS woman,  $i = 1, \dots, 45742$ ,  $p_{i2}$  = Prob(duration E&P > 0, 1994–2006) for the  $i$ th NHS woman,  $I = 1, \dots, 45742$ ,  $x_{ik}$  =  $k$ th breast cancer risk factor for the  $i$ th subject,  $I = 1, \dots, 45742$ ,  $k = 1, \dots, K$  and let  $\hat{p}_{i1}, \hat{p}_{i2}$  be the corresponding estimated probabilities obtained by substituting the estimated parameters  $\hat{\alpha}_1, \hat{\alpha}_2$  and  $\hat{\beta}_{k1}, \hat{\beta}_{k2}$ ,  $k = 1, \dots, K$  for the true parameters in Eq. 1.

2. Let  $y_{i1} = \ln(\text{duration of use of estrogen alone for the } i\text{th NHS woman, where } y_{i1} > 0, y_{i2} = \ln(\text{duration of use of E \& P for the } i\text{th NHS woman, where } y_{i2} > 0.$
3. (a) Let

$$\begin{aligned} z_{i1} &= \alpha_1^* + \sum_{k=1}^K \beta_{k1}^* x_{ik} + e_{i1}^*, \\ z_{i2} &= \alpha_2^* + \sum_{k=1}^K \beta_{k2}^* x_{ik} + e_{i2}^* \end{aligned} \tag{2}$$

where

$$z_{i1} = \ln(y_{i1}), z_{i2} = \ln(y_{i2}), e_{i1} \sim N(0, \sigma_{i1}^2), e_{i2} \sim N(0, \sigma_{i2}^2)$$

be linear regressions of  $z_{i1}$  and  $z_{i2}$  on  $\tilde{x}_i$ , respectively.

(b) Let

$$\hat{z}_{i1}^* = \hat{\alpha}_1^* + \sum_{k=1}^K \hat{\beta}_{k1}^* x_{ik},$$

$$\hat{z}_{i2}^* = \hat{\alpha}_2^* + \sum_{k=1}^K \hat{\beta}_{k2}^* x_{ik}$$

be the corresponding predicted values of  $z_{i1}$  and  $z_{i2}$ , respectively.

(c) Let  $\hat{\sigma}_1^2, \hat{\sigma}_2^2$  be the estimated residual variances corresponding to  $z_{i1}$  and  $z_{i2}$ , in Eq. 2, respectively.

4. Let  $U_{i1}, U_{i2}$  be  $U(0,1)$  random variables and let  $u_{i1}, u_{i2}$  be the realization of these random variables generated using the RANUNI function of SAS.

Let  $V_{i1}, V_{i2}$  be corresponding  $N(0,1)$  random variables and let  $v_{i1}, v_{i2}$  be the realization of these random variables generated using the RANNOR function of SAS.

5. Let  $\hat{y}_{i1}, \hat{y}_{i2}$  = imputed estimate of duration of estrogen alone and duration of E&P, respectively, for the  $i$ th NHS woman.

(a) If  $\hat{p}_{i1} < u_{i1}$ , then  $\delta_{i1} = 1$ , else  $\delta_{i1} = 0$ .

(b) If  $\hat{p}_{i2} < u_{i2}$ , then  $\delta_{i2} = 1$ , else  $\delta_{i2} = 0$ .

(c) Let  $\hat{w}_{i1} = \hat{z}_{i1}^* + \hat{\sigma}_1 v_{i1}$ ,  
 $\hat{w}_{i2} = \hat{z}_{i2}^* + \hat{\sigma}_2 v_{i2}$ .

(d) (i) If  $\delta_{i1} = \delta_{i2} = 0$ , then  $\hat{y}_{i1} = \hat{y}_{i2} = 0$ ,

(ii) If  $\delta_{i1} = 1$  and  $\delta_{i2} = 0$ , then  $\hat{y}_{i1} = \min[\exp(\hat{w}_{i1}), 12], \hat{y}_{i2} = 0$ ,

(iii) If  $\delta_{i1} = 0$  and  $\delta_{i2} = 1$ , then  $\hat{y}_{i1} = 0, \hat{y}_{i2} = \min[\exp(\hat{w}_{i2}), 12]$ ,

(iv) If  $\delta_{i1} = 1$  and  $\delta_{i2} = 1$ , and  $\hat{w}_{i1} > \hat{w}_{i2}$ , then  $\hat{y}_{i1} = \min[\exp(\hat{w}_{i1}), 12], \hat{y}_{i2} = 0$ ,

(v) If  $\delta_{i1} = 1$  and  $\delta_{i2} = 1$ , and  $\hat{w}_{i1} < \hat{w}_{i2}$ , then  $\hat{y}_{i1} = 0, \hat{y}_{i2} = \min[\exp(\hat{w}_{i2}), 12]$ .

6. (a) Based on steps 1–5, we estimate duration of estrogen alone and duration of E&P for all NHS women and fit the log-incidence model described in step 1, assuming that duration of HT is continuous starting in 1994 and ceases after duration  $\hat{y}_{i1}$  or  $\hat{y}_{i2}$ , respectively.

(b) We repeat (a) five times and obtain five separate estimates of  $\alpha$  and  $\beta$ , respectively. Let  $\hat{\alpha}_m, \hat{\beta}_{k,m}$  be the estimates of  $\alpha$  and  $\beta_k$  for the  $m$ th imputation,  $k = 1, \dots, K$ .

The overall estimates are given by:

$$\hat{\alpha} = \sum_{m=1}^5 \hat{\alpha}_m / 5, \hat{\beta}_k = \sum_{m=1}^5 \hat{\beta}_{k,m} / 5,$$

with

$$\text{var}(\hat{\alpha}) = \sum_{m=1}^5 \text{var}(\hat{\alpha}_m) / 5 + (6/5) \sum_{m=1}^5 (\hat{\alpha}_m - \hat{\alpha})^2 / 4,$$

$$\text{var}(\hat{\beta}_k) = \sum_{m=1}^5 \text{var}(\hat{\beta}_{k,m}) / 5 + (6/5) \sum_{m=1}^5 (\hat{\beta}_{k,m} - \hat{\beta}_k)^2 / 4.$$

7. A similar strategy as in step 6 is used to obtain multiple imputation estimates of AUC based on imputed data, where  $\text{Var}(AUC)$  is obtained using the methods in Rosner and Glynn [20].
8. For imputation with the CTS data, we use the NHS prediction equations in (1) and (2) and proceed as in steps 5 and 6 to obtain overall estimates of  $\alpha$  and  $\beta$  for CTS as given in Table 4.
9. Estimates of AUC are obtained for CTS based on the coefficients  $\hat{\alpha}$  and  $\hat{\beta}$  obtained from NHS data in step 6. The AUC estimates were obtained for five imputed CTS datasets and combined using the methods in step 6.
10. Similarly, calibration of the NHS model based on the coefficients  $\hat{\alpha}$  and  $\hat{\beta}$  in step 6 were applied to five imputed CTS datasets. Separate estimates of observed counts, expected counts and Poisson regression parameters were obtained for each imputed CTS dataset and combined using multiple imputation methods described in Appendix 1.

### References

1. Rosner B, Colditz GA (1996) Nurses' health study: log-incidence mathematical model of breast cancer incidence. *J Natl Cancer Inst* 88(6):359–364
2. Colditz G, Rosner B (2000) Cumulative risk of breast cancer to age 70 years according to risk factor status: data from the Nurses' Health Study. *Am J Epidemiol* 152(10):950–964
3. Colditz G, Rosner B, Chen WY, Holmes M, Hankinson SE (2004) Risk factors for breast cancer: according to estrogen and progesterone receptor status. *J Natl Cancer Inst* 96:218–228
4. Trichopoulos D, Hsieh CC, MacMahon B et al (1983) Age at any birth and breast cancer risk. *Int J Cancer* 31(6):701–704
5. Rosner B, Colditz GA, Willett WC (1994) Reproductive risk factors in a prospective study of breast cancer: the Nurses' Health Study. *Am J Epidemiol* 139(8):819–835
6. Lambe M, Hsieh C-c, Trichopoulos D, Ekblom A, Pavia A, Adami H-O (1994) Transient increase in risk of breast cancer after giving birth. *N Engl J Med* 331:5–9
7. Colditz GA, Rosner BA (2006) What can be learnt from models of incidence rates? *Breast Cancer Res* 8(3):208
8. Moons KG, Kengne AP, Grobbee DE et al (2012) Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 98(9):691–698
9. Moons KG, Kengne AP, Woodward M et al (2012) Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart* 98(9):683–690

10. Meads C, Ahmed I, Riley RD (2012) A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance. *Breast Cancer Res Treat* 132(2):365–377
11. Gail MH, Brinton LA, Byar DP et al (1989) Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst* 81:1879–1886
12. Rosner B, Colditz GA, Iglehart JD, Hankinson SE (2008) Risk prediction models with incomplete data with application to prediction of estrogen receptor-positive breast cancer: prospective data from the Nurses' Health Study. *Breast Cancer Res* 10(4):R55
13. Lilienfeld AM (1956) The relationship of cancer of the female breast to artificial menopause and marital status. *Cancer* 9:927–934
14. Trichopoulos D, MacMahon B, Cole P (1972) Menopause and breast cancer risk. *J Natl Cancer Inst* 48(3):605–613
15. International Agency for Research on Cancer (2008) Monograph on the evaluation of carcinogenic risk to humans: combined estrogen/progestogen contraceptives and combined estrogen/progestogen menopausal therapy. Combined estrogen-progestogen contraceptives and combined estrogen-progestogen menopausal therapy, vol 91. IARC Press, Lyon.
16. International Agency for Research on Cancer (2002) Weight control and physical activity, vol 6. International Agency for Research on Cancer, Lyon
17. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans (2007) Alcohol consumption and ethyl carbamate. International Agency for Research on Cancer, Lyon (Distributed by WHO Press, 2010)
18. Colditz GA, Hankinson SE, Hunter DJ et al (1995) The use of estrogens and progestins and the risk of breast cancer in postmenopausal women. *N Engl J Med* 332:1589–1593
19. Bernstein L, Allen M, Anton-Culver H et al (2002) High breast cancer incidence rates among California teachers: results from the California Teachers Study (United States). *Cancer Causes Control* 13:625–635
20. Rosner B, Glynn RJ (2009) Power and sample size estimation for the Wilcoxon rank sum test with application to comparisons of C statistics from alternative prediction models. *Biometrics* 65(1):188–197
21. Boyle P, Mezzetti M, La Vecchia C, Franceschi S, Decarli A, Robertson C (2004) Contribution of three components to individual cancer risk predicting breast cancer risk in Italy. *Eur J Cancer Prev* 13(3):183–191
22. Tyrer J, Duffy SW, Cuzick J (2004) A breast cancer prediction model incorporating familial and personal risk factors. *Stat Med* 23(7):1111–1130
23. Amir E, Evans DG, Shenton A et al (2003) Evaluation of breast cancer risk assessment packages in the family history evaluation and screening programme. *J Med Genet* 40(11):807–814
24. van Dieren S, Beulens JW, Kengne AP et al (2012) Prediction models for the risk of cardiovascular disease in patients with type 2 diabetes: a systematic review. *Heart* 98(5):360–369
25. Rosner B, Colditz GA (2011) Age at menopause: imputing age at menopause for women with a hysterectomy with application to risk of postmenopausal breast cancer. *Ann Epidemiol* 21(6):450–460
26. Seddon JM, Reynolds R, Yu Y, Daly MJ, Rosner B (2011) Risk models for progression to advanced age-related macular degeneration using demographic, environmental, genetic, and ocular factors. *Ophthalmology* 118(11):2203–2211
27. Evans DG, Warwick J, Astley SM et al (2012) Assessing individual breast cancer risk within the U.K. National Health Service Breast Screening Program: a new paradigm for cancer prevention. *Cancer Prev Res* 5(7):943–951
28. Tice JA, Cummings SR, Smith-Bindman R, Ichikawa L, Barlow WE, Kerlikowske K (2008) Using clinical factors and mammographic breast density to estimate breast cancer risk: development and validation of a new predictive model. *Ann Intern Med* 148(5):337–347